Mitigating the Bias of Heterogeneous Human Behavior in Affective Computing

Shen Yan, Hsien-Te Kao, Kristina Lerman, Shrikanth Narayanan, Emilio Ferrara

University of Southern California, Information Sciences Institute

Email: {shenyan, hsiente, lerman, shri, ferrarae}@isi.edu

Abstract—Affective computing is broadly applied to decision making systems ranging from mental health assessment to employability evaluation. The heterogeneity of human behavioral data poses challenges for both model validity and fairness. The limited access to sensitive attributes (e,g., race, gender) in real-world settings makes it more difficult to mitigate the unfairness of the model outcomes. In this work, we focus on the heterogeneity of human behavioral signals and analyze its impact on model fairness. We design a novel method named *multi-layer factor analysis* to automatically identify the heterogeneity patterns in high-dimensional behavioral data and propose a framework to enhance fairness of behavioral modeling without accessing sensitive attributes.

Index Terms—Fairness, Bias, Heterogeneity

I. INTRODUCTION

Data fuels every aspect of today's world. Social and human systems generate unparalleled amounts of data that are used, in conjunction with machine learning techniques, to understand individual and collective behavior, for decision making, and to inform policies and laws. Various studies use behavioral data and machine learning techniques to implement systems that aim to understand and track human affects. Example of such systems exist in various application domains, from health assessment [1] to job performance evaluation [2]. However, individuals' physiological and psychological differences may introduce various sources of biases in the data, most prominently heterogeneity [3], [4]. This, in turn, can affect machine learning models' accuracy [5], [6] and model fairness [7], [8].

An example of heterogeneity in human behavioral data is the Simpson's paradox [9], a phenomenon wherein an association or a trend observed at the level of the entire population disappears or even reverses when data is disaggregated by its underlying subgroups. This phenomenon is common in human behavioral data [10]. Failure to take the heterogeneous patterns into account during the modeling process might impair both the utility and fairness of the system [11]. Researchers have explored different techniques to test and discover such patterns in the past [12]–[14]. However, previous methods either rely on the group labels or are only able to capture simplified patterns, which are not suitable for the fairness-aware models when the group labels (i.e., sensitive attributes) are not observable.

In addition, most of previous work on fairness-aware machine learning requires the access to sensitive attributes. However, in many real-world applications, sensitive attributes, like gender, race, etc., are not observable due to privacy concerns or legal restrictions. Recent regulations such as the European General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), or the Health Insurance Portability and Accountability Act (HIPAA) regulate the usage of personal data. For example, credit institutions cannot ask or access information about race to applicants who apply for credit [Equal Credit Opportunity Act: 15, 12 CFR §1002.5(b)]. Similarly, insurance companies can no longer request race information from the individuals they insure [15]. From January 2020, potential employers will no longer be allowed to request previous-salary information from perspective employees. Thus, fairness-aware models that don't rely upon, and don't access sensitive attributes are better suited to real-world applications.

Contributions of this work

Motivated by the above challenges in affective computing, in this work we focus on mitigating the unfair impact of heterogeneous behavioral features without accessing sensitive attributes. In summary, our contributions in this paper are as follows:

- We analyze the impact of different behavioral patterns on model utility and model fairness.
- We propose a method to identify heterogeneity patterns without sensitive attributes, named *multi-layer factor analysis*.
- We propose a framework combining *multi-layer factor analysis* and feature rescaling to mitigate the bias in affective computing without accessing sensitive attributes. Experimental results show that the proposed framework improves model fairness.

II. RELATED WORK

The problem of fairness in machine learning has been drawing increasing research interests over the course of recent years. Most of the previous work focus on classification tasks [16], [17]. A few papers considered fairness in regression problems. Convex [18] and non-convex [19] frameworks have been explored to add fairness constraints to regression models. Quantitative definitions and theoretical performance guarantees in fair regression have also been discussed in recent work [20]. All of the strategies above require the access to sensitive attributes in order to mitigate the source of bias. However, collecting that type of information might be difficult, or even forbidden by laws, in real-world applications.

Recently, a few studies have explored different strategies to address this issue. One typical solution is using *non-sensitive information* as proxy for sensitive attributes. Previous work [21] has shown that non-sensitive information can be highly correlated with sensitive attributes. *Proxy fairness* [22] leverages the correlations between proxy features and true sensitive attributes. Proxy features are used as the alternative to sensitive attribute(s) when applying a standard fairness-improving strategy. Although the existence of proxy features gives the hope to improve fairness with unobserved sensitive attributes, identifying perfect proxy groups is still challenging.

Some researchers have explored methods to uncover latent heterogeneous patterns in the data [12]. Several recent studies also investigated the use of disaggregation methods without sensitive attributes [13], [14]. However, previous work mainly focuses on finding the optimal partition of each feature, which fails to capture more complex scenarios such as when different groups share overlapping feature ranges. Our proposed work will also address this issue.

III. PRELIMINARIES

A. Fair regression

In this paper, we study the problem of model fairness in a regression setting, where the goal is to predict a true outcome $Y \in [a, b]$ from a feature vector X based on labeled training data. The fairness of prediction $\hat{Y}$ of model M is evaluated with respect to *sensitive groups* of individuals defined by *sensitive attributes* A, such as gender or race. Sensitive attributes A are assumed to be binary, i.e., $A \in \{0, 1\}$, where $A = 1$ represents the privileged group (e.g., male), while $A = 0$ represents the underprivileged group (e.g., female). Such simplifications are generally used in the literature, although criticisms of reductionism are widely acknowledged [23], [24].

We consider two quantitative definitions of fairness appearing in prior works on fair classification and regression [25]: *statistical parity* (SP) and *equal accuracy* (EA).

*Definition 1 (**Statistical Parity (SP) [26]**):* A model M satisfies statistical parity under a distribution over (X, A, Y) if $M(X)$ is independent of the protected attribute A. Since $M(X) \in [a, b]$, the statistical parity of M is defined as

$$\Pr[M(X) \geq z | A = 0] = \Pr[M(X) \geq z | A = 1],$$

where $z \in [a, b]$.

Statistical Parity aims to equalize the distribution differences of the outcomes across different sensitive groups. In this work, SP is measured by the distance of average outcome of each sensitive group.

*Definition 2 (**Equal Accuracy (EA)**):* Equal Accuracy rewards the model M for predicting each group as accurate at the same rate. The equal accuracy of M defined as

$$\mathbb{E}[\epsilon(Y, M(X)) | A = 0] = \mathbb{E}[\epsilon(Y, M(X)) | A = 1],$$

where $\epsilon(Y, (M(X)))$ represents the estimation error.

In this work, we choose Mean Absolute Error (MAE) as the metric for regression task, thus we evaluate the Equal Accuracy with the MAE differences between groups.

B. Fisher's linear discriminant

Suppose two groups of p-dimensional samples $\vec{x}_0$, $\vec{x}_1$ have means $\vec{\mu}_0 = [\mu_{01}, \dots, \mu_{0p}]$, $\vec{\mu}_1 = [\mu_{11}, \dots, \mu_{1p}]$ and covariance Σ_0, Σ_1. Then the linear combination of features $\vec{w} \cdot \vec{x}_i$ has means $\vec{w} \cdot \vec{\mu}_i$ and variances $\vec{w}^T \Sigma_i \vec{w}$ for $i = 0, 1$. Fisher defined the separation between these two distributions to be the ratio of the variance between the groups to the variance within the groups:

$$S = \frac{\sigma_{\text{between}}^2}{\sigma_{\text{within}}^2} = \frac{(\vec{w} \cdot \vec{\mu}_1 - \vec{w} \cdot \vec{\mu}_0)^2}{\vec{w}^T \Sigma_1 \vec{w} + \vec{w}^T \Sigma_0 \vec{w}} = \frac{(\vec{w} \cdot (\vec{\mu}_1 - \vec{\mu}_0))^2}{\vec{w}^T (\Sigma_0 + \Sigma_1) \vec{w}}.$$

It can be shown that the maximum separation occurs when

$$\vec{w} \propto (\Sigma_0 + \Sigma_1)^{-1} (\vec{\mu}_1 - \vec{\mu}_0). \tag{1}$$

C. Factor analysis

Suppose we have a set of p observable random variables $x_1, \dots, x_p$ with means $\mu_1, \dots, \mu_p$.

For some unknown constants l_{ij} and k unobserved random variables F_j (i.e., common factors), where $i \in \{1, \dots, p\}$ and $j \in \{1, \dots, k\}$, where $k < p$, we have

$$x_i - \mu_i = l_{i1} F_1 + \cdots + l_{ik} F_k + \varepsilon_i.$$

Here, the ε_i are unobserved stochastic error terms with zero mean and finite variance, which may not be the same for all i. In matrix terms, we have $\vec{x} - \vec{\mu} = LF + \vec{\varepsilon}$. F is defined as the factors, and L as the loading matrix.

Suppose the covariance matrix of $(\vec{x} - \vec{\mu})$ is Σ, we have

$$\Sigma = LL^T + \Psi. \tag{2}$$

## IV. IMPACT OF HETEROGENEITY ON FAIRNESS

Heterogeneity is often present in social and behavioral data, and its presence affects the analysis of trends as well as the accuracy of prediction tasks [10]. In this section, we analyze different heterogeneous patterns and their impact on the fairness of model outcomes. Figure 1 illustrates the bias derived from heterogeneity. If ignoring the heterogeneous patterns, the trends learned from the data can be biased against certain groups.

In this section, we conduct our analysis on synthetic datasets. Using synthetic data is important because we can arbitrarily control the characteristics of the paradox, which will allow us to understand what effects its presence has on model fairness and accuracy.

Our synthetic datasets have 1000 samples and 10 informative features from two sensitive groups with same number of samples. For each of the dataset, N out of the 10 features exhibit a pattern compatible with heterogeneity. We focus on six different common heterogeneous patterns listed in Table I. Pattern #1-3 appear in the datasets when the target variable of the two sensitive groups shares the same range; this set of examples is hence named *Shared-Range* data; furthermore, Pattern #4-6 consider the situation when group 1 has overall lower ranges of target variables than group 2, named *Different-Range* data.
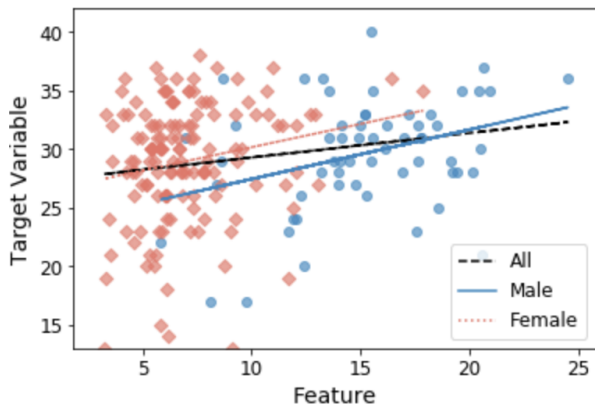
(a)
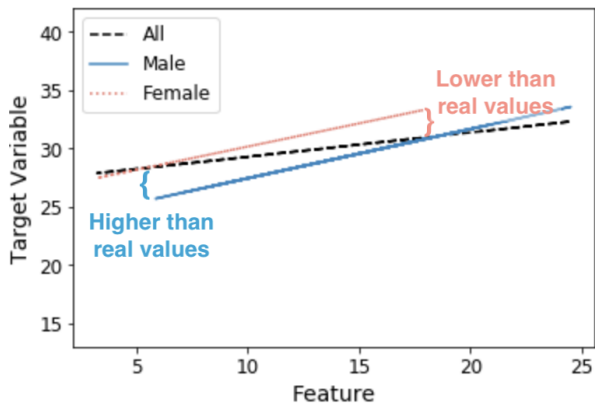
(b)

Fig. 1. **Example of bias from heterogeneity.** The plots illustrate the bias derived from heterogeneity. (a) shows a heterogeneous pattern exists in a real-life dataset. If the model ignores the heterogeneity, the learned trend (i.e., black dash line in (a)) will discriminate against the female samples (i.e., red) as shown in (b).

(1) Shared-Range

Pattern		N	Metrics		
			MAE	EA	SP
No Heterogeneity		-	1.13	-0.04	0.06
#1	group 1 / group 2	1	1.79	-0.13	-1.51
		2	2.35	0.05	-1.67
		3	2.75	-0.04	-1.66
		4	2.89	-0.02	-1.57
		5	3.05	-0.05	-1.53
#2	group 1 / group 2	1	1.67	-0.02	-2.04
		2	2.17	0.05	-2.55
		3	2.51	0.05	-2.64
		4	2.74	-0.01	-2.55
		5	2.93	0.12	-2.54
#3	group 2 / group 1	1	1.92	0.02	-0.07
		2	2.40	-0.02	0.07
		3	2.77	-0.01	-0.04
		4	2.97	-0.03	-0.01
		5	3.17	0.05	-0.01

(2) Different-Range

Pattern		N	Metrics		
			MAE	EA	SP
No Heterogeneity		-	5.07	0.01	0.02
#4	group 1 / group 2	1	2.30	0.05	-8.59
		2	2.30	0.17	-8.69
		3	2.36	0.18	-8.62
		4	2.38	0.17	-8.55
		5	2.30	0.24	-8.74
#5	group 1 / group 2	1	2.82	0.19	-6.71
		2	2.83	0.21	-6.70
		3	2.85	0.22	-6.61
		4	2.76	0.08	-6.80
		5	2.73	0.06	-6.93
#6	group 2 / group 1	1	4.09	0.04	-4.03
		2	4.07	0.03	-4.17
		3	4.11	0.05	-3.94
		4	4.11	0.03	-4.07
		5	4.10	0.05	-4.04

A linear regression model is trained on each dataset and evaluated with 10-fold cross validation. We adopt mean absolute errors (MAE) to measure the overall accuracy of the predictions. *Statistical parity* (SP) is measured by the distance of average outcome of each sensitive group; *equal accuracy* (EA) is measured by the distance of MAE across different groups.

Based on the results in Table I, different patterns show different impact on the model outcomes. In *Shared-Range* data, Pattern #1-3 have negative impact on the model accuracy with respect to the overall MAE. As the number N of features with heterogeneity increases, the overall MAE increases accordingly. Pattern #2 shows the most significant impact on the *statistical parity* of outcomes. The heterogeneity causes more pronounced negative impact on both fairness metrics in *Different-Range* data.

V. METHODS

In this section, we propose a method to mitigate the unfairness caused by heterogeneity. Our method includes two parts:

identifying the heterogeneous patterns based on factor analysis and feature rescaling.

A. Identifying Heterogeneous Patterns

Unveiling the heterogeneous patterns in complex behavioral data is challenging, especially when the group labels are not available. In this work, we propose a method to identify heterogeneous patterns in multi-variate behavioral data leveraging factor analysis, named *multi-layer factor analysis* (MLFA). MLFA is based on the effectiveness of factor analysis to separate subgroups with heterogeneity on balanced correlated features, as shown in Theorem 1.

Theorem 1: Let X be a dataset of n variables $x_1, \ldots, x_n$ exhibiting heterogeneous patterns between two groups X_0 and X_1. F represents the factor matrix of X after factor analysis.
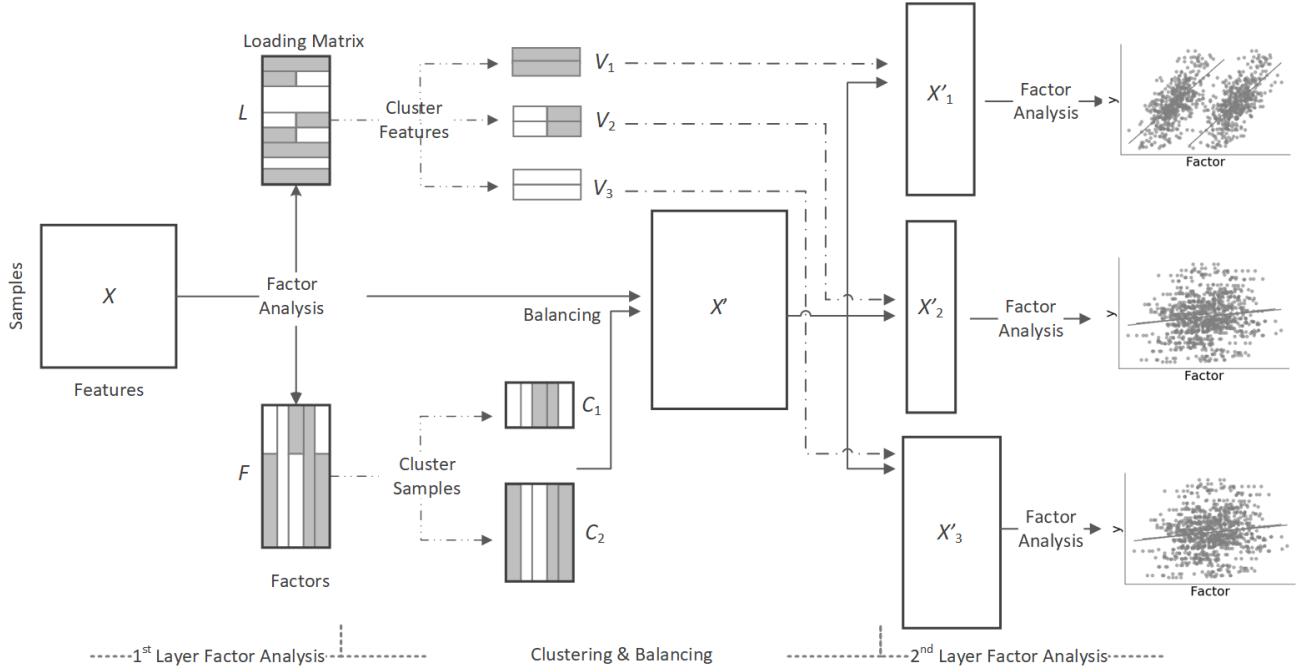
Fig. 2. **Multi-Layer Factor Analysis (MLFA) framework.** The first layer factor analysis discovers the feature clusters V_i and sample clusters C_i. The original dataset X is balanced based on C_i and then separated into subsets $X'_1, \ldots, X'_k$, where each subset X'_i only contains the features in V_i. We then conduct the second layer factor analysis on $X'_1, \ldots, X'_k$.

Assume the two groups of observations follow $\mathcal{N}(\vec{\mu}_0, \Sigma_0)$ and $\mathcal{N}(\vec{\mu}_1, \Sigma_1)$, respectively. F shows the maximum separation of Fisher's linear discriminant between groups when X_0 and X_1 have same size and $x_1, \ldots, x_n$ are highly correlated with each other.

Proof 1: Since X consists of X_0 and X_1 following $\mathcal{N}(\vec{\mu}_0, \Sigma_0)$ and $\mathcal{N}(\vec{\mu}_1, \Sigma_1)$, respectively, X can be viewed as a mixture Gaussian distribution $\mathcal{N}(\vec{\mu}, \Sigma) = \sum_{i=0}^{1} \alpha_i \mathcal{N}(\vec{\mu}_i, \Sigma_i)$. Thus,

$$\vec{\mu} = \sum_{i=0}^{1} \alpha_i \vec{\mu}_i$$

$$\Sigma = \sum_{i=0}^{1} \alpha_i \Sigma_i + \sum_{i=0}^{1} \alpha_i (\vec{\mu}_i - \vec{\mu})(\vec{\mu}_i - \vec{\mu})^T.$$

If X_0 and X_1 have same size, $\alpha_0 = \alpha_1 = \frac{1}{2}$,

$$\vec{\mu} = \frac{\vec{\mu}_0 + \vec{\mu}_1}{2},$$

$$\Sigma = \frac{1}{2}(\Sigma_0 + \Sigma_1 + (\frac{\vec{\mu}_0 - \vec{\mu}_1}{2})(\frac{\vec{\mu}_0 - \vec{\mu}_1}{2})^T + (\frac{\vec{\mu}_1 - \vec{\mu}_0}{2})(\frac{\vec{\mu}_1 - \vec{\mu}_0}{2})^T)$$

$$= \frac{1}{2}(\Sigma_0 + \Sigma_1 + \frac{(\vec{\mu}_1 - \vec{\mu}_0)^2}{2}).$$

Let F and L respectively represent the factor matrix and loading matrix of X after factor analysis. The transformation matrix from X to F is $\Sigma^{-1} \cdot L$, thus

$$\vec{w}_j = \Sigma^{-1} \cdot l_j,$$

where $\vec{w}_j$ represents the projection vector from X to factor F_j.

When $x_1, \ldots, x_n$ are highly correlated with each other, $\Sigma \approx \beta_1 J_{n,n}$, $\vec{\mu}_1 - \vec{\mu}_0 \approx \beta_2 J_{n,1}$, where β_1 and β_2 are constant values and J is all-ones matrix. According to Eqn. 2, $l_j = \beta'_1 J_{n,1}$,

$$\Sigma = \frac{1}{2}(\Sigma_0 + \Sigma_1 + \frac{(\vec{\mu}_1 - \vec{\mu}_0)^2}{2})$$

$$= \frac{1}{2}(\Sigma_0 + \Sigma_1 + \frac{\beta_2^2 J_{n,n}}{2}) \approx \beta_1 J_{n,n}.$$

Thus, $(\Sigma_0 + \Sigma_1) = (2\beta_1 - \frac{\beta_2^2}{2}) J_{n,n}$.

According to Fisher's linear discriminant theory, the ratio of $\vec{w}_j$ to Eqn. 1 is

$$\frac{\vec{w}_j}{(\Sigma_0 + \Sigma_1)^{-1}(\vec{\mu}_1 - \vec{\mu}_0)} = \frac{\Sigma^{-1} \cdot l_j}{(\Sigma_0 + \Sigma_1)^{-1}(\vec{\mu}_1 - \vec{\mu}_0)}$$

$$= \frac{(2\beta_1 - \frac{\beta_2^2}{2}) J_{n,n} l_j}{\beta_1 J_{n,n}(\vec{\mu}_1 - \vec{\mu}_0)}$$

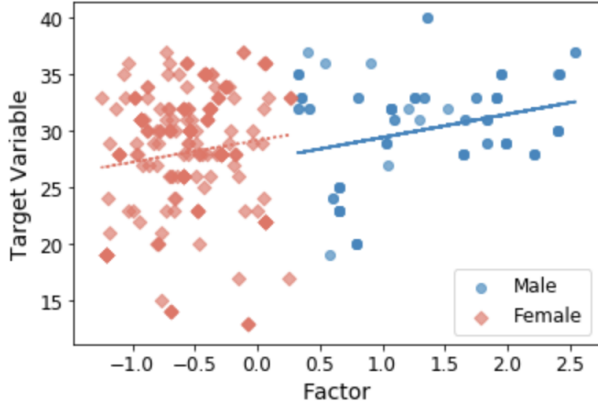$$= \frac{(2\beta_1 - \frac{\beta_2^2}{2})\beta'_1}{\beta_1 \beta_2}.$$

Therefore, $\vec{w}_j \propto (\Sigma_0 + \Sigma_1)^{-1}(\vec{\mu}_1 - \vec{\mu}_0)$, the maximum separation occurs.

B. Multi-Layer Factor Analysis (MLFA) framework

Inspired by Theorem 1, we propose the framework of *multi-layer factor analysis* (MLFA) as shown in Figure 2. This framework has three steps:
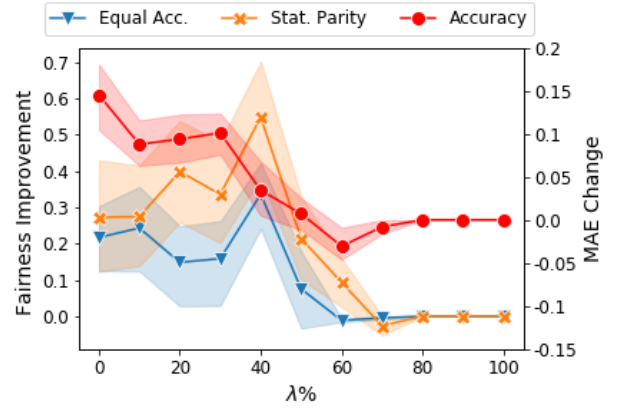
(a) Traditional factor analysis
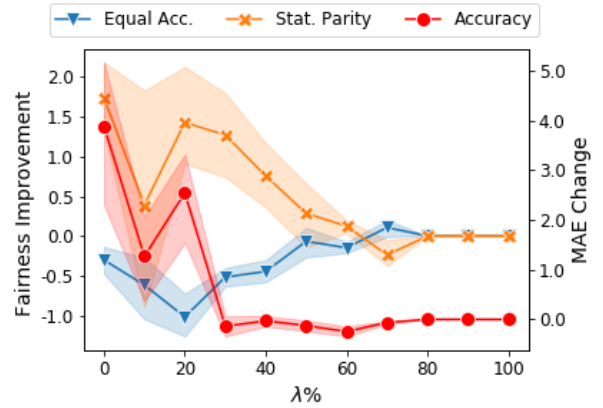

(a) TILES

(b) MLFA

(b) Older Adults
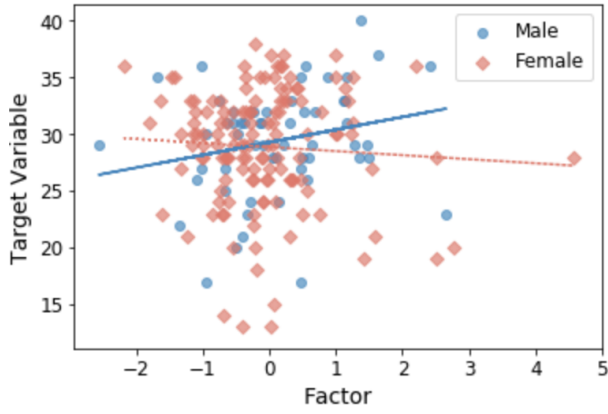
Fig. 3. **Example of MLFA Outcomes.** The plots compare the extracted factors using (a) traditional factor analysis and (b) the MLFA framework. The factor from MLFA shows more separable structure.

Fig. 4. **Effects of the parameter** λ. The plots illustrate the performance change on fairness and accuracy, according to three metrics (*Equal Accuracy*, *Statistical Parity*, *Accuracy*), for *TILES* (left) and *Older Adults* (right) data.

- **First layer factor analysis:** conduct factor analysis on the original dataset X. Get factor matrix F and loading matrix L.
- **Clustering and balancing:** cluster the samples based on F into m sample clusters $C_1, \ldots, C_m$ and cluster the features based on L into k feature clusters $V_1, \ldots, V_k$; balance X based on the size of sample clusters, making sure the balanced the dataset X' has same number of samples from each sample cluster C_i; then divided X' based on the feature clusters. Get $X'_1, \ldots, X'_k$. The clustering and balancing step aims to balance the dataset, making the input subsets $X'_1, \ldots, X'_k$ of the second layer factor analysis satisfies the assumptions of equal-sized groups and correlated features.
- **Second layer factor analysis:** conduct factor analysis on $X'_1, \ldots, X'_k$. Get factor matrix $F_1, \ldots, F_k$.

If features within V_j exhibit heterogeneity, F_j will show a clustered structure, $j \in [1, k]$. Figure 3 gives a empirical illustration of the performance of MLFA. Figure 3 (a) and (b) show the most informative factor of the heterogeneous features in a real life behavioral dataset *TILES* (See details in §VI-A). Comparing to traditional factor analysis, MLFA extracts a more separable factor.

We use *Gaussian Mixture Model* (GMM) as the clustering algorithm with Bayesian information criterion (BIC) and average sum of squared distances (SSD) within clusters as evaluation metrics to testify the assumption on F_j. If F_j exist a clustered structure, BIC and SSD will decrease after clustering.

We further introduce a parameter λ as the criteria of identifying good cluster structures. Clusters with SSD decrease more than $\lambda\%$ will be considered as meaningful clusters, the corresponding features are identified with heterogeneity.

C. Feature rescaling

After identifying the heterogeneous features, we rescale those features to mitigate their impact on the model outcomes. We adopt the *disparate impact remover* [27] as the rescaling method. *Disparate impact remover* is a preprocessing technique that edits feature values to improve group fairness while preserving rank-ordering within groups. After the rescaling, the feature distributions across groups are hard to distinguish while the individual's ranking within their group will be preserved.

TABLE II
EXPERIMENTAL PERFORMANCE ($\lambda = 0$). COMPARE THE UTILITY AND FAIRNESS PERFORMANCE ON THE DATA WITH AND WITHOUT OUR PROPOSED METHOD. THE RESULTS SHOW THAT OUR METHOD YIELDS DECENT IMPROVEMENTS ON FAIRNESS AND CAN EVEN IMPROVE MODEL ACCURACY.

Dataset	Method	Linear Regression			Decision Tree			Random Forest		
		MAE	EA	SP	MAE	EA	SP	MAE	EA	SP
Synthetic 1	Original	**1.76**	**0.44**	0.64	4.08	0.29	0.27	3.27	0.23	0.16
	MLFA (Ours)	1.78	0.50	**0.28**	**4.01**	**0.26**	**0.14**	**3.14**	**0.23**	**0.05**
Synthetic 2	Original	**1.76**	0.29	0.45	3.96	0.27	0.69	3.15	**0.19**	0.19
	MLFA (Ours)	1.78	0.48	**0.39**	**3.92**	0.43	**0.18**	**3.10**	0.25	**0.15**
TILES	Original	**3.20**	0.74	1.17	3.34	0.65	**0.39**	3.19	0.67	0.53
	MLFA (Ours)	3.42	**0.74**	**0.66**	**3.17**	**0.59**	0.41	**3.15**	**0.61**	**0.38**
Older Adults	Original	**6.43**	**2.30**	4.95	**6.43**	**2.39**	5.24	**6.11**	**1.70**	4.43
	MLFA (Ours)	7.06	2.46	**4.12**	6.73	3.82	**2.57**	6.78	2.61	**2.71**

TABLE III
COMPARISON OF THE DEBIASING PERFORMANCE OF RANDOM FOREST MODELS WITH AND WITHOUT SENSITIVE ATTRIBUTES (SA).

Type	Dataset	Metric	Method			
			w/ SA		w/o SA	
Shared-Range	*TILES*	MAE	3.17	-0.6%	**3.15**	**-1.2%**
		EA	0.87	+29.8%	**0.61**	**-8.9%**
		SP	**0.12**	**-77.3%**	0.38	-28.3%
Different-Range	*Older Adult*	MAE	6.90	+12.9%	**6.78**	**+10.9%**
		EA	3.07	+80.5%	**2.61**	**+53.5%**
		SP	**0.39**	**-91.1%**	2.71	-38.8%

Fig. 5. **The proposed fair modeling pipeline.** The train dataset is pre-processed by MLFA and feature rescaling to mitigate the bias as well as learn the pre-processing models. The pre-processing models are used to rescale the test dataset.

D. Modeling Pipeline

Figure 5 shows the pipeline of our proposed fair machine learning pipeline for affective computing.

The train dataset will be pre-processed by our proposed MLFA and feature rescaling methods to mitigate the biases embedded in the feature dimension. After the pre-processing step, the processed data are used for model training. The pre-processing step also generates clustering models used for test dataset processing.

In the testing stage, samples in the test set are clustered by the GMM models learned from the training set and be processed based on the clustered groups. Note that the proposed MLFA method also has a clustering step after the 1st layer of factor analysis, the clustering method used in this step is not restricted.

VI. EXPERIMENTS

A. Datasets

1) Synthetic: We generate two synthetic datasets with 40 features and 1000 samples, where each sensitive group has 500 samples and 20 out of the 40 features exhibit different patterns. *Synthetic 1* is *Shared-Range* data. The target variable Y has the same range $[0, 30]$ across groups. *Synthetic 2* is *Different-Range* data, $Y \in [0, 20]$ in group 1 and $Y \in [10, 30]$ in group 2.

2) TILES: Tracking Individual Performance with Sensors (TILES) [28], [29] is a 10 weeks longitudinal study with 212 hospital worker volunteers in a large Los Angeles hospital, where 30% of the participants are male and 70% are female. Bio-behavioral signal data were collected from continuous sensing of garment-based sensor. Physiological signals and physical activities information including heart rate, breathing rate, step counts, accelerometer, etc. are used as the predictive features. In this work, we focus on predicting the *cognitive ability*. The target variables are collected from the pre-study survey. We use the gender information of each participant as the sensitive attribute. The distribution of *cognitive ability* shows no statistical difference (p-value=0.34 under t-test) between genders. This dataset is an example for *Shared-Range* patterns.

3) Older Adults: The older adults dataset [30] is collected to study the relationship between physical fitness and cognitive performance. 70 older adults (28 male and 42 female) with a mean age 71±4.7 years were recruited. Physical activity tests included the 6-minute walk test, Bicep Curls, Static and Dynamic Balance, Timed Up and Go, Sit to Stand, Grip strength, and Functional Reach. The Stroop Task [31] is used to measure cognitive performance. In this work, we use the number of total mistakes in Stroop Task as the target variable. Female participants have 12.3 mistakes in average, while male participants only have in average 7.32 mistakes, which has similar properties to the *Different-Range* data.

B. Performance assessment

For all our experiments, each dataset is randomly split into 90% development set and 10% test set with 50 repeats.

We use *KMeans* as the clustering method in our *multi-layer factor analysis* framework. As for pre-processing, we
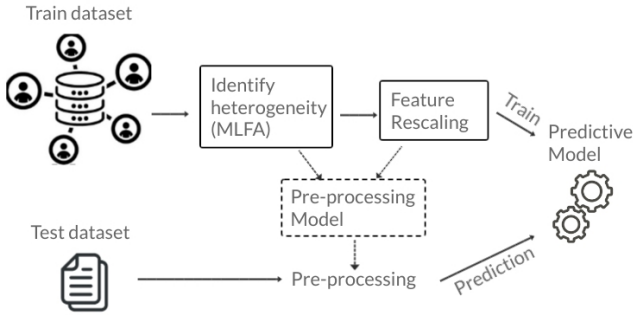
evaluate our proposed method with three different types of standard regression models: *linear regression*, *decision tree*, and *random forest*. All models are implemented by using the `scikit-learn` library.

We validate the performance of our method on both synthetic and real-life datasets. Among the 4 datasets, *Synthetic 1* and *TILES* are *Shared-Range* data, *Synthetic 2* and *Older Adults* are *Different-Range* data.

Our evaluation metrics include *accuracy*, *statistical parity*, and *equal accuracy*:

- **Accuracy**: We adopt *mean absolute errors* (MAE) to measure the overall accuracy of the predictions.
- **Statistical parity (SP)** is measured by the distance of outcomes of each sensitive group.
- **Equal accuracy (EA)** is measured by the distance of MAE across different groups.

To validate the performance of our proposed method, we also compare the model performance with traditional pre-processing fair machine learning strategies. Table II and Table III compare the performance of the three following model strategies:

- **Original:** The model is trained based on the original datasets without any pre-processing.
- **Debiasing with true sensitive groups:** The model is trained based on the datasets processed by *disparate impact remover* [27]. To evaluate the true performance of our method, we use the real sensitive attributes in the pre-processing step. That implies that this model is significantly advantaged compared to models that do not access the sensitive attributes.
- **Debiasing without true sensitive groups (Ours):** The MLFA model is trained based on the datasets processed by our proposed fair modeling pipeline shown in Fig. 5. No sensitive attributes are needed in the process. That puts our model at significant disadvantage compared to e.g., the *disparate impact remover*.

Our proposed method can yield improved fairness metrics in all experimental settings. MLFA is mainly aiming to balance the outcome distribution across groups (i.e., statistical parity), as shown in Figure 1. For *Shared-Range* datasets, for instance, *Synthetic 1* and *TILES* datasets, our method can improve all three metrics including model accuracy. Comparing to the debiasing method with true sensitive attributes, as shown in Table III, our method can achieve better fairness with less accuracy loss (i.e., MAE increase).

Effects of λ. In our method, we introduce the parameter λ as the criteria to identify heterogeneity. As λ increases, only the factors having more separable clusters would be considered as the indicators of heterogeneity.

However, the best λ is highly dependent on the properties of the datasets. Too small λ might have negative impact on model accuracy due to the mis-identification of heterogeneity. Rescaling wrong features might lose informative information for the regression model. Too large λ might have negative

impact on model fairness due to the failure of identifying heterogeneity.

Figure 4 shows how performance changes as a function of λ. For all the three metrics of interest, i.e., MAE, SP, and EA, the closer they are to zero, the better the model performs. Therefore, the performance improvements are defined as the values reduced on each metric after applying our method.

Both *TILES* and *Older Adults* datasets show decresing trends of MAE change when λ increases, indicating better model accuracy. In terms of the fairness improvements, the *TILES* dataset (i.e., *Shared-Range* data) shows similar trends for both EA and SP metrics, the most improvement happens when $\lambda = 40$. For the *Older Adults* dataset (i.e., *Different-Range* data), the improvements of SP metrics come at the cost of EA metrics as our method is aiming to optimize the SP. The SP improvement has a decreasing trend as λ increases.

VII. Conclusions

Affective computing has found broad applicability in many decision making domains, including health and job performance evaluation, financial and employment scrutiny, etc.

Hence, guaranteeing model fairness in affective computing is an open challenge for real-world applications. In this work, we investigated how heterogeneous behavioral patterns can impact the fairness of model outcomes. We proposed a method to identify heterogeneous features based on *multi-layer factor analysis* (MLFA), which can also be combined with feature rescaling techniques to mitigate the unfair impact of heterogeneity when sensitive attributes are unobserved.

Experimental results on synthetic and real-world datasets show that our proposed method can improve both the accuracy and fairness of models compared to using the original datasets. Our method can in fact be used as a pre-processing step for different regression models.

There are a few ways forward to improve our method in the future. Currently, the method is applied to continuous variables, but it could be extended to categorical variables, which are often seen in behavioral data. In addition, the impact of heterogeneity on fairness in classification models should also to be investigated.

References

[1] A. Pantelopoulos and N. G. Bourbakis, "A survey on wearable sensor-based systems for health monitoring and prognosis," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 1, pp. 1–12, 2009.

[2] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell, "Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones," in *Proceedings of the 2014 international joint conference on pervasive and ubiquitous computing*. ACM, 2014, pp. 3–14.

[3] C. Hernández-Quevedo, A. M. Jones, N. Rice *et al.*, "Reporting bias and heterogeneity in self-assessed health. evidence from the british household panel survey," *Health, Econometrics and Data Group (HEDG) Working paper 05*, vol. 4, 2004.

[4] A. Wood, M. Rychlowska, and P. M. Niedenthal, "Heterogeneity of long-history migration predicts emotion recognition accuracy." *Emotion*, vol. 16, no. 4, p. 413, 2016.

[5] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2017.

[6] R. D. Riley, J. Ensor, K. I. Snell, T. P. Debray, D. G. Altman, K. G. Moons, and G. S. Collins, "External validation of clinical prediction models using big datasets from e-health records or ipd meta-analysis: opportunities and challenges," *bmj*, vol. 353, p. i3140, 2016.

[7] J. H. Dulebohn, R. B. Davison, S. A. Lee, D. E. Conlon, G. McNamara, and I. C. Sarinopoulos, "Gender differences in justice evaluations: Evidence from fmri." *J Appl Psychol*, vol. 101, 2016.

[8] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[9] C. H. Wagner, "Simpson's paradox in real life," *The American Statistician*, vol. 36, no. 1, pp. 46–48, 1982.

[10] K. Lerman, "Computational social scientist beware: Simpson's paradox in behavioral data," *Journal of Computational Social Science*, vol. 1, no. 1, pp. 49–58, 2018.

[11] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *arXiv preprint arXiv:1908.09635*, 2019.

[12] C. C. Fabris and A. A. Freitas, "Discovering surprising patterns by detecting occurrences of simpson's paradox," in *Research and Development in Intelligent Systems XVI*. Springer, 2000, pp. 148–160.

[13] N. Alipourfard, P. G. Fennell, and K. Lerman, "Using simpson's paradox to discover interesting patterns in behavioral data," in *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[14] P. G. Fennell, Z. Zuo, and K. Lerman, "Predicting and explaining behavioral data with structured feature space decomposition," *EPJ Data Science*, vol. 8, no. 1, p. 23, 2019.

[15] M. N. Elliott, A. Fremont, P. A. Morrison, P. Pantoja, and N. Lurie, "A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity," *Health services research*, vol. 43, no. 5p1, pp. 1722–1736, 2008.

[16] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 2018, pp. 335–340.

[17] M. P. Kim, A. Ghorbani, and J. Zou, "Multiaccuracy: Black-box post-processing for fairness in classification," in *Proceedings of the 2019 Conference on AI, Ethics, and Society*. ACM, 2019, pp. 247–254.

[18] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth, "A convex framework for fair regression," *arXiv preprint arXiv:1706.02409*, 2017.

[19] J. Komiyama, A. Takeda, J. Honda, and H. Shimao, "Nonconvex optimization for regression with fairness constraints," in *International Conference on Machine Learning*, 2018, pp. 2742–2751.

[20] A. Agarwal, M. Dudik, and Z. S. Wu, "Fair regression: Quantitative definitions and reduction-based algorithms," in *International Conference on Machine Learning*, 2019, pp. 120–129.

[21] S. Hajian and J. Domingo-Ferrer, "A methodology for direct and indirect discrimination prevention in data mining," *IEEE transactions on knowledge and data engineering*, vol. 25, pp. 1445–1459, 2012.

[22] M. Gupta, A. Cotter, M. M. Fard, and S. Wang, "Proxy fairness," *arXiv preprint arXiv:1806.11212*, 2018.

[23] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in *KDD*, 2017.

[24] F. Hamidi, M. K. Scheuerman, and S. M. Branham, "Gender recognition or gender reductionism? the social implications of embedded gender recognition systems," in *CHI '18*, 2018, pp. 1–13.

[25] B. Fish, J. Kun, and Á. D. Lelkes, "A confidence-based approach for balancing fairness and accuracy," in *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 2016, pp. 144–152.

[26] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[27] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 259–268.

[28] B. M. Booth, K. Mundnich, T. Feng, A. Nadarajan, T. H. Falk, J. L. Villatte, E. Ferrara, and S. Narayanan, "Multimodal human and environmental sensing for longitudinal behavioral studies in naturalistic settings: Framework for sensor selection, deployment, and management," *Journal of medical Internet research*, vol. 21, no. 8, p. e12832, 2019.

[29] K. Mundnich, B. M. Booth, M. l'Hommedieu, T. Feng, B. Girault, J. L'hommedieu, M. Wildman, S. Skaaden, A. Nadarajan, J. L. Villatte *et al.*, "Tiles-2018, a longitudinal physiologic and behavioral data set of hospital workers," *Scientific Data*, vol. 7, no. 1, pp. 1–26, 2020.

[30] U. Ramnath, L. Rauch, E. Lambert, and T. Kolbe-Alexander, "The relationship between functional status, physical fitness and cognitive performance in physically active older adults: A pilot study," *PloS one*, vol. 13, no. 4, p. e0194918, 2018.

[31] J. R. Stroop, "Studies of interference in serial verbal reactions." *Journal of experimental psychology*, vol. 18, no. 6, p. 643, 1935.