Finding Interpretable Word Embedding Subspaces using Covariance and Correlation Maximization

Anonymous ACL submission

Abstract

This paper proposes a new method for estimating a direction in a word embedding space corresponding to an interpretable semantic property such as gender, race, or religion. Our technique assumes that words can be assigned numerical scores that quantify their association with the target property. We estimate the subspace by maximizing the covariance or correlation of these scores with the projection of word embeddings along the subspace. Using our technique, we show that word embedding spaces in English, French, and Chinese contain subspaces that encode gender, race, religion, sentiment, word length, and national population. We then apply our technique to the mitigation of gender and racial bias from word embeddings. We find that using our technique to estimate a gender or race subspace improves performance on several benchmarks.

1 Introduction

One of the most famous empirical results in natural language processing is the discovery that a range of semantic properties are encoded by distinguished *interpretable subspaces* of the word embedding space (Mikolov et al., 2013a; Rothe and Schütze, 2016; Jang and Myaeng, 2017; Arora et al., 2018; Şenel et al., 2018; Shin et al., 2018; Ethayarajh et al., 2019a). A simple way to probe the structure of these subspaces is through linear analogies: letting $[\![w]\!] \in \mathbb{R}^d$ denote the embedding of a word w, we expect words participating in an analogy such as *king* : *queen* :: *man* : *woman* to exhibit the relation $[\![\text{king}]\!] - [\![\text{queen}]\!] \approx [\![\text{man}]\!] - [\![\text{woman}]\!] \approx \boldsymbol{g}$, where the subspace $\text{span}(\boldsymbol{g})$ represents the concept that relates the word pairs. A rich body of literature on the interpretation of word embedding spaces has identified subspaces corresponding to syntactic and semantic features (Mikolov et al., 2013b; Baroni et al., 2014), quantificational features (Linzen et al., 2016), and specific lexical properties such as national capitals (Mikolov et al., 2013a) and gender and ethnic stereotypes (Bolukbasi et al., 2016; Manzini et al., 2019).

Linear analogies provide a simple and intuitive method for intrinsic evaluation of word embeddings by validating the existence of interpretable subspaces (Yaghoobzadeh and Schütze, 2016). But recent techniques in NLP, particularly in social bias mitigation (Bolukbasi et al., 2016; Zhao et al., 2018; Ravfogel et al., 2020), require not only that interpretable subspaces exist, but also that a basis for these subspaces can be precisely estimated. Unfortunately, it is difficult to estimate subspaces using analogy-based methods because of the requirement that words be paired. This requirement is difficult to satisfy in domains such as race where there is no obvious way to define word pairs. In domains that are more amenable to analogies, the labor intensity of constructing word pairs limits the amount of data that can be used in the estimation of an interpretable subspace. For example, Bolukbasi et al.'s (2016) estimation of the gender subspace only uses ten word pairs.

This paper presents two novel algorithms for identifying interpretable embedding subspaces, with particular focus on applications to bias mitigation. In contrast to previous methods, our approach does not require word pairs or manually crafted sets of words designed to capture some semantic concept. Instead, we assume that properties encoded by embedding subspaces are numerically valued, and that these values can be measured empricially through human judgments, public datasets, or world knowledge databases. Our two methods are therefore applicable to any property for which each word can be assigned a numerical value, and they can incorporate large, existing sets of labeled words at little to no additional annotation cost. Given a corpus of words annotated with numerical scores, our first method, *covariance maximization* (MaxCov), estimates an interpretable

subspace by maximizing the covariance between the scores and the projections of word embeddings to this subspace. Our second method, *correlation maximization* (MaxCorr), is similar to MaxCov, but maximizes correlation instead of covariance.

After introducing our two algorithms, we test them using two sets of experiments. First, we conduct an intrinsic evaluation of MaxCov and MaxCorr by attempting to find interpretable subspaces that represent gender, race, religion, sentiment, word length, and national population in English, French, and Chinese. In almost all cases, we are able to identify subspaces that correlate with our data with $\rho > .6$. Then, we apply our method to the downstream task of social bias mitigation, and show that using MaxCov or MaxCorr to identify the gender or race subspace improves performance on several benchmarks.

2 Related Work

Interpretable embedding subspaces play an important role in *projection-based debiasing*,[1] a two-step pipeline for bias mitigation proposed by Bolukbasi et al. (2016). The first step is to estimate an interpretable subspace encoding a social attribute such as gender. Then, word embeddings are "debiased" by surgically altering their projections onto this subspace. The contribution of the present paper is to improve upon the first step of this process.

Estimating Gender Spaces. Most work on projection-based debiasing focuses on removing gender bias from English-language embeddings. The simplest conceivable method for finding a "gender subspace" is to use the difference vector $\text{span}(\llbracket\text{she}\rrbracket - \llbracket\text{he}\rrbracket)$. This method is used by Dev et al. (2020). Generalizations of this approach are used by Ethayarajh et al. (2019b), who take several difference vectors to form a basis for a multidimensional subspace, and Dev and Phillips (2019), who use the first principal componenent of a set of difference vectors as a one-dimensional gender basis.

Bolukbasi et al. (2016) use a method similiar to Dev and Phillips (2019); but instead of using difference vectors, they take pairs of word embeddings and center them around the origin. Another method involves using the weight vector of a support vector machine (Ravfogel et al., 2020). The method most similar to ours is the *DensRay* algorithm proposed by Dufter and Schütze (2019). Like our approach, DensRay employs maximization, but treats gender as a binary rather than continuous variable. Their objective maximizes the distance between opposite-gender words and minimizes the distance between same-gender words along the gender subspace.

Other Properties and Languages. Bolukbasi et al.'s (2016) method has been generalized to removing gender bias from Swedish embeddings (Sahlgren and Olsson, 2019) as well as racial and religious bias from English embeddings (Manzini et al., 2019). Ravfogel et al. (2020) also explore removing racial bias, but at the sequence level rather than the word level.

3 Estimating Interpretable Subspaces

Our approach to estimating interpretable word embedding subspaces assumes that we have access to a set of *reference words* $\mathbb{W}$, such that each word $w \in \mathbb{W}$ is associated with a *score* $s(w)$ along some semantic dimension. Our goal is to find a unit vector $\boldsymbol{g}$ such that for each $w \in \mathbb{W}$, $\boldsymbol{g}^\top \llbracket w \rrbracket \approx s(w)$. In MaxCov, we choose $\boldsymbol{g}$ to be the vector that maximizes the covariance between $\boldsymbol{g}^\top \llbracket w \rrbracket$ and $s(w)$. MaxCorr works similarly, except we maximize correlation instead of covariance. We show that these two methods can be implemented straightforwardly using efficient algorithms.

3.1 Covariance Maximization

Formally, MaxCov estimates $\boldsymbol{g}$ as follows:

$$\boldsymbol{g} = \underset{\|\boldsymbol{v}\|=1}{\text{argmax}} \underset{w \in \mathbb{W}}{\text{cov}}(\boldsymbol{v}^\top \llbracket w \rrbracket, s(w)).$$

This method is similar to PCA, except that instead of finding the direction of greatest variance in the embeddings, we find the direction of *greatest covariance* with the scores assigned to the reference words. It turns out that MaxCov is computed by the formula $\boldsymbol{g} = \boldsymbol{a}/\|\boldsymbol{a}\|$, where

$$\boldsymbol{a} = \sum_{w \in \mathbb{W}} (s(w) - \overline{s})(\llbracket w \rrbracket - \overline{\boldsymbol{w}})$$

and the variables $\overline{s} = \text{mean}_{w \in \mathbb{W}}(s(w))$ and $\overline{\boldsymbol{w}} = \text{mean}_{w \in \mathbb{W}}(\llbracket w \rrbracket)$ denote the average score and average embedding of the reference words, respectively. We derive this formula in Appendix A.1.

3.2 Correlation Maximization

In MaxCorr, $\boldsymbol{g}$ is estimated as

$$\boldsymbol{g} = \underset{\|\boldsymbol{v}\|=1}{\text{argmax}} \underset{w \in \mathbb{W}}{\text{corr}}(\boldsymbol{v}^\top \llbracket w \rrbracket, s(w)).$$

[1] This term is due to Stańczak and Augenstein (2021).

To compute MaxCorr, we fit a linear regression model

$$s(w) = \boldsymbol{a}^\top [\![w]\!] + b,$$

and take $\boldsymbol{g} = \boldsymbol{a}/\|\boldsymbol{a}\|$. The validity of this approach is proven in Appendix A.2, where we verify formally that $\boldsymbol{a}/\|\boldsymbol{a}\|$ is indeed the direction of greatest correlation with $s(w)$.

4 Exploring Embedding Subspaces

We begin by using MaxCov and MaxCorr to determine what kinds of continuous properties are represented in embedding spaces by an interpretable subspace. In this experiment, we search for subspaces encoding information about gender, race, religion, and sentiment, as well as population and orthography. We apply our method to word embeddings in English, French, and Chinese.

4.1 Data

To fit an interpretable subspace, we obtain reference words and scores from publicly available datasets. These datasets are enumerated in Table 1, which identifies each dataset by an abbreviated name. We use three different kinds of data for extracting scores.

Human Judgments. Human judgment studies from psychology, social science, and behavioral science provide a direct measure of stereotypical associations between words and semantic properties. For this experiment, we use human judgment data for gender (female vs. male), race (African American vs. European American), and sentiment (positive valence vs. negative valence). All scores were elicited from participants using a Likert scale, with the exception of Mo18-S (Mohammad, 2018), which elicited valence rankings that were then interpolated using best–worst scaling.

Frequencies. Certain words, such as personal names or country names, are associated with particular social identities. For example, most people named *Mary* are female, while most people named *John* are male. We leverage demographic statistics in order to extract scores for gender, race, and religion. For gender, we use census data on given names from the United States Social Security Administration (SSA, 2019), the French National Institute of Statistics and Economic Studies (INSEE, 2019), and the Chinese National Citizen Identity Information Center[2] (NCIIC, Bao, 2021). For race, we use data on surnames from the 2010 United States Census (United States Census Bureau, 2021), which reports the frequencies among six racial categories of the 1,000 most common names. For religion, we use statistics from the Pew Research Center (2012) on the religious composition of 233 countries and territories.

For the SSA, INSEE, and NCIIC data, we convert the reported frequency counts into a gender rating by estimating the probability that a person with a given name is female according to Laplace's (1814) rule of succession:

$$s(w) = \frac{\#\text{ female individuals named } w + 1}{\#\text{ individuals named } w + 2}.$$

Since the United States Census and Pew Research data do not report exact counts for each demographic group, we instead use percentages reported in those datasets, which are precise to one-tenth of a percentage point.

Counts. Our data on national population come from World Bank Open Data (2022). For orthographic word length, we simply compute the length of the 1,000 most frequent words that have not been filtered out from our word embedding spaces (see Subsection 4.2).[3] We define the length of an English or French word to be the number of characters in that word; we define the length of a Chinese word to be the total combined stroke count of all characters in that word. Unlike the datasets based on frequencies, we directly use the scores reported in these datasets without converting them into percentages.

4.2 Procedure

The goal of this experiment is to determine the extent to which semantic properties are represented by interpretable embedding subspaces that can be discovered by MaxCov and MaxCorr. We fit an interpretable subspace with MaxCov and MaxCorr using 75% of each dataset in Table 1, and measure how well the subspace predicts the scores assigned to the remaining 25%. We measure the quality of an interpretable subspace $\mathrm{span}(\boldsymbol{g})$ via the correlation between $s(w)$ and $\boldsymbol{g}^\top [\![w]\!]$.

[2] Since Chinese given names are unique to the individual, the NCIIC dataset does not report frequency of given names *per se*, but rather the number of occurrences of individual Chinese characters in given names assigned to men and women.

[3] The word length data are not listed in Table 1.

Name	Source	*N*	Property	Word Type	Locale	Type	Range
KT03-G	Kennison and Trofe (2003)	232	Gender	Professions	EN-US	Judgments	1–7
Ga08-G	Gabriel et al. (2008)	127	Gender	Professions	EN-GB	Judgments	0–100
Ga08-G	Gabriel et al. (2008)	127	Gender	Professions	FR-CH	Judgments	0–100
SKB19-G	Scott et al. (2019)	5,553	Gender	Miscellaneous	EN-GB	Judgments	1–7
SSA-G	SSA (1880–2019)	99,444	Gender	Given Names	EN-US	Frequencies	0–1
INSEE-G	INSEE (1900–2019)	35,010	Gender	Given Names	FR-FR	Frequencies	0–1
Ba21-G	Bao (2021)	2,614	Gender	Given Names	ZH-CN	Frequencies	0–1
Census-Ra	US Census (2010)	1,000	Race	Surnames	EN-US	Frequencies	0–1
SD18-Ra	Stelter and Degner (2018)	159	Race	Given Names	EN-US	Judgments	1–7
Pew-Re	Pew Research Center (2010)	233	Religion	Countries	N/A	Frequencies	0–1
SKB19-S	Scott et al. (2019)	5,553	Sentiment	Miscellaneous	EN-GB	Judgments	1–9
Mo18-S	Mohammad (2018)	19,971	Sentiment	Miscellaneous	EN-CA	Judgments	0–1
Gi12-S	Gilet et al. (2012)	835	Sentiment	Miscellaneous	FR-FR	Judgments	1–7
Ba21-S	Bao (2021)	2,614	Sentiment	Given Names	ZH-CN	Judgments	1–5
Ya17-S	Yao et al. (2017)	1,100	Sentiment	Miscellaneous	ZH-CN	Judgments	1–9
Population	World Bank Open Data (2022)	217	Population	Countries	N/A	Counts	N/A

Table 1: Datasets used to assign scores to words. "*N*" denotes the total number of words (including compounds) provided by each dataset. For human judgment and frequency count datasets, "Locale" denotes the language the words are presented in and the country where the data were elicited; for datasets involving countries, we translate country names into target languages, using single-token names whenever possible. For datasets from the SSA, INSEE, US Census, and the Pew Research Center, dates denote the time period over which data were collected.

Embeddings. We use 300-dimensional word embeddings for all three languages. We use GloVe embeddings trained on the 42-billion-token Common Crawl corpus (Pennington et al., 2014) for English and fastText embeddings (Grave et al., 2018) for French. For Chinese, we use embeddings provided by Li et al. (2018), which are trained using Skip-gram with negative sampling (Mikolov et al., 2013b) with character-level features (Chen et al., 2015) on the Mixed-large dataset.

Preprocessing. We filter out all English and French words containing non-alphanumeric characters, as well as Chinese words containing non-Chinese characters. We then filter out all but the 50,000 most frequent words in each language before normalizing the embeddings to unit length.

Unlike the English GloVe embeddings, the French fastText embeddings are case-sensitive. For datasets based on given names and country names, we capitalize each word according to orthographic conventions in French; for datasets based on common nouns, we consider both capitalized and all-lowercase versions of each word. For word length data, we follow the capitalization used in the word embeddings.

Validation. To account for the possibility of overfitting to the reference words, we perform 4-fold cross validation and report the mean result obtained across the folds. We measure the significance of our results using a two-sided permutation test in which the experiment is repeated 1,000 times with the scores randomly shuffled for each dataset. We use the average result of the permutation test as a baseline for comparison.

4.3 Results

The results of our exploration are shown in Table 2. Baseline values from the permutation test range from $-.008$ to .005 for both MaxCov and MaxCorr for all three languages. All results are statistically significant, with $p \leq .008$.

In most cases, we are able to find an interpretable embedding subspace that robustly encodes the scores given by our datasets. The only exception is the Ya17-S sentiment dataset, for which MaxCov achieves a cross-validated correlation of only .297. Two other datasets with weaker results are the Unaffiliated religion data from Pew-Re and the Native American race data from Census-Ra. The high quality of the interpretable subspace for word length relative to the permutation test baseline seems surprising at first glance, given that none of the three word embedding models contain features that explicitly encode this information. However, this result is explained by the observation that word length is inversely correlated with frequency and other statistics, which are accessible to word embedding models during training (Zipf, 1936; Piantadosi et al., 2011).

Between MaxCov and MaxCorr, it does not appear that either algorithm consistently outperforms

Dataset	*N*	MaxCov	MaxCorr
English			
KT03-G	202	**.678**	.675
Ga08-G	81	.465	**.605**
SKB19-G	4,517	.707	**.793**
SSA-G	7,949	.693	**.745**
SD18-Ra	117	**.834**	.826
Census-Ra (White)	924	.827	**.865**
Census-Ra (Black)	924	**.660**	.635
Census-Ra (Asian)	924	.793	**.844**
Census-Ra (Native)	924	**.483**	.410
Census-Ra (Multi.)	924	**.786**	.726
Census-Ra (Latino)	924	**.917**	.916
Pew-Re (Christian)	194	.773	**.817**
Pew-Re (Muslim)	194	.776	**.819**
Pew-Re (Unaff.)	194	**.539**	.453
SKB19-S	4,517	.756	**.852**
Mo18-S	16,834	.671	**.776**
Population	187	**.523**	.409
Word Length	1,000	.607	**.622**
French			
Ga08-G	96	.737	**.758**
INSEE-G	3,139	**.720**	.276
Pew-Re (Christian)	183	.682	**.726**
Pew-Re (Muslim)	183	.619	**.642**
Pew-Re (Unaff.)	183	**.410**	.396
Gi12-S	530	**.820**	.534
Population	174	**.434**	.431
Word Length	1,000	.676	**.691**
Chinese			
Ba21-G	1,798	.593	**.596**
Pew-Re (Christian)	183	**.753**	.663
Pew-Re (Muslim)	183	**.755**	.738
Pew-Re (Unaff.)	183	**.564**	.394
Ba21-S	1,798	.672	**.746**
Ya17-S	914	**.297**	.116
Population	181	**.576**	.470
Word Length	1,000	.563	**.638**

Table 2: Results for the embedding subspace exploration experiment. "*N*" represents the number of words in each dataset for which an embedding is available after filtering.

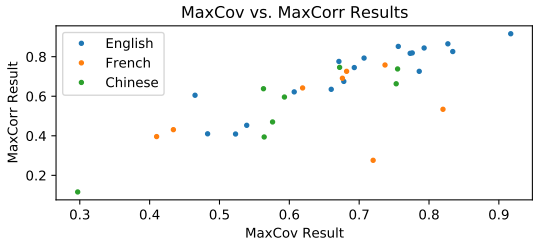

Figure 1: The relationship between MaxCov and MaxCorr results ($\rho = .902$ for English, $\rho = .371$ for French, and $\rho = .908$ for Chinese). The two French outliers represent the INSEE-G and Gi12-S datasets.

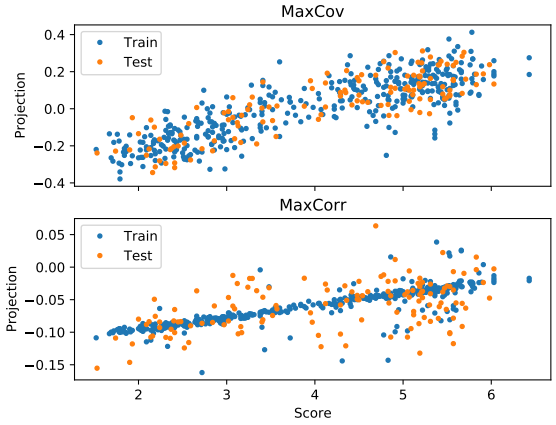

Figure 2: MaxCorr overfitting on Gi12-S. Plots compare scores from the dataset with the projections of embeddings onto the interpretable subspace.

the other. Figure 1 shows, in fact, that the two algorithms are strongly correlated with one another in performance. The two outliers in this plot represent the INSEE-G and Gi12-S datasets. In both cases, MaxCorr resulted in significant overfitting, as illustrated in Figure 2.

5 Social Bias Mitigation

In this second experiment, we apply MaxCov and MaxCorr to the task of removing social bias from word embeddings. Our goal is to assess the suitability of interpretable subspaces estimated using MaxCov and MaxCorr for this task, compared to existing methods. We do this by replicating a number of different evaluations of bias from the literature and observing the extent to which each subspace estimation method reduces the amount of bias measured. Our evaluations are implemented using the WEFE libary (Badilla et al., 2020).

5.1 Experimental Setup

In each of the following analyses, we compare the pre-trained embeddings from the previous section with embeddings that have been debiased according to an interpretable subspace identified using MaxCov, MaxCorr, or a baseline method. For all subspace estimation methods, debiasing is accomplished using the *strong debiasing* technique proposed by Prost et al. (2019). Let $\text{span}(\mathbb{B})$ be an interpretable subspace, where $\mathbb{B}$ is orthonormal. The strongly debiased version of a word embedding $[\![w]\!]$ with respect to $\mathbb{B}$ is given by

$$\boldsymbol{w} = [\![w]\!] - \sum_{\boldsymbol{b} \in \mathbb{B}} \text{proj}_{\boldsymbol{b}}([\![w]\!]),$$

where $\text{proj}_{\boldsymbol{b}}([\![w]\!])$ is the projection of $[\![w]\!]$ onto $\boldsymbol{b}$.

5.2 Baselines

For comparison, we consider three baseline methods of estimating interpretable subspaces for bias mitigation.

Tuples. The *Tuples method* of Bolukbasi et al. (2016) uses a combination of several word pairs such as *she–he* in order to estimate a subspace. We use a generalization of this method by Manzini et al. (2019), which allows for word tuples of arbitrary length such as *African–Caucasian–Asian* in order to compute subspaces for multivalent properties such as race or religion. Given a set $\mathbb{T} \subseteq \mathbb{W}^n$ of n-tuples of words, a basis $\mathbb{B}$ for a k-dimensional interpretable subspace is estimated by taking the first k principal components of the set

$$\bigcup_{(w_1,w_2,\ldots,w_n)\in\mathbb{T}} \left\{ [\![w_i]\!] - \operatorname*{mean}_{1\le j\le n}([\![w_j]\!]) \right\}_{i=1}^{n},$$

normalized to unit length.

SVM. The *SVM method* of Ravfogel et al. (2020) estimates an interpretable subspace $\boldsymbol{g} = \boldsymbol{a}/\|\boldsymbol{a}\|$ by taking $\boldsymbol{a}$ to be the weight vector of a linear support vector machine (SVM) that classifies word embeddings into one or more semantic categories. This makes $\boldsymbol{g}$ orthogonal to the hyperplane that separates the categories. When debiasing embeddings with respect to gender, Ravfogel et al. (2020) construct the training dataset for the SVM by taking the word embeddings with the 7,500 highest and 7,500 lowest values in the $[\![\text{she}]\!] - [\![\text{he}]\!]$ direction.

DensRay. Like our approach, the *DensRay method* (Dufter and Schütze, 2019) uses a set of words $\mathbb{W}$ accompanied by scores $s(w)$. The interpretable subspace is estimated by minimizing

$$\boldsymbol{g} = \operatorname*{argmin}_{\|\boldsymbol{v}\|=1} \sum_{u,w\in\mathbb{W}} s(u)s(w) \left\| \boldsymbol{v}^\top([\![u]\!] - [\![w]\!]) \right\|^2$$

where for all $w \in \mathbb{W}$, $s(w)$ is either 1 or -1. For example, if $\boldsymbol{g}$ represents gender, then $s(\text{he}) = 1$ and $s(\text{she}) = -1$. Intuitively, DensRay attempts to find the subspace that maximizes similarity between same-category words while minimizing similarity between opposite-category words.

Appendix B describes implementational details for MaxCov, MaxCorr, and the three baseline methods, including word lists and datasets used to estimate each subspace.

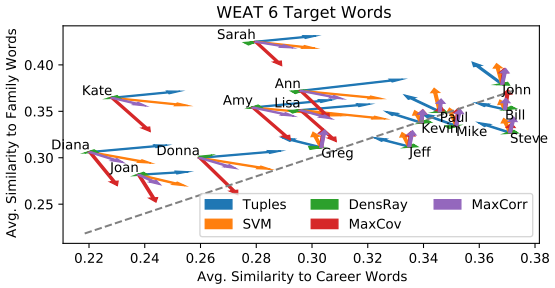

Figure 3: Target words from the English WEAT 6 test, visualized according to their mean cosine similarity to the two sets of attribute words. Arrows connect undebiased word embeddings to debiased ones.

5.3 Word Embedding Association Test

The *Word Embedding Association Test* (WEAT, Caliskan et al., 2017) compares two sets of *target words* $\mathbb{T}_1, \mathbb{T}_2$ in terms of their distance to two sets of *attribute words* $\mathbb{A}_1, \mathbb{A}_2$. For a target word $w \in \mathbb{T}_1 \cup \mathbb{T}_2$, the quantity

$$s_{\mathbb{A}_1,\mathbb{A}_2}(w) = \operatorname*{mean}_{a\in\mathbb{A}_1}(\cos([\![w]\!],[\![a]\!])) - \operatorname*{mean}_{b\in\mathbb{A}_2}(\cos([\![w]\!],[\![b]\!]))$$

measures the degree to which w is, on average, closer to words in $\mathbb{A}_1$ than $\mathbb{A}_2$. For example, if $\mathbb{A}_1 = \{\text{he, him, male}\}$ and $\mathbb{A}_2 = \{\text{she, her, female}\}$, then $s_{\mathbb{A}_1,\mathbb{A}_2}(\text{science}) > 0$ means that the word *science* is closer to "male" words than "female" words. From this individual bias metric, an *effect size* is computed by aggregating bias measures across all target words:

$$\frac{\operatorname{mean}_{u\in\mathbb{T}_1}(s_{\mathbb{A}_1,\mathbb{A}_2}(u)) - \operatorname{mean}_{v\in\mathbb{T}_2}(s_{\mathbb{A}_1,\mathbb{A}_2}(v))}{\operatorname{stdev}_{w\in\mathbb{T}_1\cup\mathbb{T}_2}(s_{\mathbb{A}_1,\mathbb{A}_2}(w))}.$$

A positive effect size indicates that $\mathbb{T}_1$ has a greater bias towards $\mathbb{A}_1$ and against $\mathbb{A}_2$ than $\mathbb{T}_2$ does. A negative effect size is interpreted analogously. An effect size of 0 indicates lack of bias.

Results. Table 3 shows effect sizes computed for the six WEAT tests from Caliskan et al. (2017) that measure bias in terms of binary race and binary gender. We also run the French- and Chinese-language WEAT tests from Kurpicz-Briki (2020) and Jiao (2021), respectively. In almost all cases for English, strong debiasing with MaxCov is the most effective method for debiasing, while MaxCorr is more effective for French. Note in particular that the baseline methods struggle when applied to race or to French-language word embeddings, and that

	Targets	Attributes	None	Tuples	SVM	DensRay	MaxCov	MaxCorr
English (Caliskan et al., 2017)								
3	White vs. Black	Pleasant vs. Unpleasant	1.40	1.40	1.40	1.40	**.96**	1.26
4	White vs. Black	Pleasant vs. Unpleasant	1.51	1.52	1.47	1.52	**.79**	1.30
5	White vs. Black	Pleasant vs. Unpleasant	1.37	1.37	1.35	1.37	**.25**	1.19
6	Male vs. Female	Career vs. Family	1.69	1.23	1.06	1.69	**.97**	1.38
7	Math vs. Arts	Male vs. Female	1.50	.33	.11	1.53	**.06**	.27
8	Science vs. Arts	Male vs. Female	1.05	−1.36	−.53	1.06	−.52	**−.22**
French (Kurpicz-Briki, 2020)								
6-fr1	Male vs. Female	Career vs. Family	.77	1.05	.90	.80	.84	**.43**
6-fr2	Male vs. Female	Career vs. Family	1.07	1.25	1.18	1.07	1.11	**.85**
7-fr	Math vs. Arts	Male vs. Female	.64	1.52	.61	.68	**.49**	.53
8-fr	Science vs. Arts	Male vs. Female	−.33	**.18**	−.35	−.33	−.54	−.45
Chinese (Jiao, 2021)								
7	Math vs. Arts	Male vs. Female	1.49	**.63**	1.49	1.50	1.07	1.22
8	Science vs. Arts	Male vs. Female	1.04	.67	1.08	1.05	.11	**.09**

Table 3: Effect sizes from the WEAT test (closer to 0 is better), calculated from embedding spaces debiased using various methods. The leftmost column shows identifiers assigned to individual WEAT tests by Caliskan et al. (2017), Kurpicz-Briki (2020), and Jiao (2021).

	Female	Asian	Hispanic
None	.474	.098	**.004**
Tuples	.315	.094	.006
SVM	.385	.096	.020
DensRay	.476	.103	.007
MaxCov	.326	**.078**	.011
MaxCorr	**.283**	.112	.015

Table 4: R^2 values from the occupation statistics analysis (lower is better).

the SVM-based method additionally struggles with Chinese-language embeddings. In contrast, MaxCov and MaxCorr are able to reduce bias in almost every case, regardless of language or semantic property.

Figure 3 shows visually that debiasing has a greater effect on "female" words than on "male" words. As indicated by the angles of the arrows, debiasing methods differ in terms of the extent to which they make target words more similar to one of the two attributes versus the other.

5.4 Relation to Occupation Statistics

Next, we apply an analysis due to Garg et al. (2018), which relates gender and racial bias in word embeddings with occupational statistics. The analysis tests the hypothesis that gender and racial bias measured in word embeddings linearly predict the demographics of various professions.

Let $1, 2, \ldots, n$ be a collection of demographic groups, represented by attribute word sets $\mathbb{A}_1, \mathbb{A}_2, \ldots, \mathbb{A}_n$. Assume that group n is designated as the "unmarked" group (males for gender and Whites for race). Given a profession word p, let p_i be the percentage of workers in p belonging to group i. The bias of the word embedding $[\![p]\!]$ with respect to group i is measured by *relative norm distance* (RND), defined as

$$\mathrm{rnd}_i(p) = \operatorname*{mean}_{j \neq i} \left(\|[\![p]\!] - \overline{\boldsymbol{a}}_j\| \right) - \|[\![p]\!] - \overline{\boldsymbol{a}}_i\|$$

where $\overline{\boldsymbol{a}}_i = \mathrm{mean}_{a \in \mathbb{A}_i}([\![a]\!])$ for all i. RND is similar to the quantity $s_{\mathbb{A}_1, \mathbb{A}_2}(p)$ from Caliskan et al. (2017), except that it uses Euclidean distance instead of cosine similarity and compares a single target word with several attribute word sets. The *occupational bias* of group i in profession p is measured by

$$\text{occ-bias}_{i;n}(p) = \frac{p_i - p_n}{p_i + p_n}.$$

Given a set of professions, we compute a linear regression between occupation bias and RND score for each non-majority group (i.e., not male or White), and report the R^2 value of the regression. If R^2 is close to 0, then RND and occupation bias are not linearly related, indicating that bias measured by RND does not reflect bias empirically measured through demographics.

Setup. Following Garg et al. (2018), we conduct two versions of the experiment. In the first run, we use two groups: female and male; and in the second run, we use three groups: Asian, Hispanic, and White. Gender debiasing is implemented in the same way as in the WEAT test. For race debiasing, we use a 3-dimensional embedding subspace

	Gender		Race
	Original	WEAT 6–8	WEAT 3–5
None	.898	.869	.584
Tuples	**1.000**	**.988**	.589
SVM	.987	.978	.567
DensRay	.897	.869	.581
MaxCov	.945	.940	**.825**
MaxCorr	.976	.965	.627

Table 5: Spearman correlations from the ECT test (higher is better).

representing all three racial categories.[4] We use occupation statistics from 2015, which are the most recent data provided by Garg et al. (2018).

Results. The results are reported in Table 4. Prior to debiasing, we measure a high R^2 value for female bias and lower R^2 values for racial bias, with RND almost completely uncorrelated with occupational bias for the Hispanic group. Among the five debiasing methods, MaxCorr is most effective in removing female bias, while MaxCov is most effective in removing Asian bias. For the Hispanic group, most debiasing methods worsen the R^2 value rather than improving it. This suggests that debiasing methods may unintentionally introduce bias if it is not already detected in the embedding space.

5.5 Embedding Coherence Test

Finally, we subject our subspace estimation methods to the *embedding coherence test* (ECT, Dev and Phillips, 2019). Whereas the WEAT test and the occupation statistics analysis measure bias by the *distance* of target words to attribute words, the ECT test instead considers the *ranking* of target words in terms of their distance to attribute words.

In the ECT test, we are given sets of attribute words $\mathbb{A}_1$ and $\mathbb{A}_2$ along with a set of profession words $\mathbb{P}$. Two rankings of the profession words are computed, based on $\cos(\text{mean}_{w \in \mathbb{A}_i}(\llbracket w \rrbracket), \llbracket p \rrbracket)$ for $p \in \mathbb{P}$ and $i \in \{1, 2\}$. Bias is measured by the Spearman correlation between the two rankings.

Setup. We replicate the original conditions of Dev and Phillips (2019), which measures gender bias using the list of profession words and the gendered word pairs from Bolukbasi et al. (2016). This setup gives an unfair advantage to the Tuples method, however, because the gendered word pairs are the same as the word pairs used in the Tuples method, meaning that the Tuples method estimates bias precisely with respect to the specific words used in the test. To alleviate this, we run a second ECT test for gender bias using the "male" and "female" words from WEAT 6–8 as attributes. To measure racial bias, we run a third ECT test using the "European American" and "African American" words from WEAT 3–5 as attributes.

Results. The results of the test are shown in Table 5. This time, more gender bias is measured before debiasing than racial bias. With the exception of DensRay, all subspace estimation methods perform remarkably well in improving ECT results for gender, with the Tuples method performing the best. Unsurprisingly, the Tuples method performs perfectly when its defining words are used as the attribute sets. On race, none of the estimation methods substantially impact ECT results, except for MaxCov and MaxCorr, with the former removing most of the measured bias.

6 Conclusion

Most approaches to the estimation of interpretable subspaces rely on the exploitation of lexical structures such as male–female word pairings. Unfortunately, such symmetries are only available for a small range of concepts: Black–White word pairs, for example, are not readily available in most languages. However, by treating semantic properties as continuous rather than binary, MaxCov and MaxCorr instead exploit the structure of the real numbers. As our experiments show, numerical scores can be obtained for a large number of reference words at a low cost and with great flexibility. The strength of our approach is especially impactful in racial debiasing, where previous methods suffer from a lack of lexical symmetries that encode race.

As Gonen and Goldberg (2019) point out, interpretable subspaces are not the only way for embedding spaces to encode semantic properties, and projection-based debiasing does not fully solve the problem of bias on its own. Nonetheless, our technique improves the quality of interpretable subspaces and expands their applicability to a greater range of properties. We plan to further explore these applications in future work.

References

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2018. Linear Algebraic Struc-

[4] Details are provided in Appendix B.

ture of Word Senses, with Applications to Polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495.

Pablo Badilla, Felipe Bravo-Marquez, and Jorge Pérez. 2020. WEFE: The Word Embeddings Fairness Evaluation Framework. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence Main Track*, volume 1, pages 430–436, Yokohama, Japan. International Joint Conferences on Artificial Intelligence Organizatio.

Han-Wu-Shuang Bao. 2021. ChineseNames: Chinese Name Database 1930-2008. Software Package ChineseNames, Comprehensive R Archive Network.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers, pages 238–247, Baltimore, MD, USA. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, volume 29, pages 4349–4357, Barcelona, Spain. Curran Associates, Inc.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Xinxiong Chen, Lei Xu, Zhiyuan Liu, Maosong Sun, and Huanbo Luan. 2015. Joint Learning of Character and Word Embeddings. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pages 1236–1242, Buenos Aires, Argentina. AAAI Press/International Joint Conferences on Artificial Intelligence.

Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. 2020. On Measuring and Mitigating Biased Inferences of Word Embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7659–7666.

Sunipa Dev and Jeff Phillips. 2019. Attenuating Bias in Word vectors. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 879–887, Naha, Japan. PMLR.

Philipp Dufter and Hinrich Schütze. 2019. Analytical Methods for Interpretable Ultradense Word Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1185–1191, Hong Kong, China. Association for Computational Linguistics.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019a. Towards Understanding Linear Word Analogies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3253–3262, Florence, Italy. Association for Computational Linguistics.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019b. Understanding Undesirable Word Embedding Associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. Association for Computational Linguistics.

Ute Gabriel, Pascal Gygax, Oriane Sarrasin, Alan Garnham, and Jane Oakhill. 2008. Au pairs are rarely male: Norms on the gender perception of role names across English, French, and German. *Behavior Research Methods*, 40(1):206–212.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

A. L. Gilet, D. Grühn, J. Studer, and G. Labouvie-Vief. 2012. Valence, arousal, and imagery ratings for 835 French attributes by young, middle-aged, and older adults: The French Emotional Evaluation List (FEEL). *European Review of Applied Psychology*, 62(3):173–181.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1 (Long and Short Papers), pages 609–614, Minneapolis, MN, USA. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3483–3487, Miyazaki, Japan. European Language Resources Association (ELRA).

Kyoung-Rok Jang and Sung-Hyon Myaeng. 2017. Elucidating Conceptual Properties from Word Embeddings. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and Their Applications*, pages 91–95, Valencia, Spain. Association for Computational Linguistics.

Meichun Jiao. 2021. Investigating Gender Bias in Word Embeddings for Chinese. Master's thesis, Uppsala University, Uppsala, Sweden, December.

Shelia M. Kennison and Jessie L. Trofe. 2003. Comprehending Pronouns: A Role for Word-Specific Gender Stereotype Information. *Journal of Psycholinguistic Research*, 32(3):355–378.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, LA, USA. Association for Computational Linguistics.

Mascha Kurpicz-Briki. 2020. Cultural Differences in Bias? Origin and Gender Bias in Pre-Trained German and French Word Embeddings. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*, volume 2624 of *CEUR Workshop Proceedings*, Online. Sun SITE Central Europe.

Pierre-Simon Laplace. 1814. *Essai philosophique sur les probabilités*, first edition. Mme. Ve. Courcier, Paris, France.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical Reasoning on Chinese Morphological and Semantic Relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143, Melbourne, Australia. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Benjamin Spector. 2016. Quantificational features in distributional word representations. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 1–11, Berlin, Germany. Association for Computational Linguistics.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, MN, USA. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. In *ICLR 2013 Workshop Proceedings*, Scottsdale, AZ. OpenReview.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, GA, USA. Association for Computational Linguistics.

Saif Mohammad. 2018. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

National Institute of Statistics and Economic Studies. 2019. Fichier des prénoms de 1900 à 2019. Dataset 5bf42c958b4c4144b0110ce8, Direction interministérielle du numérique, Paris, France.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Pew Research Center. 2012. Table: Religious Composition by Country, in Percentages. Dataset, Pew Research Center, Washington, DC, USA.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Flavien Prost, Nithum Thain, and Tolga Bolukbasi. 2019. Debiasing Embeddings for Reduced Gender Bias in Text Classification. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 69–75, Florence, Italy. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Sascha Rothe and Hinrich Schütze. 2016. Word Embedding Calculus in Meaningful Ultradense Subspaces. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 512–517, Berlin, Germany. Association for Computational Linguistics.

Magnus Sahlgren and Fredrik Olsson. 2019. Gender Bias in Pretrained Swedish Embeddings. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 35–43, Turku, Finland. Linköping University Electronic Press.

Graham G. Scott, Anne Keitel, Marc Becirspahic, Bo Yao, and Sara C. Sereno. 2019. The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, 51(3):1258–1270.

Lütfi Kerem Şenel, İhsan Utlu, Veysel Yücesoy, Aykut Koç, and Tolga Çukur. 2018. Semantic Structure and Interpretability of Word Embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1769–1779.

Jamin Shin, Andrea Madotto, and Pascale Fung. 2018. Interpreting Word Embeddings with Eigenvector Analysis. In *Interpretability and Robustness in Audio, Speech, and Language (NIPS 2018 Workshop)*, Montreal, Canada. OpenReview.

Karolina Stańczak and Isabelle Augenstein. 2021. A Survey on Gender Bias in Natural Language Processing. *Computing Research Repository*, abs/2112.14168v1.

Marleen Stelter and Juliane Degner. 2018. Recognizing Emily and Latisha: Inconsistent Effects of Name Stereotypicality on the Other-Race Effect. *Frontiers in Psychology*, 9:486.

United States Census Bureau. 2021. Frequently Occurring Surnames from the 2010 Census. Dataset, United States Census Bureau, Suitland, MD, USA.

United States Social Security Administration. 2019. Baby Names from Social Security Card Applications—National Data. Dataset US-GOV-SSA-338, General Services Administration, Washington, DC, USA.

World Bank Open Data. 2022. Population, total. Dataset SP.POP.TOTL, World Bank, Washington, DC, USA.

Yadollah Yaghoobzadeh and Hinrich Schütze. 2016. Intrinsic Subspace Evaluation of Word Embedding Representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers, pages 236–246, Berlin, Germany. Association for Computational Linguistics.

Zhao Yao, Jia Wu, Yanyan Zhang, and Zhenhong Wang. 2017. Norms of valence, arousal, concreteness, familiarity, imageability, and context availability for 1,100 Chinese words. *Behavior Research Methods*, 49(4):1374–1385.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning Gender-Neutral Word Embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.

George Kingsley Zipf. 1936. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. George Routledge & Sons, Ltd., London, United Kingdom.

A Derivation of MaxCov and MaxCorr

In this appendix we derive the algorithms for MaxCov and MaxCorr. We assume that we have access to a set of *reference words* $\mathbb{W}$, where each reference word $w \in \mathbb{W}$ is associated with a *score* $s(w) \in \mathbb{R}$.

A.1 Covariance Maximization

The formula for MaxCov is straightforwardly derived from the definition of covariance. We assume a uniform distribution on $\mathbb{W}$, so that expected values over $\mathbb{W}$ can be identified with means over $\mathbb{W}$.

Theorem 1. *Let $\mathbb{W}$ be a set of reference words, and let $s : \mathbb{W} \to \mathbb{R}$ be a scoring function. Let*

$$\boldsymbol{a} = \sum_{w \in \mathbb{W}} (s(w) - \overline{s})([\![w]\!] - \overline{\boldsymbol{w}}),$$

where

$$\overline{s} = \operatorname*{mean}_{w \in \mathbb{W}}(s(w)) \text{ and } \overline{\boldsymbol{w}} = \operatorname*{mean}_{w \in \mathbb{W}}([\![w]\!]).$$

Then,

$$\frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} = \operatorname*{argmax}_{\|\boldsymbol{v}\|=1} \operatorname*{cov}_{w \in \mathbb{W}}(\boldsymbol{v}^\top [\![w]\!], s(w)).$$

Proof. By definition,

$$\begin{aligned}
&\operatorname*{cov}_{w \in \mathbb{W}}(\boldsymbol{v}^\top [\![w]\!], s(w)) \\
&= \operatorname*{mean}_{w \in \mathbb{W}} \left((\boldsymbol{v}^\top [\![w]\!] - \boldsymbol{v}^\top \overline{\boldsymbol{w}})(s(w) - \overline{s}) \right) \\
&= \frac{1}{|\mathbb{W}|} \left(\sum_{w \in \mathbb{W}} (s(w) - \overline{s})([\![w]\!] - \overline{\boldsymbol{w}}) \right)^\top \boldsymbol{v}.
\end{aligned}$$

For any vector $\boldsymbol{u} \neq \boldsymbol{0}$, the unit vector $\boldsymbol{v}$ that maximizes $\boldsymbol{u}^\top \boldsymbol{v}$ is $\boldsymbol{v} = \boldsymbol{u}/\|\boldsymbol{u}\|$. Thus, writing

$$\boldsymbol{u} = \frac{1}{|\mathbb{W}|} \left(\sum_{w \in \mathbb{W}} (s(w) - \overline{s})([\![w]\!] - \overline{\boldsymbol{w}}) \right),$$

we have

$$\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|} = \operatorname*{argmax}_{\|\boldsymbol{v}\|=1} \operatorname*{cov}_{w \in \mathbb{W}}(\boldsymbol{v}^\top [\![w]\!], s(w)).$$

The theorem follows from the observation that $\boldsymbol{u} = \boldsymbol{a}/|\mathbb{W}|$, thus $\boldsymbol{u}/\|\boldsymbol{u}\| = \boldsymbol{a}/\|\boldsymbol{a}\|$. □

A.2 Correlation Maximization

In MaxCorr, we fit a linear model

$$s(w) = \boldsymbol{a}^\top [\![w]\!] + b,$$

and take our interpreted direction to be $\boldsymbol{g} = \boldsymbol{a}/\|\boldsymbol{a}\|$. Intuitively, $\operatorname{corr}_{w \in \mathbb{W}}(\boldsymbol{v}^\top [\![w]\!], s(w))$ measures the extent to which a linear relationship exists between $\boldsymbol{v}^\top [\![w]\!]$ and $s(w)$. When fitting the above linear regression, we are finding the vector $\boldsymbol{a}$ that maximizes the linear relationship between $\boldsymbol{a}^\top [\![w]\!]$ and $s(w)$. Thus, $\boldsymbol{a}$ gives us the direction of the subspace that maximizes correlation.

Theorem 2. *Let $\mathbb{W}$ be a set of reference words, and let $s : \mathbb{W} \to \mathbb{R}$ be a scoring function. Let*

$$\boldsymbol{a}^*, b^* = \operatorname*{argmin}_{\boldsymbol{a},b} \operatorname*{MSE}_{w\in\mathbb{W}}(\boldsymbol{a}^\top [\![w]\!] + b, s(w)),$$

where MSE *is the mean squared error, given by*

$$\operatorname*{MSE}_{x\in\mathbb{X}}(\phi(x), \psi(x)) = \frac{1}{|\mathbb{X}|}\sum_{x\in\mathbb{X}}(\psi(x) - \phi(x))^2$$

for a finite set $\mathbb{X}$ and functions $\psi, \phi : \mathbb{X} \to \mathbb{R}$. Then,

$$\frac{\boldsymbol{a}^*}{\|\boldsymbol{a}^*\|} = \operatorname*{argmax}_{\|\boldsymbol{v}\|=1} \operatorname*{corr}_{w\in\mathbb{W}}(\boldsymbol{v}^\top [\![w]\!], s(w)).$$

Proof. Fix a unit vector $\boldsymbol{v}$. Let $a_{\boldsymbol{v}}, b_{\boldsymbol{v}}$ be the result of fitting a linear regression model between $\boldsymbol{v}^\top [\![w]\!]$ and $s(w)$:

$$a_{\boldsymbol{v}}, b_{\boldsymbol{v}} = \operatorname*{argmin}_{a,b} \operatorname*{MSE}_{w\in\mathbb{W}}(a\boldsymbol{v}^\top [\![w]\!] + b, s(w)).$$

We use the fact that

$$\begin{aligned}
&\operatorname*{corr}_{w\in\mathbb{W}}(\boldsymbol{v}^\top [\![w]\!], s(w)) \\
&= \operatorname*{corr}_{w\in\mathbb{W}}(a_{\boldsymbol{v}}\boldsymbol{v}^\top [\![w]\!] + b_{\boldsymbol{v}}, s(w)) \\
&= \sqrt{1 - \frac{\operatorname{MSE}_{w\in\mathbb{W}}(a_{\boldsymbol{v}}\boldsymbol{v}^\top [\![w]\!] + b_{\boldsymbol{v}}, s(w))}{\operatorname{var}_{w\in\mathbb{W}}(s(w))}},
\end{aligned}$$

which follows from the invariance of correlation under linear mappings and the partition of sums of squares. Since

$$\begin{aligned}
&\operatorname*{MSE}_{w\in\mathbb{W}}(a_{\boldsymbol{v}}\boldsymbol{v}^\top [\![w]\!] + b_{\boldsymbol{v}}, s(w)) \\
&\geq \operatorname*{MSE}_{w\in\mathbb{W}}(\boldsymbol{a}^{*\top} [\![w]\!] + b^*, s(w))
\end{aligned}$$

by the definition of $\boldsymbol{a}^*$ and b^*, we have

$$\begin{aligned}
&\operatorname*{corr}_{w\in\mathbb{W}}(a_{\boldsymbol{v}}\boldsymbol{v}^\top [\![w]\!] + b_{\boldsymbol{v}}, s(w)) \\
&\leq \operatorname*{corr}_{w\in\mathbb{W}}(\boldsymbol{a}^{*\top} [\![w]\!] + b^*, s(w)) \\
&= \operatorname*{corr}_{w\in\mathbb{W}}(a_{\boldsymbol{v}^*}\boldsymbol{v}^{*\top} [\![w]\!] + b_{\boldsymbol{v}^*}, s(w)),
\end{aligned}$$

where $\boldsymbol{v}^* = \boldsymbol{a}^*/\|\boldsymbol{a}^*\|$. Thus,

$$\begin{aligned}
&\operatorname*{corr}_{w\in\mathbb{W}}(a_{\boldsymbol{v}^*}\boldsymbol{v}^{*\top} [\![w]\!] + b_{\boldsymbol{v}^*}, s(w)) \\
&= \max_{\|\boldsymbol{v}\|=1} \operatorname*{corr}_{w\in\mathbb{W}}(a_{\boldsymbol{v}}\boldsymbol{v}^\top [\![w]\!] + b_{\boldsymbol{v}}, s(w)) \\
&= \max_{\|\boldsymbol{v}\|=1} \operatorname*{corr}_{w\in\mathbb{W}}(\boldsymbol{v}^\top [\![w]\!], s(w)),
\end{aligned}$$

whence the theorem immediately follows. □

B Implementation of Interpretable Subspace Estimation Methods

This appendix provides implementational details for MaxCov, MaxCorr, the Tuples method, the SVM method, and DensRay, as they are used in the social bias mitigation experiment of Section 5. Across the three evaluation methods considered in the experiment, a total of five interpretable subspaces are estimated: 1-dimensional male–female gender subspaces in English, French, and Chinese; a 1-dimensional Black–White English race subspace used for WEAT 3–5 and the ECT test; and a 3-dimensional White–Asian–Hispanic English race subspace used in the occupation statistics analysis. For the purposes of this appendix, the terms "Black," "White," "Asian," and "Hispanic" are based on racial categories in the United States (African Americans, European Americans, Asian Americans, and Hispanic Americans, respectively). In all cases except when the Tuples Method is used, a basis for the 3-dimensional White–Asian–Hispanic subspace is estimated by orthonormalizing bases for three separate 1-dimensional subspaces.

B.1 MaxCov and MaxCorr

The four 1-dimensional subspaces are estimated using the following datasets.

- **English Gender:** SKB19-G
- **French Gender:** Ga08-G
- **Chinese Gender:** Ba21-G
- **Black–White Race:** SD18-Ra

The White–Asian–Hispanic race subspace is constructed by combining subspaces estimated from the "White," "Asian," and "Hispanic" columns of the Census-Ra dataset.

B.2 Tuples Method

Table 6 shows all word tuples used for the Tuples method. The English Gender word pairs are from Bolukbasi et al. (2016), while the first three columns of the Race words are from Manzini et al. (2019). The French Gender and Chinese Gender words were obtained by manually translating the English Gender words. The three gender subspaces are estimated using the pairs from the English Gender, French Gender, and Chinese Gender sections

English Gender		French Gender		Chinese Gender		Race			
she	he	elle	il	她	他	black	caucasian	asian	hispanic
her	his	sa	son	她们	他们	african	caucasian	asian	latino
woman	man	femme	homme	女人	男人	black	white	asian	hispanic
mary	john	Marie	Jean	玛丽	约翰	africa	america	asia	mexico
herself	himself	la	le	女儿	儿子	africa	america	china	mexico
daughter	son	fille	fils	母亲	父亲	africa	europe	asia	mexico
mother	father	mère	père	姐姐	哥哥				
gal	guy	meuf	mec	姊妹	兄弟				
girl	boy	fille	garçon	女孩	男孩				
female	male	féminine	masculin	女	男				

Table 6: Words used to estimate gender and race subspaces via the Tuples method. The English Gender words are from Bolukbasi et al. (2016); the first three columns of Race words are from Manzini et al. (2019).

of Table 6. The Black–White race subspace is estimated using the procedure of Manzini et al. (2019): the first three columns of the Race section are used, and the subspace is given by the first principal component of the difference vectors. The White–Asian–Hispanic subspace is estimated using the last three columns of the Race section and taking the first three principal components of the difference vectors.

B.3 SVM Method

As described in Subsection 5.2, for each 1-dimensional subspace we train an SVM using word embeddings with the 7,500 highest and lowest projections onto some subspace $\mathrm{span}(\llbracket w_1 \rrbracket - \llbracket w_2 \rrbracket)$, where (w_1, w_2) is a pair of reference words representing the target semantic concept. We use the following word pairs for the four 1-dimensional spaces.

- **English Gender:** *she–he*
- **French Gender:** *elle–il*
- **Chinese Gender:** 女– 男
- **Black–White Race:** *blacks–whites*

For the White–Asian–Hispanic subspace, we combine three 1-dimensional subspaces obtained using the pairs *whites–hispanics*, *whites–asians*, and *asians–hispanics*.

B.4 DensRay

Table 8 shows the word sets used for DensRay. Following Dufter and Schütze (2019), we use the kinship terms from Mikolov et al. (2013a) in order to estimate the English gender subspace. These words are shown in the English Male Kinship Terms and English Female Kinship Terms sections of Table 8. Likewise, the French and Chinese gender subspaces are also estimated using kinship terms. The French kinship terms were translated by the authors, while the Chinese kinship terms were obtained from the Wikipedia page *Chinese kinship*.[5] For Black–White race we use the given names from Kiritchenko and Mohammad (2018), shown in the African American Given Names and European American Given Names sections of Table 8. The remaining sections of the table consist of family names from Garg et al. (2018). Using these sets of names, we estimate the White–Asian–Hispanic subspace by combining a White–Asian subspace, an Asian–Hispanic subspace, and a Hispanic–White subspace.

C Dataset Permissions

Table 7 lists the permissions for the datasets used to estimate interpretable subspaces with MaxCov and MaxCorr. Table 1 contains a mixture of public-domain government publications, open access academic publications, and closed access publications. For each open access dataset, we provide a link to its associated license.

[5] https://en.wikipedia.org/wiki/Chinese_kinship

Data Source	Open Access?	License
Kennison and Trofe (2003)	No	N/A
Gabriel et al. (2008)	No	N/A
Gilet et al. (2012)	No	N/A
Pew Research Center (2012)	Yes	Pew Research Center Terms of Use
Yao et al. (2017)	No	N/A
Mohammad (2018)	Yes	CC BY 4.0
Stelter and Degner (2018)	Yes	CC BY 4.0
National Institute of Statistics and Economic Studies (2019)	Yes	License Ouverte/Open License 2.0
Scott et al. (2019)	Yes	CC BY 4.0
United States Social Security Administration (2019)	Yes	CC0 1.0
Bao (2021)	Yes	GPL-3
United States Census Bureau (2021)	Yes	Public Domain
World Bank Open Data (2022)	Yes	CC BY 4.0

Table 7: Permissions for the data sources listed in Table 1, including hyperlinks to licenses for open-access datasets.

English Male Kinship Terms (Mikolov et al., 2013a)	boy, brother, brothers, dad, father, grandfather, grandpa, grandson, groom, he, his, husband, king, man, nephew, policeman, prince, son, sons, stepbrother, stepfather, stepson, uncle
English Female Kinship Terms (Mikolov et al., 2013a)	aunt, bride, daughter, daughters, girl, granddaughter, grandma, grandmother, her, mom, mother, niece, policewoman, princess, queen, she, sister, sisters, stepdaughter, stepmother, stepsister, wife, woman
French Male Kinship Terms	beau-fils, beau-frère, beau-père, fils, fils, frère, frères, garçon, grand-père, homme, il, mari, marié, neveu, oncle, papa, papi, petit-fils, policier, prince, père, roi, son
French Female Kinship Terms	belle-fille, belle-mère, belle-sœur, copine, elle, femme, fille, filles, fillette, grand-mère, maman, mamie, mariée, mère, nièce, petite-fille, policière, princesse, reine, sa, sœur, sœurs, tante
Chinese Male Kinship Terms	丈夫, 亲王, 他, 他们, 伯伯, 伯父, 侄儿, 侄女婿, 侄子, 儿子, 兄, 兄弟, 公公, 内兄, 内弟, 叔叔, 叔父, 哥哥, 国王, 堂兄, 堂弟, 外孙儿, 外孙女婿, 外孙子, 外父, 外甥, 外祖父, 大伯, 大舅, 女婿, 妹夫, 姊夫, 姐夫, 姑夫, 姑父, 姥爷, 姨夫, 姨父, 姨甥, 孙儿, 孙女婿, 孙子, 家公, 小叔, 小舅, 岳丈, 岳父, 弟, 弟弟, 新郎, 父亲, 爷爷, 爸爸, 王子, 男人, 男孩, 祖父, 老公, 老爷, 舅父, 舅舅, 表兄, 表弟, 襟兄, 襟弟
Chinese Female Kinship Terms	丈母, 伯娘, 伯母, 侄女, 侄媳妇, 儿媳, 公主, 堂妹, 堂姊, 堂姐, 堂嫂, 外孙女, 外孙媳妇, 外母, 外甥女, 外祖母, 大姑, 大姨, 大嫂, 太太, 女亲王, 女人, 女儿, 女孩, 女王, 奶奶, 她, 她们, 妈妈, 妯娌, 妹, 妹妹, 妻子, 姊, 姊姊, 姐姐, 姑妈, 姑姐, 姑姑, 姑母, 姑表姊, 姑表嫂, 姥姥, 姨妈, 姨妹, 姨姐, 姨母, 姨甥女, 婆婆, 婶婶, 婶母, 媳妇, 嫂, 嫂子, 孙女, 孙媳妇, 家姑, 家婆, 小姑, 小姨, 小婶, 岳母, 弟妇, 新郎, 母亲, 皇后, 祖母, 老婆, 舅妈, 舅母, 表妹, 表姊, 表姐, 表嫂, 闺女, 阿姨
African American Given Names (Kiritchenko and Mohammad, 2018)	alonzo, alphonse, darnell, ebony, jamel, jasmine, jerome, lakisha, lamar, latisha, latoya, leroy, malik, nichelle, shaniqua, shereen, tanisha, terrence, tia, torrance
European American Given Names (Kiritchenko and Mohammad, 2018)	adam, alan, amanda, andrew, betsy, courtney, ellen, frank, harry, heather, jack, josh, justin, katie, kristin, melanie, nancy, roger, ryan, stephanie
European American Family Names (Garg et al., 2018)	adams, allen, anderson, clark, davis, harris, jackson, johnson, jones, lewis, martin, moore, nelson, robinson, scott, taylor, thompson, williams, wilson, wright
Asian American Family Names (Garg et al., 2018)	chang, chen, cho, chu, chung, hong, huang, khan, kim, li, lin, liu, ng, shah, singh, tang, wang, wong, wu, yang
Hispanic American Family Names (Garg et al., 2018)	alvarez, castillo, castro, cruz, diaz, garcia, gomez, gonzalez, lopez, martinez, medina, mendoza, perez, rivera, rodriguez, ruiz, sanchez, soto, torres, vargas

Table 8: Words used to estimate gender and race subspaces via DensRay.