IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 October 2023, accepted 13 October 2023, date of publication 17 October 2023, date of current version 24 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3325235

RESEARCH ARTICLE

Controlling Bias Between Categorical Attributes in Datasets: A Two-Step Optimization Algorithm Leveraging Structural Equation Modeling

ENRICO BARBIERATO, ANDREA POZZI, (Member, IEEE), AND DANIELE TESSERA, (Member, IEEE)
Faculty of Mathematical, Physical and Natural Sciences, Catholic University of Sacred Heart, 25133 Brescia, Italy
Corresponding author: Andrea Pozzi (andrea.pozzi@unicatt.it)

ABSTRACT In the realm of data-driven systems, understanding and controlling biases in datasets emerges as a critical challenge. These biases, defined in this study as systematic discrepancies, have the potential to skew algorithmic outcomes and even compromise data privacy. Mutual information serves as a key tool in the analysis, discerning both direct and indirect relationships between variables. Utilizing structural equation modeling, this paper introduces a synthetic dataset generation method founded on a two-step optimization algorithm that aims to fine-tune variable relationships and achieve targeted mutual information levels between attribute pairs. The algorithm's first phase utilizes gradient-less optimization, focusing on individual variables. The subsequent phase harnesses gradient-based methods to unravel deeper variable interdependencies. The approach is dual-purpose: it refines existing datasets for bias mitigation and creates synthetic datasets with defined bias levels, addressing a crucial research gap. Two case studies showcase the methodology. One emphasizes the finesse of network parameter adjustments in a simulated setting. The other applies the methodology to a realistic job hiring dataset, effectively reducing bias while safeguarding key variable relationships. In summary, this paper offers a novel method for bias management, presents tools for quantitative bias adjustments, and provides evidence of the method's broad applicability through varied use cases.

INDEX TERMS Bias mitigation, data fairness, data generation, explainable AI, machine learning, optimization, statistics, structural equation modeling.

I. INTRODUCTION

The increasing availability of large datasets and the growing demand for data-driven decision-making systems have sparked significant interest in understanding and controlling the relationships between variables in datasets. Specifically, controlling bias between categorical attributes has become critical in various domains, such as privacy-preserving data analysis, fairness in machine learning, and causal inference. Left unchecked, bias in datasets can lead to skewed outcomes and unfair advantages, which can be particularly detrimental in domains such as machine learning, where algorithms learn patterns from these datasets. These biased patterns could then be unwittingly replicated and amplified, leading to potentially harmful consequences. In the field of causal inference, bias could lead to erroneous conclusions, hampering our understanding of cause-effect relationships. Reducing prejudice in datasets is vital for the development of unbiased and fair decision-making algorithms, which are increasingly used to support or even replace human judgment in various applications [1]. However, controlling bias in datasets is a formidable challenge, as it requires an in-depth understanding of the underlying causal structure and an effective method for adjusting the relationships between variables.

In response to the multifaceted implications of bias outlined above, this work contributes a systematic approach to comprehensively understand and control the bias within

The associate editor coordinating the review of this manuscript and approving it for publication was Michele Nappi.

datasets, particularly focusing on categorical attributes. By employing innovative techniques and utilizing mutual information as a measure of dependency between variables, this research paves the way for more unbiased, equitable, and reliable data-driven decision making and analysis. It seeks not only to highlight the presence of bias in datasets but also to provide practical and effective strategies for mitigating it, thus ensuring that the derived insights and patterns are reflective of a balanced and fair representation.

At this juncture, understanding the concept of 'bias' is pivotal. Here, 'bias' refers to any systematic discrepancy that alters the representation or treatment of certain categories within a dataset. This can manifest as a statistical over- or under-representation, leading to potential disparities in subsequent analyses or applications. Take, for example, 'demographic parity,' a key fairness metric in machine learning. It insists on equal positive outcome rates across all protected groups in binary classification tasks. For instance, within a loan approval scenario, demographic parity ensures equal approval rates for all groups, assuming all other factors are constant. Besides demographic parity, also mutual information emerges as another crucial metric, extensively used to measure dependency between variables. This metric is a significant asset for providing a comprehensive quantitative measure of bias, due to its capacity to delineate both direct and indirect relationships between variables.

In addition to a suitable quantitative measurement of bias such as mutual information, a method for portraying the intricate web of causal relationships within datasets is paramount. Probabilistic graphical models and Structural Equation Modeling (SEM) emerge as substantial candidates for this task, offering a robust framework for illustrating and understanding the complex interplay between different dataset attributes. As highlighted by [2], [3], [4], [5], and [6], these tools can effectively map out the causal relationships, providing a clear insight into the dynamics at play. Within this context, the work of [7] employs SEM-based probabilistic networks for qualitatively analyzing the effect of altering network parameters on the extent of bias (measured as mutual information) among various dataset attributes.

The present paper introduces a novel method for generating synthetic datasets that effectively control bias between categorical attributes, building upon the work of [7]. Unlike the approach in [7], which showed the possibility of using SEM coefficients to influence bias evaluation post hoc, this paper's method actively adjusts the relationships between variables. The goal is to ensure that a specific mutual information between pairs of attributes in a dataset is attained. Specifically, this study proposes a two-step optimization algorithm to accurately manage bias within the dataset attributes, grounded on the principles of structural equation modeling. The algorithm operates in two sequential stages. The first stage uses a gradient-less optimization method, optimizing each variable independently to avoid issues related to variable coupling. This method enhances efficiency and ensures scalability, particularly for a moderate number of attribute pairs. The second stage uses a gradient-based optimization approach, utilizing an analytical model of a probabilistic network that outlines the causal dependencies between dataset attributes. Thanks to such analytical model, this stage handles more complex interactions between variables, considering the interplay of different attributes for a comprehensive optimization solution. The combined approach of gradient-less and gradient-based methods allows the algorithm to tap into the strengths of both.

It is crucial to emphasize that the proposed methodology can serve as a component of a pre-processing algorithm designed to mitigate bias in a given dataset. Specifically, once the parameters of an appropriate graphical model have been learned from the biased data, the introduced technique can be employed to adjust the configuration of these parameters, aiming to diminish the bias while minimizing any adverse effects on the remaining causal relationships. Subsequently, an unbiased version of the dataset can be synthetically generated by utilizing the graphical model with the modified parameters. Furthermore, this method is significant for producing synthetic datasets with varying degrees of bias, addressing the scarcity of publicly available datasets for research. This feature is especially helpful for researchers focused on developing and testing bias-reducing classifiers, as it enables the creation of synthetic datasets across a broad range of bias scenarios.

The efficacy of the proposed methodology is demonstrated through two distinct case studies, presented in the results section. The first case study showcases the algorithm's capacity to finely adjust network parameters within a numerically simulated environment, ensuring that the desired mutual information among selected pairs of dataset attributes is achieved. It underscores both the strengths and the limitations of the approach, emphasizing the intrinsic relationship between the desired mutual information and the degree of freedom inherent to the topological structure of the graphical model. The second case study applies the proposed methodology to a more realistic context, specifically focusing on a dataset related to job hiring decisions. It demonstrates how the proposed method can be employed to mitigate potential bias while maintaining the relationships among other variables, thereby facilitating the synthetic generation of a dataset that closely mimics the original one, but with significantly reduced bias.

Summarizing, the key contributions of this paper include:

- Introduction of a novel synthetic dataset generation method using a two-step optimization algorithm for effective bias control between categorical attributes within datasets.
- Proposal of a quantitative approach to adjust relationships between variables, ensuring a desired level of mutual information between attribute pairs in a dataset.

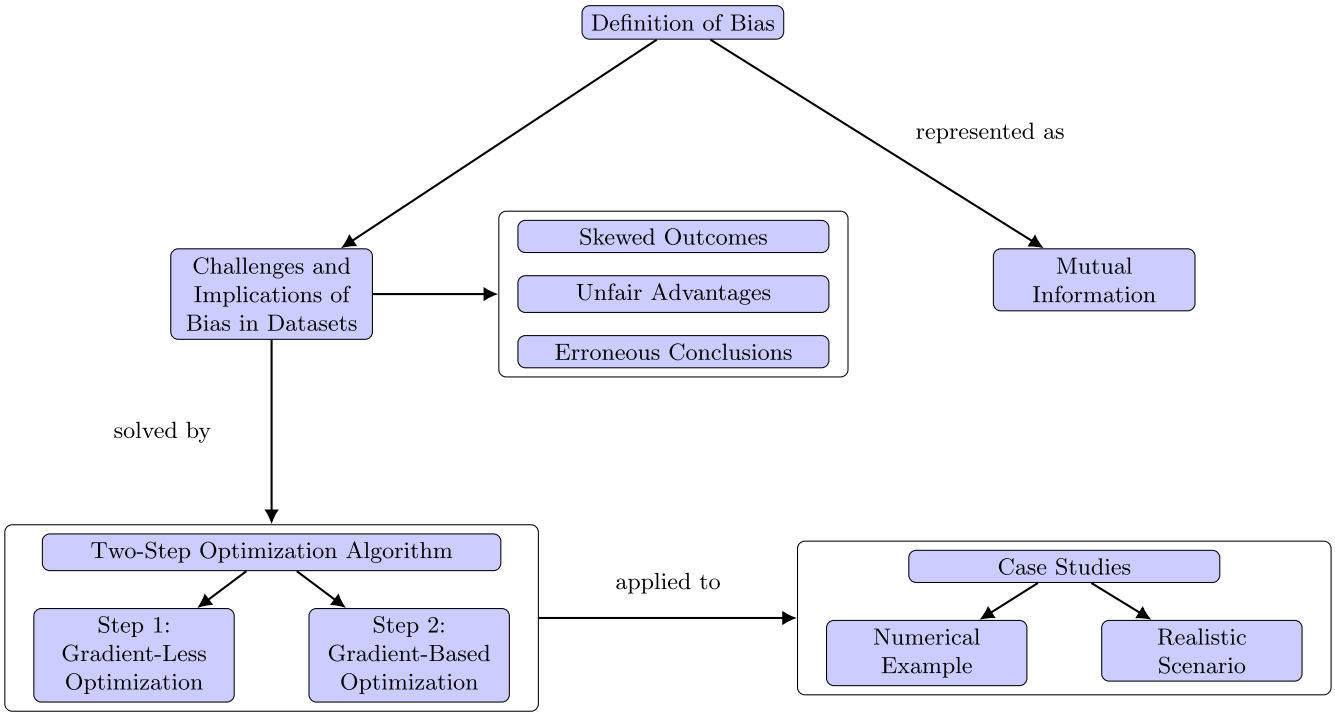


FIGURE 1. Visual representation of the key points proposed and discussed in this paper. The diagram succinctly illustrates the primary elements and their interconnections, including the definition of bias, challenges, two-step optimization algorithm, and case studies, providing a coherent and concise visual overview that complements the textual exposition.

- Development and validation of a tool for bias mitigation in existing datasets and generation of synthetic datasets with specified bias degrees.

These contributions together underline the key role of this paper in the ongoing efforts towards effective bias control within datasets. A visual representation of the key points proposed and discussed in this paper is depicted in Figure 1.

This paper is organized as follows: Section II provides an overview of related work in the field of bias and fairness in datasets. In Section III, the methodology is described in detail, including subsections on structural equation modeling (III-A), synthetic dataset generation (III-B), mutual information (III-C), and optimization (III-D). Section IV presents the aforementioned case studies. Finally, Section V concludes the paper with a summary of the main findings and potential directions for future research.

II. RELATED WORK

This section provides an overview of the related work in the field of bias and fairness in datasets, focusing on measures of dependency, synthetic dataset generation, and fairness in machine learning. A short comparative analysis is reported at the end of each subsection.

A. BIAS DEFINITIONS AND MEASURES OF DEPENDENCY

Bias in data has been defined in different ways. According to Amodei et al. [8] claim that "bias as "systematic errors in a machine learning model that arise from the data it is trained on." The authors argue that bias can have a number of side effects, including i) discrimination against certain groups of people, unfairness in decision-making, and iii) reduced trust in machine learning systems. Caliskan et al. [9] define bias as "the tendency of a machine learning model to make different predictions for different groups of people, even when those groups are statistically equivalent." Specifically, the authors argue that bias can be caused by a number of factors, varying from the data and algorithm used to the way in which the model is used. Finally, Rudin [10] denotes bias as "the difference between the predictions of a machine learning model and the ground truth."

Several measures of dependency between variables have been proposed in the literature, with mutual information being one of the most widely used [11]. Mutual information quantifies the amount of information shared between two variables, making it suitable for assessing bias between categorical attributes in datasets [12]. Other measures of dependency, such as Pearson's correlation coefficient and distance correlation, have also been employed in various contexts [13], [14], [15]. Besides mutual information and Pearson's correlation, there are several other dependency measures commonly used to quantify the relationship between variables. For example, Spearman's correlation coefficient [16] is a measure of the monotonic relationship between two variables. It calculates the correlation between the ranked values of the variables rather than their actual values and it can be used for those variables that may not have

a linear relationship. Moreover, when considering data that can be mathematically represented with dynamical systems, an alternate measure of dependency could be the sensitivity of the outputs to the system's parameters. This sensitivity essentially quantifies the degree to which changes in a given parameter can impact the system's output, serving as a useful indicator of dependency or influence within the system [17].

Kendall's Tau [18] is another rank-based correlation measure that assesses the strength of the ordinal association between two variables. It is particularly useful when dealing with ranked or ordinal data and is less sensitive to outliers compared to Pearson's correlation.

The Distance correlation [19] measures the dependency between two variables based on the idea of comparing distances between points. It captures both linear and non-linear relationships and can be deployed when dealing with high-dimensional data. Information gain [20] is a measure typically used in the field of machine learning and especially with decision trees. It quantifies the reduction in entropy (uncertainty) of a target variable by knowing the values of a predictor variable. Furthermore, Maximal Information Coefficient (MIC) [21] is a measure that captures both linear and nonlinear dependencies between two variables. It quantifies the strength of the relationship by measuring the maximum amount of information shared by the variables across all possible partitions of the data.

Mutual information is more general as captures both linear and nonlinear relationships between variables. Also, it does not assume any specific form of relationship and it is able to detect any type of dependence. It is sensitive to linear and nonlinear relationships between variables and can capture complex relationships. However, it does not offer a direct interpretation in terms of the strength of the relationship although it captures any type of dependence between variables, including both linear and nonlinear dependencies, and is widely used in various fields.

Pearson's coefficient measures the linear relationship between variables and quantifies the strength and direction of the linear association. As a drawback, it may fail to detect nonlinear relationships. It has a clear interpretation as it ranges from -1 to 1, but is limited to capturing linear dependencies and may not necessarily detect more complex relationships. It is useful when analyzing linear relationships.

Spearman's correlation coefficient aims at capturing monotonic relationships, therefore including both linear and nonlinear relationships. It is able to measure the strength and especially the direction of the monotonic association between variables, although may not detect non-monotonic relationships. Very often, it is deployed to rank ordinal variables or assess the relationship between ranked data. Again, it has a clear interpretation as it ranges from -1 to 1 and is more appropriate when considering ordinal variables.

Kendall's Tau correlation measure captures monotonic relationships, although it does not consider the shape of the relationship beyond a monotonic trend. It has to be noticed that mutual information can be applied to both discrete and continuous variables, while Kendall's Tau is commonly used for ranking ordinal variables or assessing the relationship between ranked data. It is preferred when assessing the agreement or concordance between ranked data: in this sense, it is rather limited when compared to the wide usability of mutual information.

Information gain differs from mutual information in context and especially in the calculation (the former employs the joint and marginal probability distributions of the variables; the latter measures the reduction in entropy when a specific attribute is known). Furthermore, information gain is used only for discrete variables.

In conclusion, by considering table 1, it results that mutual information is particularly suitable for the problem addressed in this paper.

B. SYNTHETIC DATASET GENERATION

Synthetic dataset generation has been an area of interest due to its potential applications in privacy-preserving data analysis, data augmentation, and model validation [22], [23], [24], [25], [26]. Several methods have been proposed for generating synthetic datasets, such as sampling-based approaches, Generative Adversarial Networks (GAN) [27], and variational autoencoders [28].

Furthermore, it is possible to generate synthetic data by fitting a parametric model, such as Gaussian, to the original data and generating synthetic samples from the learned parameters, although it is assumed that the data follow a specific distribution. Using non-parametric models, such as Kernel Density Estimation (KDE) ([29], [30], [31]) or Gaussian Mixture Models (GMM, [32], [33]), to estimate the underlying probability distribution of the data. Synthetic samples are then generated by sampling from the estimated distribution. Notably, it is possible to model the sequential dependencies in the data using Markov chains [34], [35]). The transition probabilities between states are estimated from the original data, and new synthetic sequences are generated by sampling from the Markov chain.

Data augmentation techniques based on mutual information generate new data samples by applying transformations (or perturbations) to the original data while preserving the relevant information. These techniques can involve operations such as rotation, translation, scaling, or even adding noise to the data. On the other hand, GAN can generate new samples that are similar to the original data but are not exact copies. IN this sense, they are primarily used for data synthesis or generation tasks.

KDE can be used to generate new synthetic data samples by sampling from the estimated density function. These samples are drawn randomly following a distribution modeled by KDE. The generated data points are not necessarily copies of the original data although are representative of the same distribution. KDE-based data generation is typically used in scenarios where a large amount of training data is needed.

TABLE 1. Measures of dependency evaluation.

Measure of Dependency	Type of relationship captured	Sensitivity to nonlinear relationships	Dependencies captured	Formula
Mutual Information	It is more general and captures both linear and nonlinear relationships between variables. It does not assume any specific form of relationship and can detect any kind of dependence.	Mutual information is sensitive to both linear and nonlinear relationships between variables	It captures any type of dependence between variables, including both linear and nonlinear dependencies. It can detect complex relationships such as nonlinear associations, interactions, and dependencies that go beyond a simple linear association	$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right)$, where X and Y are discrete random variables, p(x,y) is the joint probability distribution of X and Y, p(x) is the marginal probability distribution of X and p(y) is the marginal probability distribution of Y.
Pearson's coefficient	It specifically measures the linear relationship between variables. It quantifies the strength and direction of the linear association	Pearson's correlation coefficient only measures the strength of the linear relationship and may fail to detect nonlinear relationships.	It measures the linear relationship between variables. It quantifies the strength and direction of the linear association	$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$ where n is the number of data points, x_i is the ith value of the first variable, y_i is the ith value of the second variable, $\bar{x}$ is the mean of the first variable, $\bar{y}$ is the mean of the second variable.
Spearman's correlation	Spearman's correlation coefficient focuses on capturing monotonic relationships, which include both linear and nonlinear relationships that exhibit a consistent increasing or decreasing trend. It measures the strength and direction of the monotonic association between variables	It is specifically designed to capture monotonic relationships and may not detect non-monotonic relationships	It captures the strength and direction of monotonic relationships between variables, including both linear and nonlinear monotonic associations. It is robust to outliers and provides a measure of the overall agreement in the ranks of the variables. However, it is limited to capturing monotonic relationships and may not capture more intricate non-monotonic relationships.	$\rho_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$ where n is the number of data points, d_i is the difference between the ranks of the two variables for the ith data point.
Kendall's Tau correlation	It specifically measures the strength and direction of the ordinal association or concordance between two ranked variables. It captures monotonic relationships but does not account for the magnitude of the differences between the ranks.	It focuses specifically on measuring the concordance or agreement in the ordering of ranks. It does not consider the shape of the relationship beyond the monotonic trend	Rank order (the degree to which the ranks of the paired observations agree), Pairwise comparisons, and ties (i.e. when multiple observations have the same value).	$\tau = \frac{\sum_{i<j}(sgn(x_i-x_j)sgn(y_i-y_j))}{n(n-1)/2}$, where n is the number of data points, $sgn(x_i - x_j)$ is the sign of the difference between the ranks of x_i and x_j, $sgn(y_i-y_j)$ is the sign of the difference between the ranks of y_i and y_j

With regard to GMM-based methods, the samples are generated by first selecting a component from the mixture according to its weight and then drawing a data point from the selected Gaussian distribution. As a result, the generated data points are representative of the modeled distribution.

Table 2 summarizes the methods discussed in this section.

In the context of bias and fairness, the authors in [7] have focused on qualitative analysis of bias in datasets using structural equation modeling and probabilistic networks, highlighting the impact of network parameters on the amount of bias between dataset attributes. The present paper builds upon this previous work, proposing a novel method for generating synthetic datasets with controlled bias and a two-step optimization algorithm for quantitatively adjusting the relationships between variables.

C. FAIRNESS IN MACHINE LEARNING

The importance of fairness in machine learning has been widely recognized, leading to the development of various techniques for ensuring fair decision-making algorithms [36]. These techniques typically fall into three categories: pre-processing, in-processing, and post-processing. Pre-processing methods focus on transforming the input data to remove or mitigate biases before training the model, for example by adjusting the dataset to ensure balanced representation across different groups or classes, generating synthetic data points to increase representation of underrepresented groups, or constructing new features that capture important information about underrepresented groups.

Finally, techniques such as Principal Component Analysis (PCA) or t-SNE to reduce bias caused by irrelevant or redundant features (see [37], where the authors use volatility as a metric to generate better, and fairer predictions; [38] for a discussion on facial attribute recognition used to denote the attribution of model bias from imbalanced training data distribution, joined to balancing data distribution achieved by adversarial examples; and finally, [39], where the author use covariate shift to assess fair decisions).

In-processing techniques modify the learning algorithm itself to ensure fair predictions. For example, re-weighting is a technique that works by assigning different weights to different instances or groups to adjust for imbalances in the dataset and is contrapposed to rescaling, i.e. modifying the predicted outcomes or scores to account for the imbalance in the data. Other approaches work on modifying decision thresholds to achieve fairness goals for different groups or

TABLE 2. A comparison of data augmentation techniques.

Data augmentation method	Strenghts	Weaknesses
Generative Adversarial Networks (GAN)	Can generate high-quality, diverse samples that capture the underlying data distribution	Training GANs is generally challenging and require complex additional tuning. The generated samples may suffer from mode collapse, where the generator produces limited variations or fails to capture the entire data distribution.
Variational Autoencoders (VAEs)	They provide a principled approach for learning latent representations and generating new data samples	They tend to produce blurry samples, especially in image generation tasks. The generated samples might not capture the true data distribution faithfully.
Kernel Density Estimation (KDE)	It can generate new samples by randomly sampling from the estimated density.	KDE can be computationally expensive, especially for high-dimensional data. The quality and diversity of generated samples heavily depend on the quality of the estimated density function, which can be very difficult to obtain accurately.
Gaussian Mixture Model (GMM)	It provides a flexible and interpretable way to model data distributions	The number of components in the GMM needs to be determined, which can be hard in practice.

even by deploying a form of adversarial learning, i.e. training a model to simultaneously predict the target variable and detect sensitive attributes to reduce bias. Finally, regularization algorithms add fairness constraints or penalties to the model's objective function to minimize discrimination (see, for example, [40] for a survey on in-processing techniques; [41] for a study on a adversarial network limiting the bias from the data perspective and the model at the same time, and finally, [42] where the authors review how to re-deploy fairness under distribution shifts. This approach represents a sufficient condition being the basis for a theory-guided self-training algorithm founded on an intra-group expansion basis).

Post-processing methods adjust the model's predictions after training to ensure fairness. This goal is usually achieved by adjusting model outputs to achieve fairness goals by mapping them to a calibrated probability scale. It is also possible to add an additional "reject" category to the classification task, allowing uncertain predictions to be flagged rather than assigned biased outcomes. Finally, a form of equalized odds postprocessing is achieved by modifying predicted labels to ensure equal false positive and false negative rates across different groups (see [43] for a discussion on MULTIACCURACY-BOOST, a rapidly converging post-processing algorithm; [44] for a study about the trade-off between the minimization of error disparity across different population groups and the calibration between probability estimates, and finally [45], where the authors consider counterfactual equalized odds and develop a post-processed predictor that is estimated though doubly robust estimators, following a previous line of research based on postprocessing techniques). Table 3 summarizes the approaches by effectiveness, suitability, and shortcomings.

The proposed approach in this paper contributes to the pre-processing category, aiming to control the amount of bias in datasets through the generation of synthetic data and quantitative adjustment of the relationships between variables. In particular, the proposed methodology in this paper builds upon and extends previous work in the fields of measures of dependency, synthetic dataset generation, and fairness in machine learning. Utilizing structural equation modeling and graphical models, the two-step optimization algorithm is devised to efficiently manage the bias between categorical attributes in datasets, guaranteeing that the desired mutual information is attained. This, in turn, contributes to the development of unbiased and fair decision-making algorithms in various applications.

III. METHODOLOGY

This section introduces the proposed methodological framework for generating synthetic datasets, which allows users to directly specify the desired level of dependencies among sample attributes. Each record in the dataset is represented as a tuple of attributes, with each attribute being a discrete random variable with ordinal values. Specifically, we consider a dataset $\mathcal{D} = \{r_i; i = 1, .., N\}$ consisting of N records, where the i-th record r_i is represented as a tuple of K attributes, i.e., $r_i = (a_{i,1}, a_{i,2}, \ldots, a_{i,K})$. These attributes are assumed to be realizations of their corresponding discrete random variables, namely $A_1, A_2, \ldots, A_K$. Statistical dependencies among different attributes are examined using structural equation modeling [3], a graph-based approach that enables researchers to investigate direct, indirect, and mediating effects between variables within a single, comprehensive model.

The primary goal of this methodology is to generate a set of records with a desired level of causal dependency (i.e., statistical bias) between selected pairs of attributes, given the marginal probabilities of different attribute values a priori.

Section III-A describes the Structural Equation Modeling approach, while Section III-B presents an adaptation of SEM for synthetic dataset generation. Lastly, Section III-C introduces the concept of mutual information which is used in this paper to quantify the amount of bias (i.e. mutual dependence) between two variables.

A. STRUCTURAL EQUATION MODELING

Structural equation modeling is a robust statistical technique commonly used to fit probabilistic models based on observed

TABLE 3. A comparison of approaches to data fairness.

Approach	Effectiveness	Suitability	Shortcomings
Pre-processing	Fairness s not necessarily guaranteed as pre-processing addresses bias in the data distribution rather than the model itself.	It can be deployed when there is a need to mitigate bias in the data before training any model	It does not necessarily capture all aspects of bias present in the dataset. Also, it could add new biases in the data manipulation process.
In-processing	Fairness is usually guaranteed, as the fairness constraints are explicitly included during the learning process.	Used in particular where direct control over the fairness of the learning process is necessary	Modifying the learning algorithm might lead to increased computational overhead. Furthermore, the fairness constraints could degrade model performance.
Post-processing	Fairness is assured up to some extent, as it ensures the model's outputs match specific fairness criteria	It may be considered when there is a trained model that exhibits bias, and fairness is requested without retraining the model	Potentially, it may be less effective than in-processing methods in achieving fairness. Additionally, adjustments to the model's outputs could result in suboptimal performance

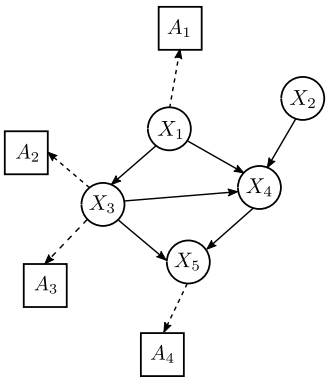


FIGURE 2. A representative example of a structural equation model (SEM).

measurements. In this context, the measurements represent the outcomes of observed variables that describe the phenomenon under investigation, while their dependencies are modeled using a graph of latent Gaussian random variables. In the analysis of datasets, the observed variables are discrete random variables corresponding to the sample attributes $A_1, A_2, \ldots, A_K$. Their dependencies are represented by a probabilistic network comprising M latent continuous random variables, namely, $X_1, X_2, \ldots, X_M$.

Figure 2 illustrates an example of an SEM, where the squared nodes labeled A_i; $i = 1, \ldots, 4$ represent the manifest variables, and the rounded nodes, X_i; $i = 1, \ldots, 5$, denote the hidden variables. The figure also displays the causal dependencies between pairs of latent variables (solid arcs) and the connections between latent and manifest variables (dotted arcs). Nodes without incoming edges (such as X_1 and X_2) represent independent latent variables, while nodes with incoming edges (such as X_3, X_4, and X_5) represent dependent latent variables.

In general, a SEM with M latent variables is described as a weighted Directed Acyclic Graph (DAG) [46] with M nodes, where the i-th node corresponds to the latent variable X_i. A DAG is a collection of nodes (or vertices) connected by directed edges (or arcs), where each edge has an initial node and a terminal node. The term acyclic indicates that there is no sequence of edges that forms a closed loop or cycle, i.e. a path starting and ending at the same node. The edges between nodes are described as $(i, j) \in E$ where $E \subseteq \{(i, j); 1 \leq i, j \leq M\}$ is the set of DAG edges. Within the context of SEM, such edges signify the direct relationships between the corresponding latent variables, which may arise from factors such as causality or correlation. Moreover, each $(i, j) \in E$ is associated with a weight $\alpha_{i,j} \in \mathbb{R}$ representing the regression coefficient (i.e., the the amount of direct dependency) between the variables X_i and X_j.

The variable X_i related to a node without incoming edges (such as X_1 and X_2 in Figure 2) is described by an independent Gaussian variable with probability distribution $\mathcal{N}(0, \sigma_i^2)$, where σ_i accounts for the intrinsic variability of X_i. On the other hand, a node X_i with incoming edges (e.g., nodes X_3, X_4 and X_5 in Figure 2) corresponds to a dependent latent variable that can be described as linear combinations of the parent nodes and a Normal probability distribution that accounts for its uncertainty, that is:

$$X_i = \sum_{(j,i)\in E} \alpha_{j,i} X_j + \mathcal{N}(0, \sigma_i^2) \tag{1}$$

where the regression coefficient $\alpha_{i,j}$ is set to zero if the j-th node does not influence the i-th one. In this way, the structure of the graph and these coefficients accurately depict the dependencies and relationships between latent variables.

The previous equation establishes that each dependent latent variable, X_i, is normally distributed. This is a consequence of the fact that a sum of normally distributed random variables itself follows a normal distribution. For each dependent variable X_i, the mean of this distribution is zero. The variance of the distribution, however, includes both the intrinsic variability of the latent variable σ_i^2 and the accumulated influences from all the parent variables. Each parent influence is calculated as the square of the respective regression coefficient $\alpha_{j,i}$, multiplied by the parent variable's variance σ_j^2. The summation of all these influences constitutes the overall effect of the parent variables on the variance of X_i. In particular, it holds that:

$$X_i \sim \mathcal{N}(0, \sigma_i^2 + \sum_{(j,i)\in E} \alpha_{j,i}^2 \sigma_j^2) \tag{2}$$

where the obtained variance accurately captures both the internal variability of the variable and the contribution from its dependencies.

Given the definitions and relationships previously outlined, the entire vector $\mathbf{X} = [X_1, X_2, \cdots, X_M]$ of latent variables is described as a multivariate Normal distribution. The mean of this multivariate distribution is a vector of zeros, and the variance-covariance structure is given by a matrix $\mathbf{\Sigma} \in \mathbb{R}^{M \times M}$, as denoted by:

$$\mathbf{X} \sim \mathcal{N}(0, \mathbf{\Sigma}) \tag{3}$$

The matrix $\mathbf{\Sigma}$, known as the variance-covariance matrix, characterizes the variance of each latent variable (along the diagonal of the matrix) and the covariance between each pair of latent variables (off-diagonal elements of the matrix). The computation of this matrix is achieved by the expression:

$$\mathbf{\Sigma} = \mathbf{Q}^\top \mathbf{\Sigma_n} \mathbf{Q} \tag{4}$$

where $\mathbf{\Sigma_n}$ is the diagonal matrix with the elements of the vector $[\sigma_1, \sigma_2, \ldots, \sigma_M]$ on the principal diagonal, and $\mathbf{Q} = (\mathbf{I} - \mathbf{A})^{-1}$, with $\mathbf{A}$ being the adjacency matrix of the considered DAG. Note that it is possible to obtain the Pearson correlation coefficient $\rho_{i,j} \in [-1, 1]$ between the latent variables X_i and X_j from the variance-covariance matrix as follows:

$$\rho_{i,j} = \frac{\Sigma_{i,j}}{\sqrt{\Sigma_{i,i}}\sqrt{\Sigma_{j,j}}} \tag{5}$$

where $\Sigma_{i,j}$ is the j-th elements of the i-th row of the variance-covariance matrix $\mathbf{\Sigma}$. Note that $\rho_{i,j} = \rho_{j,i}$ since the variance-covariance matrix is symmetric. In order to simplify the following discussion, it is also useful to express the covariance matrix $\mathbf{\Sigma}_{[i,j]} \in \mathbb{R}^{2\times 2}$ between the two variables X_i and X_j as a function of their variances and their correlation coefficient:

$$\mathbf{\Sigma}_{[i,j]} = \begin{bmatrix} \Sigma_{i,i} & \rho_{i,j}\sqrt{\Sigma_{i,i}}\sqrt{\Sigma_{j,j}} \\ \rho_{i,j}\sqrt{\Sigma_{i,i}}\sqrt{\Sigma_{j,j}} & \Sigma_{j,j} \end{bmatrix} \tag{6}$$

considering that $\rho_{i,j} = 1$ corresponds to a singular value matrix.

Building upon the previous discussion, it is important to highlight the integral role that the selection of latent variables plays in the effective application of the proposed model. These latent variables capture complex and often non-linear interactions between the categorical attributes of the dataset, interactions that might be challenging to model directly. Therefore, the judicious selection of these variables is critical for accurately representing the attribute dependencies and, consequently, for controlling bias effectively. In this context, domain knowledge assumes a paramount role. A comprehensive understanding of the dataset and the potential interactions between its attributes can guide the selection of latent variables. Knowing which attributes might have significant dependencies can help identify the dimensions that capture the inherent biases in the data. Further, statistical techniques such as exploratory factor analysis can complement domain knowledge and provide valuable guidance in the selection process. These methods can help uncover the underlying structure of the data and suggest potential latent variables that might not be immediately apparent.

In summary, while the proposed model's sensitivity to the choice of latent variables underlines their importance, a thorough understanding of the domain and the use of suitable statistical techniques can ensure their optimal selection, thereby enhancing the model's effectiveness in bias control.

B. DATASET GENERATION

The multivariate distribution expressed in (3) can be employed to generate a synthetic dataset where the causal relationships between the sample attributes are depicted by the graphical model introduced in III-A. Specifically, the latent representation of each sample in the dataset is an M-dimensional vector, which is derived by sampling the aforementioned multivariate normal distribution. The observed variables (i.e. the attributes), assumed to be categorical in this context, are subsequently obtained from their latent representation through the application of appropriately selected thresholds, referred to as cutoffs. These cutoffs are determined uniquely, provided that the marginal distributions of the various attributes are assumed to be given. This implies that the probability $\mathcal{P}(A_i = \tilde{A}_i^k)$ is known for each attribute A_i, with $\tilde{A}_i^k$ representing its admissible categorical values for $k = 1, \ldots, n_i$.

The cutoffs $c_{i,1}, \ldots, c_{i,n_i+1}$ are used to partition the cumulative probability function $F : \mathbb{R} \to [0, 1]$ of the i-th latent variable X_i into n_i regions each of them associated with a single categorical class, in order to satisfy the following equation:

$$\begin{aligned} \mathcal{P}(A_i = \tilde{A}_i^k) &= \mathcal{P}(c_{i,k} \leq X_i < c_{i,k+1}) \\ &= F(c_{i,k+1}) - F(c_{i,k}) \end{aligned} \tag{7}$$

for $i = 1, \ldots, M$ and $k = 1, \ldots, n_i$. Note that it is assumed that the lower bound $c_{i,1}$ represents negative infinity while the upper bound c_{i,n_i+1} represents positive infinity. An example of cutoffs setting is depicted in Figure 3, where the probability density function $f(X_i)$ of a generic latent variable X_i is partitioned into four different regions, whose area corresponds to the probabilities $\mathcal{P}(A_i = \tilde{A}_i^k)$ for $i = 1, \ldots, 4$.

A sample in the dataset is generated by transforming a randomly drawn vector from the multivariate Gaussian distribution into a vector of categorical features based on the region each entry of the continuous vector belongs to. Consequently, a dataset composed of categorical variables is obtained. It is important to note that the proposed formulation is limited to the case of binary variables or ordinal categorical variables A_i with $n_i > 2$, as the latent

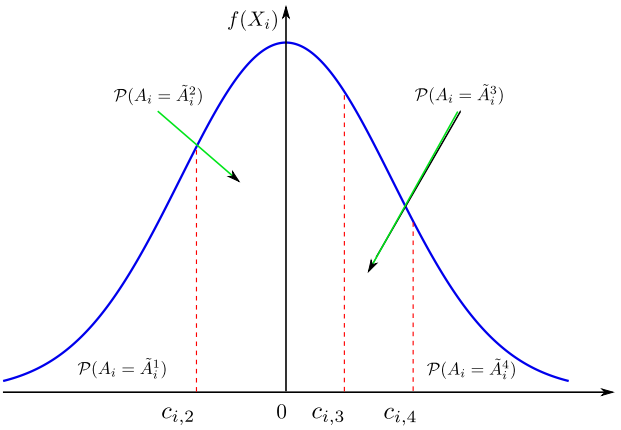


FIGURE 3. An illustration of a cutoffs setting for a generic latent variable X_i, with the area under its probability density function $f(X_i)$ partitioned into four distinct regions.

continuous representation inherently provides an ordering for the variables under consideration.

C. MEASUREMENT OF BIAS

In the preceding sections, we have outlined the primary objective of this research article, which is to effectively choose the network parameters ($\alpha_{i,j}$ and σ_i) introduced in III-A to facilitate the generation of synthetic datasets containing a predefined amount of dependencies between sample attributes. To achieve this, it is essential to employ a quantitative measure to assess these dependencies. Of the various available options, mutual information has been chosen as an appropriate candidate for this purpose due to its ability to capture both linear and non-linear relationships between variables, making it a versatile and powerful measure.

However, before delving into the details of mutual information, it is important to articulate a more formal definition of 'bias'. In the context of this study, 'bias' is characterized as any systematic inaccuracy that misrepresents or manipulates certain categories of attributes in our dataset. This could materialize as statistical overrepresentation or underrepresentation of these categories, consequently leading to unfair results in subsequent data analysis or application. A practical example of this concept can be seen in 'demographic parity', a widely accepted fairness metric in machine learning. This principle states that the rate of positive outcomes should remain consistent across all protected groups (such as different races, genders, and ages) in binary classification tasks. Mathematically, demographic parity can be expressed as follows. Let Y be the decision (0 for negative outcome, 1 for positive outcome) and A be the protected attribute (e.g., gender or race). We say that a decision-making process satisfies demographic parity if $P(Y = 1|A = a) = P(Y = 1)$ for all values 'a' of the attribute A. Yet, while this metric is simple and interpretable, it does not accommodate potential legitimate dependencies between the protected attribute and the decision, which could lead to unrealistic fairness expectations in some cases.

In the subsequent sections, an exploration of the critical properties of mutual information will be undertaken, highlighting its role in mitigating bias and advancing fairness in AI applications. Specifically, mutual information is utilized to offer a quantifiable measure of bias, accomplished by calculating the statistical dependencies between pairs of categorical attributes both before and after the implementation of the proposed method. This provides a tangible measure of bias mitigation. However, it is essential to underline that mutual information inherently measures dependency, not bias per se. Consequently, to interpret mutual information accurately as a measure of bias, it is vital to verify that the dependencies measured are indeed indicative of bias, a determination that is typically context-dependent and necessitates expert domain knowledge.

In the realms of probability and information theory, mutual information serves as a metric to measure the degree of mutual dependence between a pair of random variables. This concept is closely related to the entropy of a random variable, a fundamental idea in information theory that quantifies the expected amount of information contained within a random variable.

The mutual information between two categorical random variables, A_i and A_j, is defined as:

$$I(A_i, A_j) = \sum_{A_i, A_j} P(A_i, A_j) \log \left(\frac{P(A_i, A_j)}{P(A_i)P(A_j)} \right) \tag{8}$$

where $P(A_i)$ denotes the marginal probability of variable A_i, and $P(A_i, A_j)$ represents the joint probability of the two variables. Additionally, the following relationship holds true:

$$I(A_i, A_j) = H(A_i) + H(A_j) - H(A_i, A_j) \tag{9}$$

where

$$H(A_k) = -\sum_{A_k} P(A_k) \log \left(P(A_k)\right), \quad k = i, j \tag{10}$$

$$H(A_i, A_j) = -\sum_{A_i, A_j} P(A_i, A_j) \log \left(P(A_i, A_j)\right) \tag{11}$$

and $H(A_i)$ and $H(A_j)$ represent the marginal entropies for the variables A_i and A_j, respectively, while $H(A_i, A_j)$ corresponds to their joint entropy.

In the following, a discussion is conducted in order to highlight the relationship between the network parameters ($\alpha_{i,j}$ and σ_i) and the mutual information between two observed variables A_i and A_j, exploiting the aforementioned definition. While the marginal probability of a feature A_i is assumed to be given since it corresponds to $\mathcal{P}(A_i = \tilde{A}_i^k)$, the joint probability mass function $\mathcal{P}(A_i, A_j)$ can be represented by a table with $n_i n_j$ entrances, each of which is computed by integrating over specific rectangles (defined according to the cutoffs) the following probability density

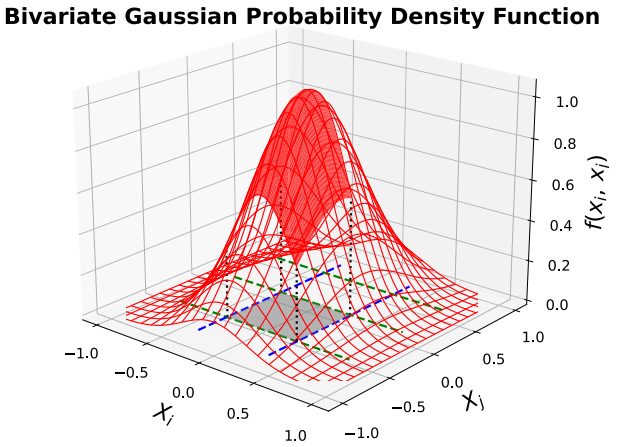


FIGURE 4. Graphical representation of a bivariate Gaussian probability density function $f(X_i, X_j)$, where variables X_i and X_j have 2 and 3 levels respectively. Cutoffs are shown with green and blue dashed lines. A specific rectangle $\mathcal{R}_{i,j}^{2,2}$ is highlighted in gray on the plane $f(X_i, X_j) = 0$. Note that the volume under the red surface within this rectangle represents the probability $\mathcal{P}((X_i, X_j) \in \mathcal{R}_{i,j}^{2,2})$.

function $f(X_i, X_j)$:

$$f(X_i, X_j) = \frac{1}{2\pi\sqrt{det(\boldsymbol{\Sigma}_{[i,j]})}} \exp\left(-\frac{1}{2}[X_i\, X_j]\, \boldsymbol{\Sigma}_{[i,j]}\, [X_i\, X_j]^\top\right) \tag{12}$$

which corresponds to the zero mean bi-variate Gaussian associated with the latent variables X_i and X_j. Specifically, the probability that A_i and A_j assume two specific values $\tilde{A}_i^h$ and $\tilde{A}_j^k$, with $h \in \{1, \ldots, n_i\}$ and $k \in \{1, \ldots, n_j\}$, corresponds to the probability $\mathcal{P}((X_i, X_j) \in \mathcal{R}_{i,j}^{h,k})$ that a pair (X_i, X_j), extracted from the latent bi-variate Gaussian with probability density function in (12), lies in the rectangle $\mathcal{R}_{i,j}^{h,k}$ (see Figure 4, where an example of such a rectangle is depicted in gray) defined as:

$$\mathcal{R}_{i,j}^{h,k} = \{(X_i, X_j) : c_{i,h-1} \leq X_i < c_{i,h},\ c_{j,k-1} \leq X_j < c_{j,k}\} \tag{13}$$

However, a closed form expression for the probability $\mathcal{P}((X_i, X_j) \in \mathcal{R}_{i,j}^{h,k})$ does not exist, except for the trivial case in which $n_i = n_j = 2$, and therefore it should be properly approximated as described in the following.

Specifically, an approximation can be obtained by computing an estimation $\hat{p}$ of the success probability p of the following Bernoulli variable $B \sim \mathcal{B}(1, p)$ over N samples:

$$P\left((X, Y) \in \mathcal{R}_{i,j}^{h,k}\right) = p \tag{14}$$

$$P\left((X, Y) \notin \mathcal{R}_{i,j}^{h,k}\right) = 1 - p \tag{15}$$

where N can be chosen in order to obtain a confidence interval for the estimation with a desired guaranteed probability. For instance, 10^6 samples guarantee a level of precision $\psi = 1.5 \cdot 10^{-3}$ with confidence level of 99.7%, i.e. $\mathcal{P}(\hat{p} - \psi < p < \hat{p} + \psi) = 0.997$.

Finally, it is interesting to show that, when the marginal distributions $\mathcal{P}(A_i)$ and $\mathcal{P}(A_j)$ are given, the mutual information $I(A_i, A_j)$ between two categorical variables A_i and A_j depends only on the correlation coefficient $\rho_{i,j}$ between the two corresponding latent variables X_i and X_j, according to the function $\phi_{i,j}(\rho_{i,j})$, where $\phi_{i,j} : [0, 1] \to \mathbb{R}^+$. In fact, as can be seen from (8), the mutual information depends only on the marginal probabilities $\mathcal{P}(A_i)$ and $P(A_j)$ and the joint probability $\mathcal{P}(A_i, A_j)$. While the former are assumed to be given, the latter depends on the matrix $\boldsymbol{\Sigma}_{[i,j]}$ in (6), which in turn depends on the following variables: $\Sigma_{i,i}$, $\Sigma_{j,j}$ and $\rho_{i,j}$. Note that the cut-off values for a variable A_i are uniquely set according to the marginal distributions and depend linearly on the standard deviation $\sqrt{\Sigma_{i,i}}$ of its latent representation, i.e. $c_{i,q} = \sqrt{\Sigma_{i,i}}\bar{c}_{i,q}$ where $\bar{c}_{i,q}$, $q = 1, \ldots, n_i + 1$ are the cutoffs for a Gaussian variable with unitary standard deviation and same marginal probability. Considering the following latent variables $X_i \sim \mathcal{N}(0, \Sigma_{i,i})$ and $X_j \sim \mathcal{N}(0, \Sigma_{j,j})$, with

$$[X_i\, X_j]^\top \sim \mathcal{N}\left([0,\ 0]^\top, \begin{bmatrix} \Sigma_{i,i} & \rho_{i,j}\sqrt{\Sigma_{i,i}\Sigma_{j,j}} \\ \rho_{i,j}\sqrt{\Sigma_{i,i}\Sigma_{j,j}} & \Sigma_{j,j} \end{bmatrix}\right) \tag{16}$$

the idea is to show that the joint probability $\mathcal{P}(A_i = \tilde{A}_i^h, A_j = \tilde{A}_j^k) = \mathcal{P}((X_i, X_j) \in \mathcal{R}_{i,j}^{h,k})$ does not depend on $\Sigma_{i,i}$ and $\Sigma_{j,j}$. Let consider Z_i, $Z_j \sim \mathcal{N}(0, 1)$ such that $X_i = \sqrt{\Sigma_{i,i}}Z_i$ and $X_j = \sqrt{\Sigma_{j,j}}Z_j$, it holds that

$$\begin{aligned} &\mathcal{P}((X_i, X_j) \in \mathcal{R}_{i,j}^{h,k}) \\ &\quad = \mathcal{P}(c_{h-1,i} < X_i < c_{h,i},\ c_{k-1,j} < X_j < c_{h,j}) \\ &\quad = \mathcal{P}(\bar{c}_{h-1,i} < Z_i < \bar{c}_{h,i},\ \bar{c}_{k-1,j} < Z_j < \bar{c}_{h,j}) \end{aligned} \tag{17}$$

with

$$[Z_i\, Z_j]^\top \sim \mathcal{N}\left([0,\ 0]^\top, \begin{bmatrix} 1 & \rho_{i,j} \\ \rho_{i,j} & 1 \end{bmatrix}\right) \tag{18}$$

thus showing that $\mathcal{P}(A_i = \tilde{A}_i^h, A_j = \tilde{A}_j^k)$ depends only on the correlation coefficient $\rho_{i,j}$, when the marginal distributions (and hence the cutoffs) are set a priori.

D. TWO-STEP OPTIMIZATION

This section outlines a method for calculating the regression coefficients ($\alpha_{i,j}$) and conditional variances (σ_i) in a network such as the one depicted in Figure 2, targeting a specific mutual information for designated variable pairs. In particular, we propose a two-step optimization process that exploits in sequence a gradient-less optimization algorithm and a gradient-based one. In detail, the first method targets the variance-covariance matrix to determine conditions, using Pearson's correlation coefficients, which align with the target mutual information. The latter is instead used to compute the regression coefficients and conditional variances in the network that lead to a variance-covariance matrix such that the aforementioned properties are satisfied.

The goal for the network parameters is to align them with a reference set of mutual information values, $I^r_{i,j}$, between particular variable pairs A_i and A_j, with $(i,j) \in \mathcal{B} \subseteq \mathcal{M} \times \mathcal{M}$, with $\mathcal{M} = \{1, \ldots, K\}$, where $\mathcal{B}$ is the set of pairs for which there is an interest in controlling the bias. Such a task can be formulated as an optimization problem whose solution corresponds to the optimal configuration of regression coefficients $\alpha^\star_{i,j}$ and conditional variances $\sigma^{2,\star}_i$ that provides the values of mutual information closest to the references, i.e.

$$\alpha^\star_{i,j},\ \sigma^{2,\star}_i = \underset{\substack{\alpha_{i,j},\ i\neq j \\ \sigma^2_i,\ i\in\mathcal{M}}}{\arg\min} \sum_{(i,j)\in\mathcal{B}} \Big(I(A_i,\ A_j) - I^r_{i,j}\Big)^2 \tag{19a}$$

$$\text{subject to:} \quad \alpha_{i,j} \in \mathbb{R},\ \sigma^2_i \in \mathbb{R}^+ \tag{19b}$$

$$\alpha_{i,j} = \bar{\alpha}_{i,j}, \quad (i,\ j) \in \mathcal{K}_\alpha \subseteq \mathcal{M} \times \mathcal{M} \tag{19c}$$

$$\sigma^2_i = \bar{\sigma}^2_i, \quad i \in \mathcal{K}_{\sigma^2} \subseteq \mathcal{M} \tag{19d}$$

where the sets $\mathcal{K}_\alpha$ and $\mathcal{K}_{\sigma^2}$ indicate the edges and the conditional variances that are assumed to be fixed and hence not controllable. For instance, the constraint $\alpha_{h,k} = 0$ is used to model the fact that no edge exists between the h-th variable and the k-th one. Specifically, the design procedure of the sets $\mathcal{K}_\alpha$ and $\mathcal{K}_{\sigma^2}$ is carried out by domain experts to incorporate prior knowledge on the structure of the network.

Given the fact that an exact analytical expression for $\phi_{i,j}(\rho_{i,j})$ is not readily obtainable, gradient-based methods cannot tackle the optimization problem in (19). Therefore, gradient-less methodologies need to be employed, see for instance the *Nelder-Mead* algorithm [47], which relies on Montecarlo simulations, or any sort of genetic algorithms. However, such an approach is not efficient in the sense that it does not exploit the knowledge of the analytical model which relates the optimization variables with the Person's coefficients. Moreover, the effectiveness of a gradient-less algorithm in finding an optimal solution drops significantly when the number of optimization variables increases, due to the fact that it can become trapped in local minima or it may struggle to efficiently explore the vast, multi-dimensional search space. Furthermore, in (19), the optimization variables appear coupled in the cost function and can not be optimized separately. For all these reasons, as a more efficient alternative, an optimization framework is here proposed which is composed of two consecutive steps. In particular, *step 1* consists in the following optimization problem:

$$\rho^\star_{i,j} = \underset{\rho_{i,j}}{\arg\min} \Big(I(A_i,\ A_j) - I^r_{i,j}\Big)^2, \quad (i,\ j) \in \mathcal{B} \tag{20a}$$

$$\text{subject to:} \quad \rho_{i,j} \in [-1,\ 1] \tag{20b}$$

while *step 2* is defined as:

$$\alpha^\star_{i,j},\ \sigma^{2,\star}_i = \underset{\substack{\alpha_{i,j},\ i\neq j \\ \sigma^2_i,\ i\in\mathcal{M}}}{\arg\min} \sum_{(i,j)\in\mathcal{B}} \Big(\rho_{i,j} - \rho^\star_{i,j}\Big)^2 \tag{21a}$$

$$\text{subject to:} \quad \alpha_{i,j} \in \mathbb{R},\ \sigma^2_i \in \mathbb{R}^+ \tag{21b}$$

$$\alpha_{i,j} = \bar{\alpha}_{i,j}, \quad (i,\ j) \in \mathcal{K}_\alpha \subseteq \mathcal{M} \times \mathcal{M} \tag{21c}$$

$$\sigma^2_i = \bar{\sigma}^2_i, \quad i \in \mathcal{K}_{\sigma^2} \subseteq \mathcal{M} \tag{21d}$$

where the optimization problem in (20) can be solved through a gradient-less algorithm for each variable $\rho_{i,j}$, $(i,\ j) \in \mathcal{B}$ independently, while the problem in (21) can be solved through any gradient-based methods (see for instance the interior-point method), since an analytical expression for $\rho_{i,j}$ as a function of $\alpha_{i,j}$ and σ^2_i exists (see equations (4) and(5)). The pseudocode to implement the proposed two-step optimization method is provided in 1, which can be further explained by the following main steps:

- **Step 1:** For each pair of attributes (i,j) within the set of attributes of interest $\mathcal{B}$:
  - Compute the Pearson's coefficient $\rho^\star_{i,j}$ that minimizes the squared difference between the mutual information $I(A_i, A_j)$ and the desired mutual information $I^r_{i,j}$.
  - This is subject to the constraint that $\rho_{i,j}$ must lie within the interval $[-1, 1]$.
- **Step 2:** For each pair of attributes (i,j) within the set of attributes of interest $\mathcal{B}$:
  - Compute the weights $\alpha^\star_{i,j}$ and variances $\sigma^{2,\star}_i$ that minimize the squared difference between the actual Pearson's coefficient $\rho_{i,j}$ and the required coefficient $\rho^\star_{i,j}$ obtained in Step 1.
  - This computation is subject to the following constraints:
    - The weights $\alpha_{i,j}$ are real numbers and the variances σ^2_i are non-negative real numbers.
    - For a specific subset $\mathcal{K}_\alpha$ of $\mathcal{M} \times \mathcal{M}$, the weights $\alpha_{i,j}$ should be equal to predefined weights $\bar{\alpha}_{i,j}$.
    - For a specific subset $\mathcal{K}_{\sigma^2}$ of $\mathcal{M}$, the variances σ^2_i should be equal to predefined variances $\bar{\sigma}^2_i$.

Additionally, the main elements of the two-step optimization algorithm proposed in this paper are visually depicted in Figure 5.

Algorithm 1 Two-Step Optimization for Bias Control

for each attribute pair $(i,j) \in \mathcal{B}$ **do**
  Compute $\rho^\star_{i,j} = \arg\min_{\rho_{i,j}} (I(A_i, A_j) - I^r_{i,j})^2$
  subject to $\rho_{i,j} \in [-1, 1]$
end for
for each attribute pair $(i,j) \in \mathcal{B}$ **do**
  Compute $\alpha^\star_{i,j}, \sigma^{2,\star}_i = \arg\min_{\substack{\alpha_{i,j},i\neq j \\ \sigma^2_i,i\in\mathcal{M}}} \sum_{(i,j)\in\mathcal{B}} (\rho_{i,j} - \rho^\star_{i,j})^2$
  subject to
    $\alpha_{i,j} \in \mathbb{R}, \sigma^2_i \in \mathbb{R}^+$
    $\alpha_{i,j} = \bar{\alpha}_{i,j}, (i,j) \in \mathcal{K}_\alpha \subseteq \mathcal{M} \times \mathcal{M}$
    $\sigma^2_i = \bar{\sigma}^2_i, i \in \mathcal{K}_{\sigma^2} \subseteq \mathcal{M}$
end for

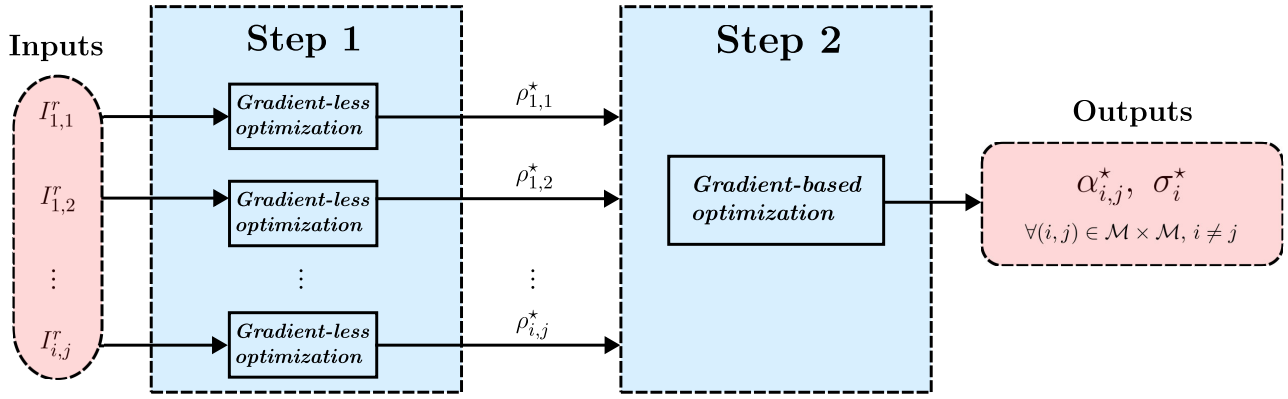


FIGURE 5. Diagram showing the two-step optimization for calculating regression coefficients and conditional variances, ensuring desired mutual information between specific variable pairs. The process utilizes both gradient-less and gradient-based methods.

It should be noted that both the approaches can retrieve the set of weights and variances that match the desired mutual information values only if the two following conditions hold:

- the target mutual information values are positive and lower than the maximum mutual information that can be achieved between the considered discrete variables;
- the target mutual information values can be achieved by acting only on the conditional variances $\sigma_i^{2,\star} \in \mathcal{M}$ and the weights $\alpha_{i,j}$, $i \neq j$ under the network constraints in (19c) and (19d). In fact, according to the network's degree of freedom, it may happen that the Pearson's coefficients between two or more pairs of variables are dependent on each other.

For the two-step algorithm, *step 1* solution from (20) yields $\rho^\star_{i,j} \in [-1, 1]$ values aligning closely with the target mutual information. In contrast, *step 2* from (21) identifies the ideal weights and variances, aiming for the smallest error between the real and desired Pearson's coefficients.

While the proposed optimization approach has its merits, it is prudent to consider also its limitations. The process employs a combination of gradient-less and gradient-based methods, each presenting its own set of advantages and potential pitfalls. The first step of the process harnesses a gradient-less method, which allows for the independent optimization of each variable. This aspect effectively addresses the challenge of coupled variables. However, even with such independence, these methods are still susceptible to local optima, which could lead to sub-optimal results. The process evolves in its second step to utilize a gradient-based method. Despite its usual efficiency, the convergence success of this method is largely contingent on the choice of initial parameters. Furthermore, the global optimality of the solution is not always guaranteed even in this case.

Beyond these considerations inherent to the optimization methods, an additional complexity emerges from the potential ripple effects of adjustments. While the objective is to control mutual information between specific attribute pairs, such adjustments may inadvertently impact mutual information between other attribute pairs. To counteract this, one might consider integrating additional constraints to preserve mutual information for unaffected pairs. However, this inclusion could reduce the degrees of freedom available for bias control, potentially leading to a situation where no feasible solution can be found. This inherent balancing act when controlling bias in intricate datasets is not exclusive to the proposed two-step optimization method but a general challenge in the field.

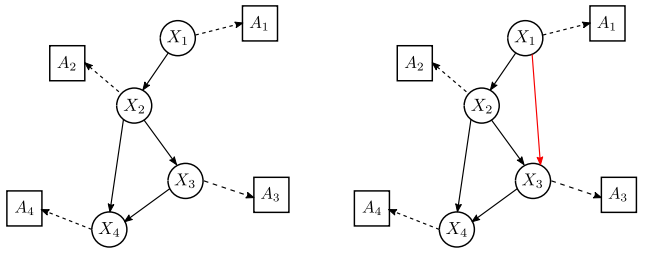
1) SCALABILITY AND COMPUTATIONAL COMPLEXITY

A critical aspect of the proposed methodology relates to its scalability and computational complexity. Specifically, the number of records in the dataset primarily influences the fitting of the SEM network rather than the proposed optimization process. Although this stage does not constitute an explicit part of the two-step optimization algorithm, it may serve as a prerequisite for identifying the interrelations amongst different dataset attributes. As the volume of data to be processed intensifies with increasing record numbers, the SEM fitting process's computational complexity naturally escalates.

The computational complexity of the two-step optimization algorithm itself, however, varies between the gradient-less and gradient-based stages. Each of these stages exhibits a unique sensitivity to the structure of attribute interrelations and the size of the corresponding graph.

In the gradient-less optimization stage, computational complexity is essentially driven by the number of attribute pairs for which control of mutual information is intended. Given the decoupled nature of these attribute pairs in this stage, the computational complexity exhibits linear scaling with the number of such pairs. This characteristic renders the gradient-less stage relatively efficient, particularly when the count of pairs of interest remains moderate.

Contrastingly, the computational complexity of the gradient-based optimization stage heavily relies on the quantity and intricacy of the elements (nodes and edges) within the graph representing attribute relationships. This dependency resonates with the computational complexity



(a) Network with $\alpha_{1,3} = 0$. (b) Network with $\alpha_{1,3} \neq 0$.

FIGURE 6. Diagram of the two networks considered as a case study. The networks differ as the left one lacks the red edge between the variables X_1 and X_3, while the right one includes it.

inherent to the interior-point method when employed in a nonlinear optimization problem. Specifically, the complexity scales roughly cubically with respect to the number of optimization variables, which correlates with the graph's nodes and edges count. Thus, larger and more intricate graphs may entail considerably high computational complexity for the gradient-based optimization stage.

Given these considerations, the proposed approach exhibits scalability for datasets with a moderate attribute count. For larger, more complex datasets, additional steps such as attribute selection or dimensionality reduction may need to be integrated to enhance efficiency.

IV. CASE STUDY

This section validates the proposed optimization algorithm in two distinct contexts. In Subsection IV-A, the performance of the methodology is examined through a numerical case study. Here, the algorithm's efficacy in identifying network parameters that yield desired mutual information among a selected subset of dataset attributes is demonstrated. Conversely, Subsection IV-B turns to a more realistic dataset, for which the corresponding graphical model is assumed to be known, where an inherent bias exists between a sensitive attribute and the target variable. In this context, the proposed methodology is employed to adjust the parameters of the Bayesian network to mitigate the undesirable bias while maintaining the relationships among all other variables. Consequently, this allows for the synthetic generation of a dataset with reduced bias, closely mimicking the original one in all other aspects.

A. VALIDATION OF THE OPTIMIZATION ALGORITHM THROUGH A NUMERICAL CASE STUDY

This subsection embarks on a detailed investigation of the proposed two-step optimization algorithm within a representative numerical context. The principal focus of this analysis is to illuminate the capability of the algorithm to fine-tune network parameters that result in desired mutual information values between defined pairs of attributes in the dataset. Through this numerical exploration, not only are the algorithm's strengths emphasized, but its limitations are also candidly addressed. As an example, it is underlined that the ability to set the desired mutual information is intrinsically related to the degree of freedom inherent to the topological structure of the given graphical model.

Specifically, the networks shown in Figure 6 are examined, each consisting of 4 nodes, that is, $\mathcal{M} = \{1, \ldots, 4\}$. The two networks differ in that the edge between variable X_1 and X_3 is assumed to be absent in Figure 6a, i.e. $\alpha_{1,3} = 0$ (which means that $(1,\ 3) \in \mathcal{K}\alpha$), while in Figure 6b the weight $\alpha_{1,3}$ can take any real value. The four categorical features are characterized by the following number of admissible values: $n_1 = 3$, $n_2 = 2$, $n_3 = 3$, and $n_4 = 2$. Additionally, for simplicity and without loss of generality, the different admissible values for the categorical features are assumed to have equal probability, meaning the marginal distribution is as follows:

$$P(A_i = \tilde{A}_i^k) = \frac{1}{n_i}, \quad k = 1, \ldots, n_i, \quad i \in \mathcal{M}. \tag{22}$$

Furthermore, in this example, the set of pairs for which the reference mutual information needs to be tracked is considered as:

$$\mathcal{B} = \{(1, 4),\ (2, 4)\}. \tag{23}$$

which corresponds to requiring the satisfaction of a specific value of mutual information $I^r_{1,4}$ between the attributes A_1 and A_4 as well as $I^r_{2,4}$ between A_2 and A_4.

In the subsequent analysis, a variety of mutual information reference combinations $I^r_{i,j}$, with $(i, j) \in \mathcal{B}$, are considered. For each combination, the two-step optimization problem given by (20)-(21) is solved. The objective is to evaluate the proposed approach's ability to effectively design network parameters under varying requirement conditions.

Specifically, a mesh is considered, defined by $I^r_{1,4}(h) = sh$ and $I^r_{2,4}(k) = sk$, with $s = 0.03$ and $h, k \in \{0,\ 1,\ \ldots,\ 30\}$. This results in a total of 961 configurations. The two-step optimization problem is solved for each configuration and for both networks depicted in Figure 6. The root mean square error between the achieved mutual information ($I^\star_{1,4}(h)$ and $I^\star_{2,4}(k)$) and the reference one is computed using the following formula:

$$e(h, k) = \frac{1}{\sqrt{2}}\sqrt{\left(I^r_{1,4}(h) - I^\star_{1,4}(h)\right)^2 + \left(I^r_{2,4}(k) - I^\star_{2,4}(k)\right)^2} \tag{24}$$

Figure 7 presents heatmaps illustrating the achieved error values for different target configurations for both the considered networks. Specifically, the left portion of Figure 7 showcases the results pertaining to the network depicted in Figure 6a, whereas the right section of Figure 5 corresponds to the results derived from the network illustrated in Figure 6b.

First of all, it is important to highlight that the mutual information between two categorical variables A_i and A_j has an upper bound $I^{\max}_{i,j}$, which depends, among other things, on the number of admissible values n_i and n_j for the two considered variables. This upper bound can be obtained by evaluating (8) with the Pearson's coefficient approaching its maximum absolute value (i.e., $|\rho_{i,j}| = 1$). Specifically,

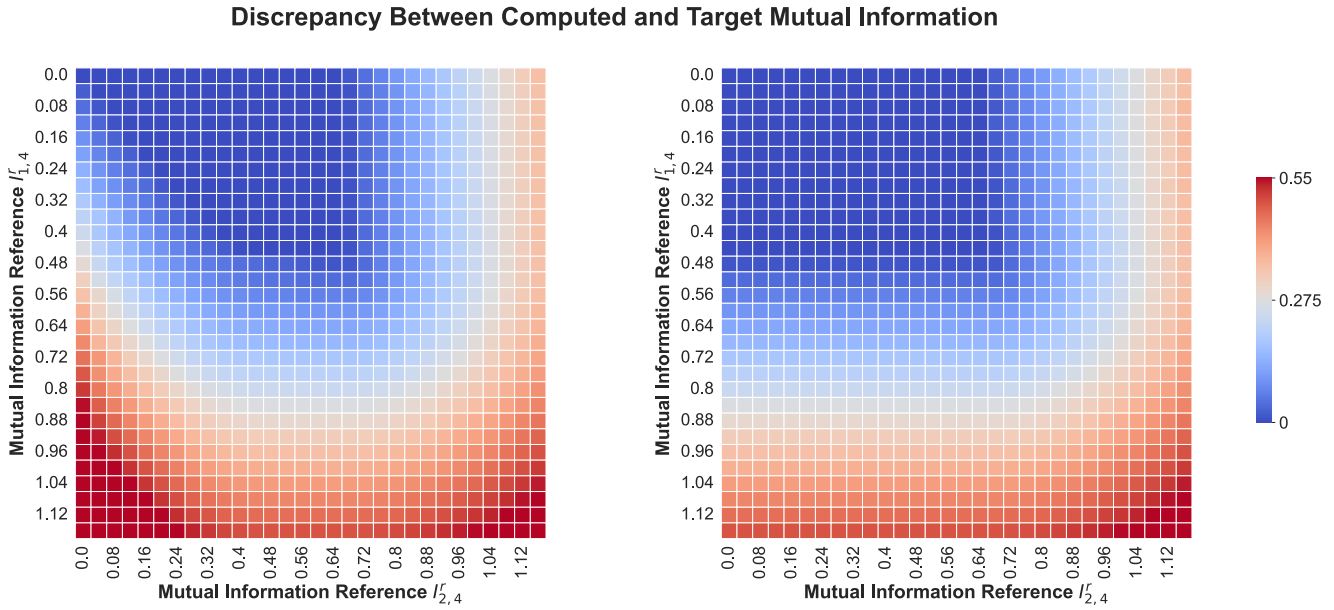


FIGURE 7. Root mean square error representing the discrepancy between the achieved mutual information and the reference one for the networks in Figure 6a (left) and Figure 6b (right), across a 2D mesh of target mutual information values. Light blue and red areas signify unachievable mutual information configurations, while dark blue indicates successful tracking of reference mutual information.

it holds that $I^{\max}_{1,4} \simeq 0.46$, while $I^{\max}_{2,4} \simeq 0.64$, for the case of a dataset with equally probable categorical features (as assumed in (22)). For this reason, the points in the 2D chart characterized by $I^r_{1,4} \geq 0.46$ and $I^r_{2,4} \geq 0.64$ exhibit a non-negligible root mean square error between the obtained mutual information and the target one for both the networks.

Interestingly, for the network in Figure 6a, the points such that $I^r_{1,4} \geq I^r_{2,4}$ cannot be tracked, as observed in the left part of Figure 7. This is because the considered network configuration, in which only the weights of certain edges are assumed to be different from zero, does not have enough degrees of freedom to independently set the values of mutual information between the considered variables. In fact, if the network in Figure 6b is considered, in which the weight $\alpha_{1,3}$ can take on any real value, the 2-dimensional region of feasible mutual information $I^r_{1,4}$ and $I^r_{2,4}$ expands to the rectangle depicted in dark blue in the right part of Figure 7. In this case, the proposed controller fails to track only the combination of mutual information references $I^r_{1,4}$ and $I^r_{2,4}$ that cannot be achieved by definition (due to the fact that the mutual information between to random variable has an upper bound, as discussed above).

As illustrated in Figure 7, the discrepancy between the target and achieved mutual information is presented as a heatmap. To provide a clearer understanding of the regions where the algorithm performs exceptionally well, Figure 8 offers an enhanced visualization. This figure highlights the points where the discrepancy is near-zero, indicating precise tracking, in dark blue. Conversely, points where the error exceeds a specific threshold (0.01) are masked, accentuating areas of successful mutual information tracking. Collectively, these figures underscore the proficiency of the proposed method in accurately tracking the reference mutual information across diverse network configurations, such as the ones considered in Figure 6a and Figure 6b.

B. MITIGATING BIAS IN A REALISTIC DATASET

In this subsection, the aim is to apply the proposed methodology to mitigate potential socioeconomic bias in a given dataset pertaining to job hiring decisions. To simplify the demonstration while preserving its applicability, it is assumed that the graphical model generating the considered dataset is fully known. In a real-life scenario, such model can be obtained by fitting the SEM's parameters on the considered dataset.

Hiring decisions are crucial determinants of individuals' career trajectories and overall economic stability. However, it is well-known that these decisions can sometimes be influenced by biases, whether conscious or unconscious. One potential source of bias is an applicant's socioeconomic status, which could unfairly influence their perceived suitability for a job. Therefore, the model consists of a set of variables that often carry weight during the hiring process. These include categorical variables such as the candidate's socio-economic status, their level of education and field of study, relevant work experience and skills, quality of references, networking skills, and the hiring decision (refer to 9 for a visual representation of the considered model). The dependencies between these variables are assumed to be known a priori and are encoded as edges in a Bayesian network. For instance, an applicant's socio-economic status can influence their level of education, relevant experience,

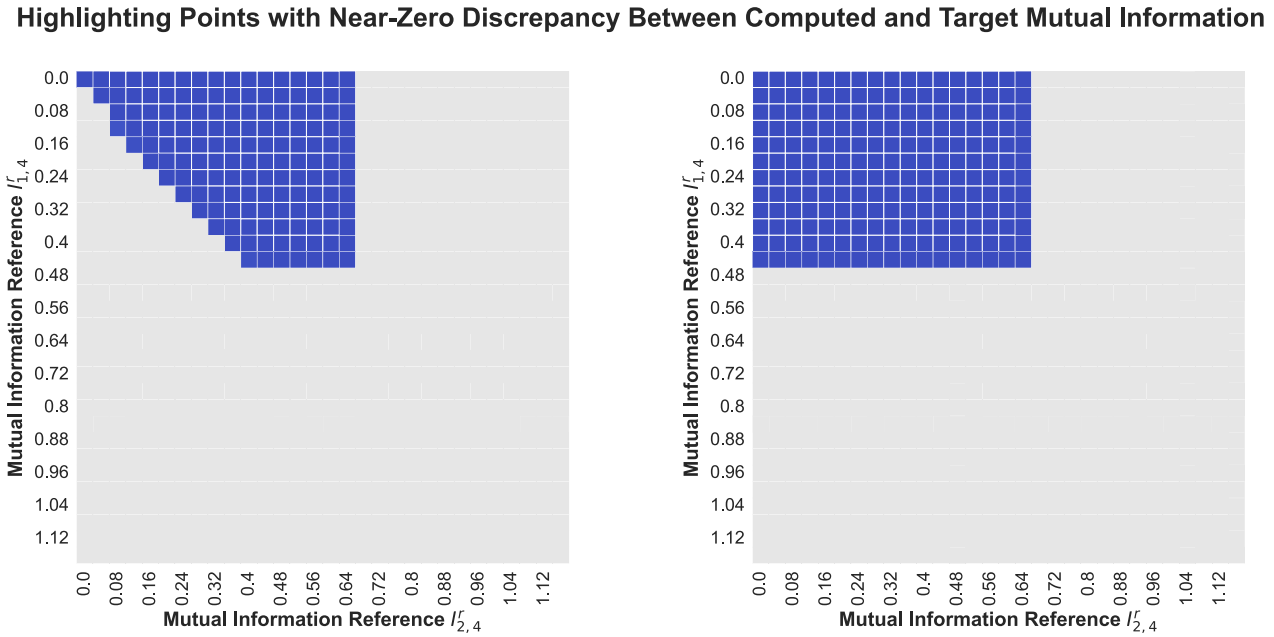


FIGURE 8. Highlighted points of successful reference mutual information tracking for the networks depicted in 6a (left) and 6b (right), respectively. This figure visualizes the heatmap in Figure 7, emphasizing near-zero discrepancy points (dark blue areas). Points exceeding a 0.01 error threshold are covered, highlighting successful algorithm tracking.

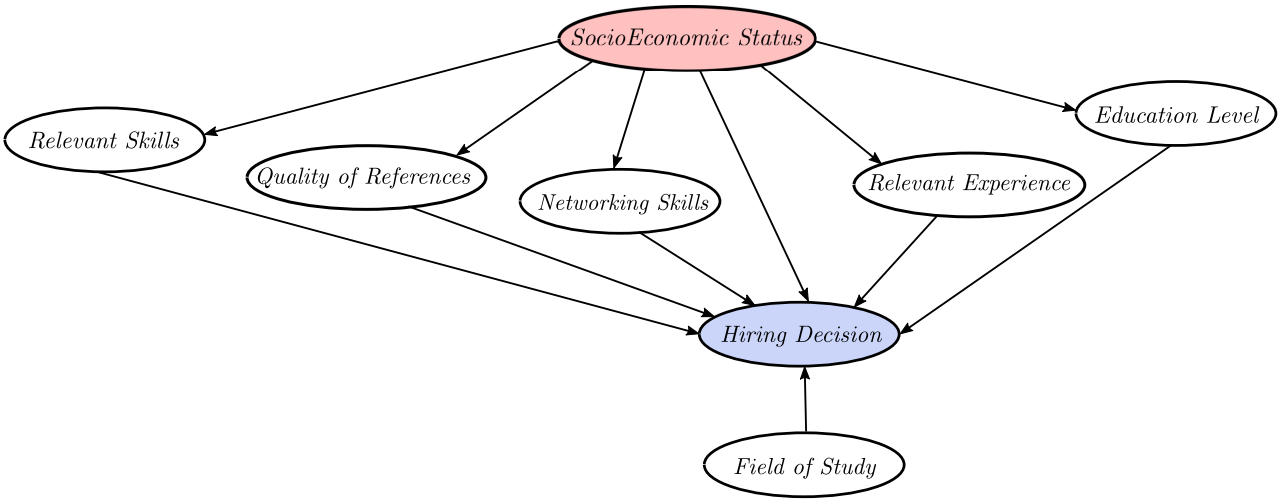


FIGURE 9. Graphical representation of the Bayesian network used in the case study in IV-B for bias mitigation. The network encompasses latent variables pertinent to the candidates' hiring process in the job market, including socio-economic status, relevant skills, and more. Only the latent variables are displayed for simplicity, omitting measured variables. Edges represent dependencies between variables, optimized using the proposed methodology to mitigate bias. Nodes corresponding to socio-economic status and hiring decision are distinctly marked in light red and light blue within the network, highlighting the presence of dataset bias between these variables in the original network parameters.

relevant skills, quality of references, and networking skills. The socio-economic status also has a direct link to the hiring decision, representing potential bias in the hiring process. The education background, job experience and skills, and references and networking skills also directly affect the hiring decision.

The original dataset is assumed to be generated by the model in Figure 9 with all the regression coefficients and variances set to 1. Interestingly, in this configuration, a mutual information of 0.356 exists between socio-economic status and hiring decision, suggesting a potential bias. This bias can be detrimental, possibly resulting in unfair hiring practices where socio-economic status unduly influences the hiring decision. The proposed two-step optimization algorithm is applied to this model to find the optimal regression coefficients and variances that can minimize such bias while preserving the other dependencies within the network. Specifically, the optimizations in 20 and 21

TABLE 4. Mutual information summary at bias mitigation levels: The table displays mutual information between key variable pairs during bias mitigation. Information between socioeconomic status and hiring decision lessens as bias mitigation rises, stabilizing for other pairs until 70% remaining bias. Beyond this, model accuracy is compromised.

Remaining Bias (%)	SocioEconomic Status / Hiring Decision	SocioEconomic Status / Relevant Skills	Relevant Skills / Hiring Decision
100%	0.357	0.207	0.194
90%	0.321	0.207	0.194
80%	0.285	0.207	0.194
70%	0.250	0.207	0.194
60%	0.222	0.207	0.189
50%	0.197	0.208	0.182
40%	0.173	0.208	0.175
30%	0.150	0.209	0.166
20%	0.126	0.209	0.154
10%	0.100	0.210	0.139
0%	0.055	0.208	0.106

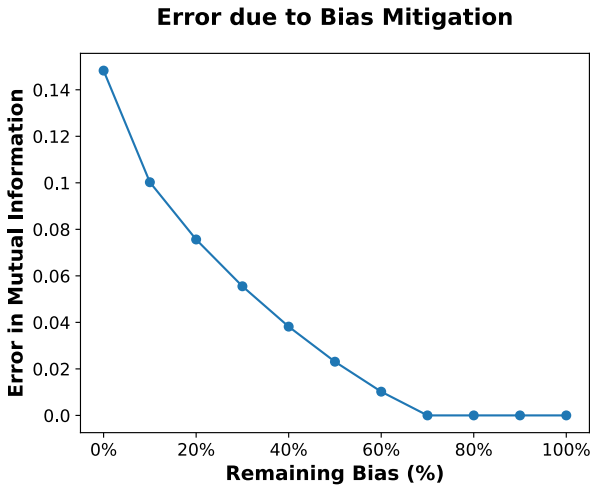


FIGURE 10. Illustration of the proposed two-step optimization algorithm's effectiveness in bias mitigation, applied to the graphical model in Subsection IV-B and Figure 9. The X-axis indicates the remaining bias as a percentage of the initial mutual information between socio-economic status and hiring decision, with 0% signifying total bias elimination and 100% no mitigation. The Y-axis denotes the root mean square error, highlighting the difference between post-optimization and reference mutual information values. These values aim to maintain consistent mutual information among other variables while achieving bias reduction. The graph demonstrates the potential to lower initial bias by up to 30%, leaving 70% residual bias, with minimal disturbance to other network dependencies. Efforts for further bias reduction beyond this limit yield significant approximation errors, unsettling the original inter-variable relationships in the network.

are repeatedly conducted with the intent to progressively diminish the mutual information between the socio-economic status variable and the hiring decision variable. Concurrently, these optimizations aim to maintain the mutual information variations among the remaining variables within acceptable boundaries. As depicted in Figure 10, the algorithm successfully mitigates up to 30%, leaving 70% of the initial bias, while minimally impacting the other dependencies within the network. This indicates that the graphical model with optimized parameters has the ability to generate a synthetic dataset that closely resembles the original dataset, while effectively reducing the amount of bias. However, attempting to further reduce the bias beyond the 30%, may result in the network losing its ability to accurately represent correlations in the dataset, as observed from an increase in approximation errors. Overall, these results highlight both the capabilities and limitations of the proposed methodology when addressing bias in realistic settings. While the methodology effectively diminishes bias up to a certain threshold, it also illuminates the inherent constraints tied to the degrees of freedom in the topological structure of the considered graphical model. Thus, while it holds substantial potential in fostering fairness in various domains reliant on data-driven decision-making, it also stresses the necessity for careful consideration of the model structure and its ability to adjust desired mutual information values.

Further details of the findings can be seen in Table 4, where mutual information for select pairs of variables is presented to maintain clarity. It becomes apparent that the mutual information between socioeconomic status and hiring decision decreases as efforts to mitigate bias are intensified. Notably, the mutual information between socioeconomic status and relevant skills remains largely unchanged throughout the entire mitigation process, as required my the optimization. Similarly, the mutual information between relevant skills level and hiring decision stays relatively stable up until a 30% reduction in the initial bias is reached, corresponding to a remaining bias of 70%. Beyond this point, changes in mutual information between these variables risk compromising the model's ability to accurately depict the dataset. Specifically, the algorithm's attempts to further minimize undesired bias inadvertently impact the essential causal relationship between relevant skills and hiring decision, a relationship that ideally should remain stable due to its significance in the hiring process.

V. CONCLUSION

In this paper, a novel two-step optimization algorithm has been presented for controlling the amount of bias between two categorical attributes in a given dataset, represented by the mutual information. Structural equation modeling and graphical models through directed acyclic graphs have been employed to represent the causal relationships between different dataset attributes, relying on latent Gaussian variables.

The proposed algorithm has been rigorously tested in two distinct contexts to demonstrate its performance under varied requirement conditions. A numerical case study provides a detailed investigation of the algorithm's capacity to adjust network parameters, resulting in desired mutual information values between specific pairs of attributes. This study underscores both the strengths of the algorithm and its limitations, revealing a fundamental relationship between the ability to achieve desired mutual information and the inherent degrees of freedom in the topological structure of the graphical model. Furthermore, the methodology is applied to a more complex scenario, involving a dataset related to job hiring decisions. This case study exemplifies how potential biases can be mitigated while preserving the relationships among other variables. It demonstrates the potential of the methodology to contribute to the generation of synthetic datasets that closely mirror the original ones, but with significantly reduced bias.

Furthermore, the impact of network configurations on the feasibility of tracking the desired mutual information has been explored. It has been observed that certain network configurations lack sufficient degrees of freedom to set the mutual information values independently, while other configurations can expand the feasible region for achieving the desired mutual information.

In conclusion, the proposed two-step optimization algorithm showcases promising potential in bias control within datasets, while adeptly maintaining other inherent relationships. The work contributes a novel and practical approach to bias mitigation in datasets and emphasizes the importance of understanding the structure of the underlying graphical model.

Looking ahead, the exploration of further applications of this methodology could be a fruitful avenue for future research. Specifically, applying the entire pipeline to real datasets - from fitting the Structural Equation Model to identifying and reducing bias - could prove highly beneficial. This work can be perceived as a pre-processing technique for bias mitigation, and the insights it yields could be invaluable when engaged with domain experts. Furthermore, the algorithm's performance could be enhanced and its applicability across various fields such as privacy-preserving data analysis, fairness in machine learning, and causal inference could be further explored.

ACKNOWLEDGMENT

The authors would like to thank Nicola Vanoli for his contribution in the analysis of the presented algorithm and for his valuable suggestions.

REFERENCES

[1] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surveys*, vol. 54, no. 6, pp. 1–35, Jul. 2022.

[2] D. Koller and N. Friedman, *Probabilistic Graphical Models*. Cambridge, MA, USA: MIT Press, 2009.

[3] D. Kaplan, *Structural Equation Modeling. Foundations and Extensions*, vol. 10, Jan. 2000.

[4] M. P. Bach, A. Topalović, and L. Turulja, "Data mining usage in Italian SMEs: An integrated SEM-ANN approach," *Central Eur. J. Oper. Res.*, vol. 34, pp. 941–973, Nov. 2022.

[5] M. Alshurideh, B. Al Kurdi, S. A. Salloum, I. Arpaci, and M. Al-Emran, "Predicting the actual use of m-learning systems: A comparative approach using PLS-SEM and machine learning algorithms," *Interact. Learn. Environments*, vol. 31, no. 3, pp. 1214–1228, Apr. 2023.

[6] W. Ahmed, "Understanding self-directed learning behavior towards digital competence among business research students: SEM-neural analysis," *Educ. Inf. Technol.*, vol. 28, no. 4, pp. 4173–4202, Apr. 2023.

[7] E. Barbierato, M. L. D. Vedova, D. Tessera, D. Toti, and N. Vanoli, "A methodology for controlling bias and fairness in synthetic data generation," *Appl. Sci.*, vol. 12, no. 9, pp. 1–15, 2022.

[8] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," 2016, *arXiv:1606.06565*.

[9] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 1334, pp. 183–186, 2017, doi: 10.1126/science.aal4230.

[10] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.

[11] T. M. Cover and J. A. Thomas, "Information theory and statistics," in *Elements of Information Theory*, vol. 1, no. 1. Hoboken, NJ, USA: Wiley, 1991, pp. 279–335.

[12] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 6, Jun. 2004, Art. no. 066138.

[13] E. van den Heuvel and Z. Zhan, "Myths about linear and monotonic associations: Pearson's r, Spearman's ρ, and Kendall's τ," *Amer. Statistician*, vol. 76, no. 1, pp. 44–52, Jan. 2022.

[14] P. Giudici and E. Raffinetti, "Shapley–Lorenz eXplainable artificial intelligence," *Expert Syst. Appl.*, vol. 167, Apr. 2021, Art. no. 114104.

[15] Y. Chen, P. Giudici, K. Liu, and E. Raffinetti, "Measuring fairness in credit scoring," *SSRN 4123413*, 2022.

[16] A. Rovetta, "Raiders of the lost correlation: A guide on using Pearson and Spearman coefficients to detect hidden correlations in medical sciences," *Cureus*, vol. 12, no. 11, p. e11794, 2020, doi: 10.7759/cureus.11794.

[17] A. Pozzi, X. Xie, D. M. Raimondo, and R. Schenkendorf, "Global sensitivity methods for design of experiments in lithium-ion battery context," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 7248–7255, 2020.

[18] D. Valencia, R. E. Lillo, and J. Romo, "A Kendall correlation coefficient between functional data," *Adv. Data Anal. Classification*, vol. 13, no. 4, pp. 1083–1103, Dec. 2019.

[19] G. J. Székely and M. L. Rizzo, *The Energy Data Distance Correlation*. Boca Raton, FL, USA: CRC Press, 2023.

[20] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[21] F. Shao and H. Liu, "The theoretical and experimental analysis of the maximal information coefficient approximate algorithm," *J. Syst. Sci. Inf.*, vol. 9, no. 1, pp. 95–104, Mar. 2021.

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[23] G. Arosio, G. Bagnara, N. Capuano, E. Fersini, and D. Toti, "Ontology-driven data acquisition: Intelligent support to legal ODR systems," in *Frontiers in Artificial Intelligence and Applications*, vol. 259. Amsterdam, The Netherlands: IOS Press, 2013, pp. 25–28.

[24] P. Del Nostro, F. Orciuoli, S. Paolozzi, P. Ritrovato, and D. Toti, "ARISTOTELE: A semantic-driven platform for enterprise management," in *Proc. 27th Int. Conf. Adv. Inf. Netw. Appl. Workshops*, Mar. 2013, pp. 44–49.

[25] A. Yale, S. Dash, R. Dutta, I. Guyon, A. Pavao, and K. P. Bennett, "Generation and evaluation of privacy preserving synthetic health data," *Neurocomputing*, vol. 416, pp. 244–255, Nov. 2020.

[26] A. Pozzi, S. Moura, and D. Toti, "A deep learning-based predictive controller for the optimal charging of a lithium-ion cell with non-measurable states," *Comput. Chem. Eng.*, vol. 173, May 2023, Art. no. 108222.

[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[28] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.
[29] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *Ann. Math. Statist.*, vol. 27, no. 3, pp. 832–837, Sep. 1956.
[30] A. Backurs, P. Indyk, and T. Wagner, "Space and time efficient kernel density estimation in high dimensions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 15799–15808.
[31] J. Guan, J. Lin, J. Guan, and E. Mokaramian, "A novel probabilistic short-term wind energy forecasting model based on an improved kernel density estimation," *Int. J. Hydrogen Energy*, vol. 45, no. 43, pp. 23791–23808, Sep. 2020.
[32] C. Chokwitthaya, Y. Zhu, S. Mukhopadhyay, and A. Jafari, "Applying the Gaussian mixture model to generate large synthetic data from a small data set," in *Proc. Construct. Res. Congr.*, Nov. 2020, pp. 1251–1260.
[33] A. Arora, N. Shoeibi, V. Sati, A. González-Briones, P. Chamoso, and E. Corchado, "Data augmentation using Gaussian mixture model on CSV files," in *Proc. Int. Symp. Distrib. Comput. Artif. Intell.* Cham, Switzerland: Springer, 2021, pp. 258–265.
[34] F. Hocaoglu, O. Gerek, and M. Kurban, "The effect of Markov chain state size for synthetic wind speed generation," in *Proc. 10th Int. Conf. Probablistic Methods Appl. Power Syst.*, May 2008, pp. 1–4.
[35] J. Dahmen and D. Cook, "SynSys: A synthetic data generation system for healthcare applications," *Sensors*, vol. 19, no. 5, p. 1181, Mar. 2019.
[36] S. Barocas and A. D. Selbst, "Big data's disparate impact," *California law Rev.*, vol. 104, no. 3, pp. 671–732, 2016.
[37] C. V. Gonzalez Zelaya, "Towards explaining the effects of data preprocessing on machine learning," in *Proc. IEEE 35th Int. Conf. Data Eng. (ICDE)*, Apr. 2019, pp. 2086–2090.
[38] Y. Zhang and J. Sang, "Towards accuracy-fairness paradox: Adversarial example-based data augmentation for visual debiasing," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 4346–4354.
[39] A. Rezaei, A. Liu, O. Memarrast, and B. D. Ziebart, "Robust fairness under covariate shift," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 11, 2021, pp. 9419–9427.
[40] M. Wan, D. Zha, N. Liu, and N. Zou, "In-processing modeling techniques for machine learning fairness: A survey," *ACM Trans. Knowl. Discovery From Data*, vol. 17, no. 3, pp. 1–27, Apr. 2023.
[41] X. Wang and H. Huang, "Approaching machine learning fairness through adversarial network," 2019, *arXiv:1909.03013*.
[42] B. An, Z. Che, M. Ding, and F. Huang, "Transferring fairness under distribution shifts via fair consistency regularization," *arXiv:2206.12796*, 2022.
[43] M. P. Kim, A. Ghorbani, and J. Zou, "Multiaccuracy: Black-box post-processing for fairness in classification," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Jan. 2019, pp. 247–254.
[44] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5680–5689.
[45] A. Mishler, E. H. Kennedy, and A. Chouldechova, "Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, Mar. 2021, pp. 386–400.
[46] I. Shrier and R. W. Platt, "Reducing bias through directed acyclic graphs," *BMC Med. Res. Methodol.*, vol. 8, no. 1, pp. 1–15, Dec. 2008.
[47] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, "Convergence properties of the Nelder–Mead simplex method in low dimensions," *SIAM J. Optim.*, vol. 9, no. 1, pp. 112–147, Jan. 1998.

ENRICO BARBIERATO received the B.S., M.Sc., and Ph.D. degrees in computer science. He is currently an Assistant Professor with the Department of Mathematics and Physics, Catholic University of Sacred Heart, Brescia, Italy. He worked for 25 years in IT consulting for the Banking, Telecommunications, and Energy&Utilities industries. His research interests include performance evaluation through multi-formalism and ethical AI.

ANDREA POZZI (Member, IEEE) received the bachelor's degree in industrial engineering, the master's degree in electrical engineering, and the Ph.D. degree in electronics, informatics, and electrical engineering from the University of Pavia, in 2015, 2017, and 2021, respectively.

He was a Visiting Scholar with TU Braunschweig and UC Berkeley, in 2016 and 2019, respectively. After serving as a Postdoctoral Researcher with the University of Pavia, he joined the Catholic University of Sacred Heart, Brescia, Italy, as an Assistant Professor of machine learning with the Faculty of Mathematical, Physical, and Natural Sciences, in January 2022. His research interests include reinforcement learning, imitation learning, machine learning, approximate dynamic programming, and advanced control theory.

DANIELE TESSERA (Member, IEEE) received the Ph.D. degree in computer engineering from the University of Pavia, Italy. He is currently an Associate Professor of computer science with the Department of Mathematics and Physics, Catholic University of Sacred Heart, Italy. He is the coauthor of more than 40 papers in international journals and conference proceedings. His research interests include performance debugging and benchmarking, workload characterization, cloud computing, and artificial intelligence applications.