Discovering Sociolinguistic Associations with Structured Sparsity

Jacob Eisenstein Noah A. Smith Eric P. Xing
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{jacobeis,nasmith,epxing}@cs.cmu.edu

Abstract

We present a method to discover robust and interpretable sociolinguistic associations from raw geotagged text data. Using aggregate demographic statistics about the authors' geographic communities, we solve a multi-output regression problem between demographics and lexical frequencies. By imposing a composite $\ell_{1,\infty}$ regularizer, we obtain structured sparsity, driving entire rows of coefficients to zero. We perform two regression studies. First, we use term frequencies to predict demographic attributes; our method identifies a compact set of words that are strongly associated with author demographics. Next, we conjoin demographic attributes into *features*, which we use to predict term frequencies. The composite regularizer identifies a small number of features, which correspond to communities of authors united by shared demographic and linguistic properties.

1 Introduction

How is language influenced by the speaker's sociocultural identity? Quantitative sociolinguistics usually addresses this question through carefully crafted studies that correlate individual demographic attributes and linguistic variables—for example, the interaction between income and the "dropped r" feature of the New York accent (Labov, 1966). But such studies require the knowledge to select the "dropped r" and the speaker's income, from thousands of other possibilities. In this paper, we present a method to acquire such patterns from raw data. Using multi-output regression with structured sparsity, our method identifies a small subset of lexical items that are most influenced by demographics, and discovers conjunctions of demographic attributes that are especially salient for lexical variation.

Sociolinguistic associations are difficult to model, because the space of potentially relevant interactions is large and complex. On the linguistic side there are thousands of possible variables, even if we limit ourselves to unigram lexical features. On the demographic side, the interaction between demographic attributes is often non-linear: for example, gender may negate or amplify class-based language differences (Zhang, 2005). Thus, additive models which assume that each demographic attribute makes a linear contribution are inadequate.

In this paper, we explore the large space of potential sociolinguistic associations using structured sparsity. We treat the relationship between language and demographics as a set of multi-input, multi-output regression problems. The regression coefficients are arranged in a matrix, with rows indicating predictors and columns indicating outputs. We apply a composite regularizer that drives entire rows of the coefficient matrix to zero, yielding compact, interpretable models that reuse features across different outputs. If we treat the lexical frequencies as inputs and the author's demographics as outputs, the induced sparsity pattern reveals the set of lexical items that is most closely tied to demographics. If we treat the demographic attributes as inputs and build a model to predict the text, we can incrementally construct a conjunctive feature space of demographic attributes, capturing key non-linear interactions.

The primary purpose of this research is exploratory data analysis to identify both the most linguistic-salient demographic features, and the most demographically-salient words. However, this model also enables predictions about demographic features by analyzing raw text, potentially supporting applications in targeted information extraction or advertising. On the task of predicting demographics from text, we find that our sparse model yields performance that is statistically indistinguishable from the full vocabulary, even with a reduction in the model complexity an order of magnitude. On the task of predicting text from author demographics, we find that our incrementally constructed feature set obtains significantly better perplexity than a linear model of demographic attributes.

	mean	std. dev.
race & ethnicity		
% white	52.1	29.0
% African American	32.2	29.1
% Hispanic	15.7	18.3
language		
% English speakers	73.7	18.4
% Spanish speakers	14.6	15.6
% other language speakers	11.7	9.2
socioeconomic		
% urban	95.1	14.3
% with family	64.1	14.4
% renters	48.9	23.4
median income ($)	42,500	18,100

Table 1: The demographic attributes used in this research.

2 Data

Our dataset is derived from prior work in which we gathered the text and geographical locations of 9,250 microbloggers on the website `twitter.com` (Eisenstein et al., 2010). Bloggers were selected from a pool of frequent posters whose messages include metadata indicating a geographical location within a bounding box around the continental United States. We limit the vocabulary to the 5,418 terms which are used by at least 40 authors; no stoplists are applied, as the use of standard or non-standard orthography for stopwords (e.g., *to* vs. *2*) may convey important information about the author. The dataset includes messages during the first week of March 2010.

O'Connor et al. (2010) obtained aggregate demographic statistics for these data by mapping geolocations to publicly-available data from the U. S. Census ZIP Code Tabulation Areas (ZCTA).[1] There are 33,178 such areas in the USA (the 9,250 microbloggers in our dataset occupy 3,458 unique ZCTAs), and they are designed to contain roughly equal numbers of inhabitants and demographically-homogeneous populations. The demographic attributes that we consider in this paper are shown in Table 1. All attributes are based on self-reports. The race and ethnicity attributes are not mutually exclusive—individuals can indicate any number of races or ethnicities. The "other language" attribute aggregates all languages besides English and Spanish. "Urban areas" refer to sets of census tracts or census blocks which contain at least 2,500 residents; our "% urban" attribute is the percentage of individuals in each ZCTA who are listed as living in an urban area. We also consider the percentage of individuals who live with their families, the percentage who live in rented housing, and the median reported income in each ZCTA.

While geographical aggregate statistics are frequently used to proxy for individual socioeconomic status in research areas such as public health (e.g., Rushton, 2008), it is clear that interpretation must proceed with caution. Consider an author from a ZIP code in which 60% of the residents are Hispanic:[2] we do not know the likelihood that the author is Hispanic, because the set of Twitter users is not a representative sample of the overall population. Polling research suggests that users of both Twitter (Smith and Rainie, 2010) and geolocation services (Zickuhr and Smith, 2010) are much more diverse with respect to age, gender, race and ethnicity than the general population of Internet users. Nonetheless, at present we can only use aggregate statistics to make inferences about the geographic communities in which our authors live, and not the authors themselves.

[1] `http://www.census.gov/support/cen2000.html`

[2] In the U.S. Census, the official ethnonym is *Hispanic or Latino*; for brevity we will use *Hispanic* in the rest of this paper.

3 Models

The selection of both words and demographic features can be framed in terms of multi-output regression with structured sparsity. To select the lexical indicators that best predict demographics, we construct a regression problem in which term frequencies are the predictors and demographic attributes are the outputs; to select the demographic features that predict word use, this arrangement is reversed. Through structured sparsity, we learn models in which entire sets of coefficients are driven to zero; this tells us which words and demographic features can safely be ignored.

This section describes the model and implementation for output-regression with structured sparsity; in Section 4 and 5 we give the details of its application to select terms and demographic features. Formally, we consider the linear equation $\mathbf{Y} = \mathbf{XB} + \epsilon$, where,

- $\mathbf{Y}$ is the dependent variable matrix, with dimensions $N \times T$, where N is the number of samples and T is the number of output dimensions (or *tasks*);
- $\mathbf{X}$ is the independent variable matrix, with dimensions $N \times P$, where P is the number of input dimensions (or *predictors*);
- $\mathbf{B}$ is the matrix of regression coefficients, with dimensions $P \times T$;
- ϵ is a $N \times T$ matrix in which each element is noise from a zero-mean Gaussian distribution.

We would like to solve the unconstrained optimization problem,

$$\text{minimize}_{\mathbf{B}} \qquad ||\mathbf{Y} - \mathbf{XB}||_F^2 + \lambda R(\mathbf{B}), \qquad (1)$$

where $||\mathbf{A}||_F^2$ indicates the squared Frobenius norm $\sum_i \sum_j a_{ij}^2$, and the function $R(\mathbf{B})$ defines a norm on the regression coefficients $\mathbf{B}$. Ridge regression applies the ℓ_2 norm $R(\mathbf{B}) = \sum_{t=1}^{T} \sum_{p}^{P} b_{pt}^2$, and lasso regression applies the ℓ_1 norm $R(\mathbf{B}) = \sum_{t=1}^{T} \sum_{p}^{P} |b_{pt}|$; in both cases, it is possible to decompose the multi-output regression problem, treating each output dimension separately. However, our working hypothesis is that there will be substantial correlations across both the vocabulary and the demographic features—for example, a demographic feature such as the percentage of Spanish speakers will predict a large set of words. Our goal is to select a small set of predictors yielding good performance across all output dimensions. Thus, we desire *structured* sparsity, in which entire rows of the coefficient matrix $\mathbf{B}$ are driven to zero.

Structured sparsity is not achieved by the lasso's ℓ_1 norm. The lasso gives element-wise sparsity, in which many entries of $\mathbf{B}$ are driven to zero, but each predictor may have a non-zero value for some output dimension. To drive entire rows of $\mathbf{B}$ to zero, we require a composite regularizer. We consider the $\ell_{1,\infty}$ norm, which is the sum of ℓ_∞ norms across output dimensions: $R(\mathbf{B}) = \sum_t^T \max_p b_{pt}$ (Turlach et al., 2005). This norm, which corresponds to a *multi-output lasso* regression, has the desired property of driving entire rows of $\mathbf{B}$ to zero.

3.1 Optimization

There are several techniques for solving the $\ell_{1,\infty}$ normalized regression, including interior point methods (Turlach et al., 2005) and projected gradient (Duchi et al., 2008; Quattoni et al., 2009). We choose the blockwise coordinate descent approach of Liu et al. (2009) because it is easy to implement and efficient: the time complexity of each iteration is independent of the number of samples.[3]

Due to space limitations, we defer to Liu et al. (2009) for a complete description of the algorithm. However, we note two aspects of our implementation which are important for natural language processing applications. The algorithm's efficiency is accomplished by precomputing the matrices $\mathbf{C} = \tilde{\mathbf{X}}^\mathsf{T}\tilde{\mathbf{Y}}$ and $\mathbf{D} = \tilde{\mathbf{X}}^\mathsf{T}\tilde{\mathbf{X}}$, where $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ are the standardized versions of $\mathbf{X}$ and $\mathbf{Y}$, obtained by subtracting the mean and scaling by the variance. Explicit mean correction would destroy the sparse term frequency data representation and render us unable to store the data in memory; however, we can achieve the same effect by computing $\mathbf{C} = \mathbf{X}^\mathsf{T}\mathbf{Y} - N\bar{\boldsymbol{x}}^\mathsf{T}\bar{\boldsymbol{y}}$, where $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{y}}$ are row vectors indicating the means

[3] Our implementation is available at `http://sailing.cs.cmu.edu/sociolinguistic.html`.

of $\mathbf{X}$ and $\mathbf{Y}$ respectively.[4] We can similarly compute $\mathbf{D} = \mathbf{X}^\mathsf{T}\mathbf{X} - N\bar{\boldsymbol{x}}^\mathsf{T}\bar{\boldsymbol{x}}$.

If the number of predictors is too large, it may not be possible to store the dense matrix $\mathbf{D}$ in memory. We have found that approximation based on the truncated singular value decomposition provides an effective trade-off of time for space. Specifically, we compute $\mathbf{X}^\mathsf{T}\mathbf{X} \approx$

$$\mathbf{U}\mathbf{S}\mathbf{V}^\mathsf{T}\left(\mathbf{U}\mathbf{S}\mathbf{V}^\mathsf{T}\right)^\mathsf{T} = \mathbf{U}\left(\mathbf{S}\mathbf{V}^\mathsf{T}\mathbf{V}\mathbf{S}^\mathsf{T}\mathbf{U}^\mathsf{T}\right) = \mathbf{U}\mathbf{M}.$$

Lower truncation levels are less accurate, but are faster and require less space: for K singular values, the storage cost is $\mathcal{O}(KP)$, instead of $\mathcal{O}(P^2)$; the time cost increases by a factor of K. This approximation was not necessary in the experiments presented here, although we have found that it performs well as long as the regularizer is not too close to zero.

3.2 Regularization

The regularization constant λ can be computed using cross-validation. As λ increases, we reuse the previous solution of $\mathbf{B}$ for initialization; this "warm start" trick can greatly accelerate the computation of the overall regularization path (Friedman et al., 2010). At each λ_i, we solve the sparse multi-output regression; the solution $\mathbf{B}_i$ defines a sparse set of predictors for all tasks.

We then use this limited set of predictors to construct a new input matrix $\hat{\mathbf{X}}_i$, which serves as the input in a standard ridge regression, thus refitting the model. The tuning set performance of this regression is the score for λ_i. Such post hoc refitting is often used in tandem with the lasso and related sparse methods; the effectiveness of this procedure has been demonstrated in both theory (Wasserman and Roeder, 2009) and practice (Wu et al., 2010). The regularization parameter of the ridge regression is determined by internal cross-validation.

4 Predicting Demographics from Text

Sparse multi-output regression can be used to select a subset of vocabulary items that are especially indicative of demographic and geographic differences. Starting from the regression problem (1), the predictors $\mathbf{X}$ are set to the term frequencies, with one column for each word type and one row for each author in the dataset. The outputs $\mathbf{Y}$ are set to the ten demographic attributes described in Table 1 (we consider much larger demographic feature spaces in the next section) The $\ell_{1,\infty}$ regularizer will drive entire rows of the coefficient matrix $\mathbf{B}$ to zero, eliminating all demographic effects for many words.

4.1 Quantitative Evaluation

We evaluate the ability of lexical features to predict the demographic attributes of their authors (as proxied by the census data from the author's geographical area). The purpose of this evaluation is to assess the predictive ability of the compact subset of lexical items identified by the multi-output lasso, as compared with the full vocabulary. In addition, this evaluation establishes a baseline for performance on the demographic prediction task.

We perform five-fold cross-validation, using the multi-output lasso to identify a sparse feature set in the training data. We compare against several other dimensionality reduction techniques, matching the number of features obtained by the multi-output lasso at each fold. First, we compare against a truncated singular value decomposition, with the truncation level set to the number of terms selected by the multi-output lasso; this is similar in spirit to vector-based lexical semantic techniques (Schütze and Pedersen, 1993). We also compare against simply selecting the N most frequent terms, and the N terms with the greatest variance in frequency across authors. Finally, we compare against the complete set of all 5,418 terms. As before, we perform post hoc refitting on the training data using a standard ridge regression. The regularization constant for the ridge regression is identified using nested five-fold cross validation within the training set.

We evaluate on the refit models on the heldout test folds. The scoring metric is Pearson's correlation coefficient between the predicted and true demographics: $\rho(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{\text{cov}(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\sigma_{\boldsymbol{y}}\sigma_{\hat{\boldsymbol{y}}}}$, with $\text{cov}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ indicating the covariance and $\sigma_{\boldsymbol{y}}$ indicating the standard deviation. On this metric, a perfect predictor will score 1 and a random predictor will score 0. We report the average correlation across all ten demo-

[4] Assume without loss of generality that $\mathbf{X}$ and $\mathbf{Y}$ are scaled to have variance 1, because this scaling does not affect the sparsity pattern.

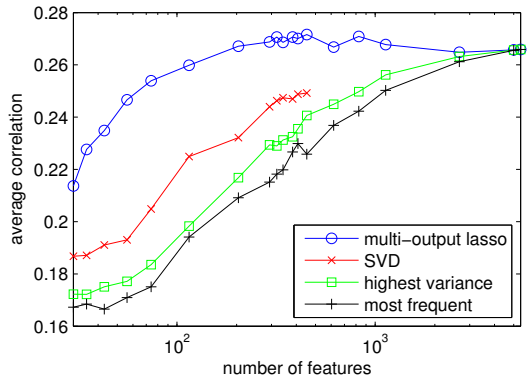

Figure 1: Average correlation plotted against the number of active features (on a logarithmic scale).

graphic attributes, as well as the individual correlations.

Results Table 2 shows the correlations obtained by regressions performed on a range of different vocabularies, averaged across all five folds. Linguistic features are best at predicting race, ethnicity, language, and the proportion of renters; the other demographic attributes are more difficult to predict. Among feature sets, the highest average correlation is obtained by the full vocabulary, but the multi-output lasso obtains nearly identical performance using a feature set that is an order of magnitude smaller. Applying the Fischer transformation, we find that all correlations are statistically significant at $p < .001$.

The Fischer transformation can also be used to estimate 95% confidence intervals around the correlations. The extent of the confidence intervals varies slightly across attributes, but all are tighter than ± 0.02. We find that the multi-output lasso and the full vocabulary regression are not significantly different on any of the attributes. Thus, the multi-output lasso achieves a 93% compression of the feature set without a significant decrease in predictive performance. The multi-output lasso yields higher correlations than the other dimensionality reduction techniques on all of the attributes; these differences are statistically significant in many—but not all—cases. The correlations for each attribute are clearly not independent, so we do not compare the average across attributes.

Recall that the regularization coefficient was chosen by nested cross-validation within the training set; the average number of features selected is 394.6. Figure 1 shows the performance of each dimensionality-reduction technique across the regularization path for the first of five cross-validation folds. Computing the truncated SVD of a sparse matrix at very large truncation levels is computationally expensive, so we cannot draw the complete performance curve for this method. The multi-output lasso dominates the alternatives, obtaining a particularly strong advantage with very small feature sets. This demonstrates its utility for identifying interpretable models which permit qualitative analysis.

4.2 Qualitative Analysis

For a qualitative analysis, we retrain the model on the full dataset, and tune the regularization to identify a compact set of 69 features. For each identified term, we apply a significance test on the relationship between the presence of each term and the demographic indicators shown in the columns of the table. Specifically, we apply the Wald test for comparing the means of independent samples, while making the Bonferroni correction for multiple comparisons (Wasserman, 2003). The use of sparse multi-output regression for variable selection increases the power of post hoc significance testing, because the Bonferroni correction bases the threshold for statistical significance on the total number of comparisons. We find 275 associations at the $p < .05$ level; at the higher threshold required by a Bonferroni correction for comparisons among all terms in the vocabulary, 69 of these associations would have been missed.

Table 3 shows the terms identified by our model which have a significant correlation with at least one of the demographic indicators. We divide words in the list into categories, which order alphabetically by the first word in each category: emoticons; standard English, defined as words with Wordnet entries; proper names; abbreviations; non-English words; non-standard words used with English. The categorization was based on the most frequent sense in an informal analysis of our data. A glossary of non-standard terms is given in Table 4.

Some patterns emerge from Table 3. Standard English words tend to appear in areas with more

vocabulary	# features	average	white	Afr. Am.	Hisp.	Eng. lang.	Span. lang.	other lang.	urban	family	renter	med. inc.
full	5418	0.260	0.337	0.318	0.296	0.384	0.296	0.256	0.155	0.113	0.295	0.152
multi-output lasso	394.6	0.260	0.326	0.308	0.304	0.383	0.303	0.249	0.153	0.113	0.302	0.156
SVD		0.237	0.321	0.299	0.269	0.352	0.272	0.226	0.138	0.081	0.278	0.136
highest variance		0.220	0.309	0.287	0.245	0.315	0.248	0.199	0.132	0.085	0.250	0.135
most frequent		0.204	0.294	0.264	0.222	0.293	0.229	0.178	0.129	0.073	0.228	0.126

Table 2: Correlations between predicted and observed demographic attributes, averaged across cross validation folds.

English speakers; predictably, Spanish words tend to appear in areas with Spanish speakers and Hispanics. Emoticons tend to be used in areas with many Hispanics and few African Americans. Abbreviations (e.g., *lmaoo*) have a nearly uniform demographic profile, displaying negative correlations with whites and English speakers, and positive correlations with African Americans, Hispanics, renters, Spanish speakers, and areas classified as urban.

Many non-standard English words (e.g., *dats*) appear in areas with high proportions of renters, African Americans, and non-English speakers, though a subset (*haha*, *hahaha*, and *yep*) display the opposite demographic pattern. Many of these non-standard words are phonetic transcriptions of standard words or phrases: *that's→dats*, *what's up→wassup*, *I'm going to→ima*. The relationship between these transcriptions and the phonological characteristics of dialects such as African-American Vernacular English is a topic for future work.

5 Conjunctive Demographic Features

Next, we demonstrate how to select conjunctions of demographic features that predict text. Again, we apply multi-output regression, but now we reverse the direction of inference: the predictors are demographic features, and the outputs are term frequencies. The sparsity-inducing $\ell_{1,\infty}$ norm will select a subset of demographic features that explain the term frequencies.

We create an initial feature set $f^{(0)}(\mathbf{X})$ by binning each demographic attribute, using five equal-frequency bins. We then constructive conjunctive features by applying a procedure inspired by related work in computational biology, called "Screen and Clean" (Wu et al., 2010). On iteration i:

- Solve the sparse multi-output regression problem $\mathbf{Y} = f^{(i)}(\mathbf{X})\mathbf{B}^{(i)} + \epsilon$.
- Select a subset of features $S^{(i)}$ such that $m \in S^{(i)}$ iff $\max_j |\boldsymbol{b}^{(i)}_{m,j}| > 0$. These are the row indices of the predictors with non-zero coefficients.
- Create a new feature set $f^{(i+1)}(\mathbf{X})$, including the conjunction of each feature (and its negation) in $S^{(i)}$ with each feature in the initial set $f^{(0)}(\mathbf{X})$.

We iterate this process to create features that conjoin as many as three attributes. In addition to the binned versions of the demographic attributes described in Table 1, we include geographical information. We built Gaussian mixture models over the locations, with 3, 5, 8, 12, 17, and 23 components. For each author we include the most likely cluster assignment in each of the six mixture models. For efficiency, the outputs $\mathbf{Y}$ are not set to the raw term frequencies; instead we compute a truncated singular value decomposition of the term frequencies $\mathbf{W} \approx \mathbf{U}\mathbf{V}\mathbf{D}^\mathsf{T}$, and use the basis $\mathbf{U}$. We set the truncation level to 100.

5.1 Quantitative Evaluation

The ability of the induced demographic features to predict text is evaluated using a traditional perplexity metric. The same test and training split is used from the vocabulary experiments. We construct a language model from the induced demographic features by training a multi-output ridge regression, which gives a matrix $\hat{\mathbf{B}}$ that maps from demographic features to term frequencies across the entire vocabulary. For each document in the test set, the "raw" predicted language model is $\hat{\boldsymbol{y}}_d = f(\boldsymbol{x}_d)\mathbf{B}$, which is then normalized. The probability mass assigned

	white	Afr. Am.	Hisp.	Eng. lang.	Span. lang.	other lang.	urban	family	renter	med. inc.
-_-	-		+	-	+	+	+			
;)		-	+	-	+					
:(		-								
:)		-								
:d	+	-	+	-	+					
as			-	+	-					
awesome	+	-					-		-	+
break			-	+	-	-				
campus			-	+	-	-				
dead	-	+		-	+		+		+	
hell			-	+	-	-				
shit	-								+	
train				-	+				+	
will			-	+	-					
would				+					-	
atlanta			-	+	-	-				
famu		+	-	+	-	-				-
harlem				-					+	
bbm	-	+		-		+	+		+	
lls		+	-	+	-	-				
lmaoo	-	+	+	-	+	+	+		+	
lmaooo	-	+	+	-	+	+	+		+	
lmaoooo	-	+	+	-	+	+			+	
lmfaoo	-		+	-	+	+			+	
lmfaooo	-		+	-	+	+			+	
lml	-	+	+	-	+	+	+		+	-
odee	-		+	-	+		+		+	
omw	-	+	+	-	+	+	+		+	
smfh	-	+	+	-	+	+	+		+	
smh	-	+					+		+	
w|	-		+	-	+	+	+		+	
con			+	-	+				+	
la		-	+	-	+					
si		-	+	-	+					
dats	-	+		-					+	-
deadass	-	+	+	-	+	+	+		+	
haha	+	-							-	
hahah	+	-								
hahaha	+	-							-	+
ima	-		+	-	+				+	
madd	-			-		+		+		
nah	-		+	-	+	+			+	
ova	-	+		-					+	
sis	-	+							+	
skool	-	+		-		+	+		+	-
wassup	-	+	+	-	+	+	+		+	-
wat	-	+	+	-	+	+	+		+	-
ya	-	+							+	
yall	-	+								
yep			-	+	-	-	-		-	
yoo	-	+	+	-	+	+	+		+	
yooo	-	+		-	+				+	

Table 3: Demographically-indicative terms discovered by multi-output sparse regression. Statistically significant ($p < .05$) associations are marked with a + or -.

term	definition	term	definition
bbm	Blackberry Messenger	omw	on my way
dats	that's	ova	over
dead(ass)	very	sis	sister
famu	Florida Agricultural and Mechanical Univ.	skool	school
		sm(f)h	shake my (fucking) head
ima	I'm going to		
lls	laughing like shit	w|	with
lm(f)ao+	laughing my (fucking) ass off	wassup	what's up
		wat	what
lml	love my life	ya	your, you
madd	very, lots	yall	you plural
nah	no	yep	yes
odee	very	yoo+	you

Table 4: A glossary of non-standard terms from Table 3. Definitions are obtained by manually inspecting the context in which the terms appear, and by consulting `www.urbandictionary.com`.

model	perplexity
induced demographic features	333.9
raw demographic attributes	335.4
baseline (no demographics)	337.1

Table 5: Word perplexity on test documents, using language models estimated from induced demographic features, raw demographic attributes, and a relative-frequency baseline. Lower scores are better.

to unseen words is determined through nested cross-validation. We compare against a baseline language model obtained from the training set, again using nested cross-validation to set the probability of unseen terms.

Results are shown in Table 5. The language models induced from demographic data yield small but statistically significant improvements over the baseline (Wilcoxon signed-rank test, $p < .001$). Moreover, the model based on conjunctive features significantly outperforms the model constructed from raw attributes ($p < .001$).

5.2 Features Discovered

Our approach discovers 37 conjunctive features, yielding the results shown in Table 5. We sort all features by frequency, and manually select a subset to display in Table 6. Alongside each feature, we show the words with the highest and lowest log-odds ratios with respect to the feature. Many of these terms are non-standard; while space does not permit a complete glossary, some are defined in Table 4 or in our earlier work (Eisenstein et al., 2010).

	feature			positive terms	negative terms
1	geo: Northeast			m2 brib mangoville soho odeee	fasho #ilovefamu foo coo fina
2	geo: NYC			mangoville lolss m2 brib wordd	bahaha fasho goofy #ilovefamu tacos
4	geo: South+Midwest	renter $\leq$ 0.615	white $\leq$ 0.823	hme muthafucka bae charlotte tx	odeee m2 lolss diner mangoville
7	Afr. Am. $>$ 0.101	renter $>$ 0.615	Span. lang. $>$ 0.063	dhat brib odeee lolss wassupp	bahaha charlotte california ikr enter
8	Afr. Am. $\leq$ 0.207	Hispanic $>$ 0.119	Span. lang. $>$ 0.063	les ahah para san donde	bmore ohio #lowkey #twitterjail nahhh
9	geo: NYC	Span. lang. $\leq$ 0.213		mangoville thatt odeee lolss buzzin	landed rodney jawn wiz golf
12	Afr. Am. $>$ 0.442	geo: South+Midwest	white $\leq$ 0.823	#ilovefamu panama midterms willies #lowkey	knoe esta pero odeee hii
15	geo: West Coast	other lang. $>$ 0.110		ahah fasho san koo diego	granted pride adore phat pressure
17	Afr. Am. $>$ 0.442	geo: NYC	other lang. $\leq$ 0.110	lolss iim buzzin qonna qood	foo tender celebs pages pandora
20	Afr. Am. $\leq$ 0.207	Span. lang. $>$ 0.063	white $>$ 0.823	del bby cuando estoy muscle	knicks becoming uncomfortable large granted
23	Afr. Am. $\leq$ 0.050	geo: West	Span. lang. $\leq$ 0.106	leno it'd 15th hacked government	knicks liquor uu hunn homee
33	Afr. Am. $>$ 0.101	geo: SF Bay	Span. lang. $>$ 0.063	hella aha california bay o.o	aj everywhere phones shift regardless
36	Afr. Am. $\leq$ 0.050	geo: DC/Philadelphia	Span. lang. $\leq$ 0.106	deh opens stuffed yaa bmore	hmmmmm dyin tea cousin hella

Table 6: Conjunctive features discovered by our method with a strong sparsity-inducing prior, ordered by frequency. We also show the words with high log-odds for each feature (postive terms) and its negation (negative terms).

In general, geography was a strong predictor, appearing in 25 of the 37 conjunctions. Features 1 and 2 (F1 and F2) are purely geographical, capturing the northeastern United States and the New York City area. The geographical area of F2 is completely contained by F1; the associated terms are thus very similar, but by having both features, the model can distinguish terms which are used in northeastern areas outside New York City, as well as terms which are especially likely in New York.[5]

Several features conjoin geography with demographic attributes. For example, F9 further refines the New York City area by focusing on communities that have relatively low numbers of Spanish speakers; F17 emphasizes New York neighborhoods that have very high numbers of African Americans and few speakers of languages other than English and Spanish. The regression model can use these features in combination to make fine-grained distinctions about the differences between such neighborhoods. Outside New York, we see that F4 combines a broad geographic area with attributes that select at least moderate levels of minorities and fewer renters (a proxy for areas that are less urban), while F15 identifies West Coast communities with large numbers of speakers of languages other than English and Spanish.

Race and ethnicity appear in 28 of the 37 conjunctions. The attribute indicating the proportion of African Americans appeared in 22 of these features, strongly suggesting that African American Vernacular English (Rickford, 1999) plays an important role in social media text. Many of these features conjoined the proportion of African Americans with geographical features, identifying local linguistic styles used predominantly in either African American or white communities. Among features which focus on minority communities, F17 emphasizes the New York area, F33 focuses on the San Francisco Bay area, and F12 selects a broad area in the Midwest and South. Conversely, F23 selects areas with very few African Americans and Spanish-speakers in the western part of the United States, and F36 selects for similar demographics in the area of Washington and Philadelphia.

Other features conjoined the proportion of African Americans with the proportion of Hispanics and/or Spanish speakers. In some cases, features selected for high proportions of both African Americans and Hispanics; for example, F7 seems to identify a general "urban minority" group, emphasizing renters, African Americans, and Spanish speakers. Other features differentiate between African Ameri-

[5] *Mangoville* and *M2* are clubs in New York; *fasho* and *coo* were previously found to be strongly associated with the West Coast (Eisenstein et al., 2010).

cans and Hispanics: F8 identifies regions with many Spanish speakers and Hispanics, but few African Americans; F20 identifies regions with both Spanish speakers and whites, but few African Americans. F8 and F20 tend to emphasize more Spanish words than features which select for both African Americans and Hispanics.

While race, geography, and language predominate, the socioeconomic attributes appear in far fewer features. The most prevalent attribute is the proportion of renters, which appears in F4 and F7, and in three other features not shown here. This attribute may be a better indicator of the urban/rural divide than the "% urban" attribute, which has a very low threshold for what counts as urban (see Table 1). It may also be a better proxy for wealth than median income, which appears in only one of the thirty-seven selected features. Overall, the selected features tend to include attributes that are easy to predict from text (compare with Table 2).

6 Related Work

Sociolinguistics has a long tradition of quantitative and computational research. Logistic regression has been used to identify relationships between demographic features and linguistic variables since the 1970s (Cedergren and Sankoff, 1974). More recent developments include the use of mixed factor models to account for idiosyncrasies of individual speakers (Johnson, 2009), as well as clustering and multidimensional scaling (Nerbonne, 2009) to enable aggregate inference across multiple linguistic variables. However, all of these approaches assume that both the linguistic indicators and demographic attributes have already been identified by the researcher. In contrast, our approach focuses on identifying these indicators automatically from data. We view our approach as an exploratory complement to more traditional analysis.

There is relatively little computational work on identifying speaker demographics. Chang et al. (2010) use U.S. Census statistics about the ethnic distribution of last names as an anchor in a latent-variable model that infers the ethnicity of Facebook users; however, their paper analyzes social behavior rather than language use. In unpublished work, David Bamman uses geotagged Twitter text and U.S. Census statistics to estimate the age, gender, and racial distributions of various lexical items.[6] Eisenstein et al. (2010) infer geographic clusters that are coherent with respect to both location and lexical distributions; follow-up work by O'Connor et al. (2010) applies a similar generative model to demographic data. The model presented here differs in two key ways: first, we use sparsity-inducing regularization to perform variable selection; second, we eschew high-dimensional mixture models in favor of a bottom-up approach of building conjunctions of demographic and geographic attributes. In a mixture model, each component must define a distribution over all demographic variables, which may be difficult to estimate in a high-dimensional setting.

Early examples of the use of sparsity in natural language processing include maximum entropy classification (Kazama and Tsujii, 2003), language modeling (Goodman, 2004), and incremental parsing (Riezler and Vasserman, 2004). These papers all apply the standard lasso, obtaining sparsity for a single output dimension. Structured sparsity has rarely been applied to language tasks, but Duh et al. (2010) reformulated the problem of reranking N-best lists as multi-task learning with structured sparsity.

7 Conclusion

This paper demonstrates how regression with structured sparsity can be applied to select words and conjunctive demographic features that reveal sociolinguistic associations. The resulting models are compact and interpretable, with little cost in accuracy. In the future we hope to consider richer linguistic models capable of identifying multi-word expressions and syntactic variation.

Acknowledgments We received helpful feedback from Moira Burke, Scott Kiesling, Seyoung Kim, André Martins, Kriti Puniyani, and the anonymous reviewers. Brendan O'Connor provided the data for this research, and Seunghak Lee shared a Matlab implementation of the multi-output lasso, which was the basis for our C implementation. This research was enabled by AFOSR FA9550010247, ONR N0001140910758, NSF CAREER DBI-0546594, NSF CAREER IIS-1054319, NSF IIS-0713379, an Alfred P. Sloan Fellowship, and Google's support of the Worldly Knowledge project at CMU.

[6] `http://www.lexicalist.com`

References

Henrietta J. Cedergren and David Sankoff. 1974. Variable rules: Performance as a statistical reflection of competence. *Language*, 50(2):333–355.

Jonathan Chang, Itamar Rosenn, Lars Backstrom, and Cameron Marlow. 2010. ePluribus: Ethnicity on social networks. In *Proceedings of ICWSM*.

John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. 2008. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *Proceedings of ICML*.

Kevin Duh, Katsuhito Sudoh, Hajime Tsukada, Hideki Isozaki, and Masaaki Nagata. 2010. n-best reranking by multitask learning. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics*.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model of geographic lexical variation. In *Proceedings of EMNLP*.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Joshua Goodman. 2004. Exponential priors for maximum entropy models. In *Proceedings of NAACL-HLT*.

Daniel E. Johnson. 2009. Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass*, 3(1):359–383.

Jun'ichi Kazama and Jun'ichi Tsujii. 2003. Evaluation and extension of maximum entropy models with inequality constraints. In *Proceedings of EMNLP*.

William Labov. 1966. *The Social Stratification of English in New York City*. Center for Applied Linguistics.

Han Liu, Mark Palatucci, and Jian Zhang. 2009. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In *Proceedings of ICML*.

John Nerbonne. 2009. Data-driven dialectology. *Language and Linguistics Compass*, 3(1):175–198.

Brendan O'Connor, Jacob Eisenstein, Eric P. Xing, and Noah A. Smith. 2010. A mixture model of demographic lexical variation. In *Proceedings of NIPS Workshop on Machine Learning in Computational Social Science*.

Ariadna Quattoni, Xavier Carreras, Michael Collins, and Trevor Darrell. 2009. An efficient projection for $\ell_{1,\infty}$ regularization. In *Proceedings of ICML*.

John R. Rickford. 1999. *African American Vernacular English*. Blackwell.

Stefan Riezler and Alexander Vasserman. 2004. Incremental feature selection and ℓ_1 regularization for relaxed maximum-entropy modeling. In *Proceedings of EMNLP*.

Gerard Rushton, Marc P. Armstrong, Josephine Gittler, Barry R. Greene, Claire E. Pavlik, Michele M. West, and Dale L. Zimmerman, editors. 2008. *Geocoding Health Data: The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice*. CRC Press.

Hinrich Schütze and Jan Pedersen. 1993. A vector model for syntagmatic and paradigmatic relatedness. In *Proceedings of the 9th Annual Conference of the UW Centre for the New OED and Text Research*.

Aaron Smith and Lee Rainie. 2010. Who tweets? Technical report, Pew Research Center, December.

Berwin A. Turlach, William N. Venables, and Stephen J. Wright. 2005. Simultaneous variable selection. *Technometrics*, 47(3):349–363.

Larry Wasserman and Kathryn Roeder. 2009. High-dimensional variable selection. *Annals of Statistics*, 37(5A):2178–2201.

Larry Wasserman. 2003. *All of Statistics: A Concise Course in Statistical Inference*. Springer.

Jing Wu, Bernie Devlin, Steven Ringquist, Massimo Trucco, and Kathryn Roeder. 2010. Screen and clean: A tool for identifying interactions in genome-wide association studies. *Genetic Epidemiology*, 34(3):275–285.

Qing Zhang. 2005. A Chinese yuppie in Beijing: Phonological variation and the construction of a new professional identity. *Language in Society*, 34:431–466.

Kathryn Zickuhr and Aaron Smith. 2010. 4% of online Americans use location-based services. Technical report, Pew Research Center, November.