An Improved Method of Automated Nonparametric Content Analysis for Social Science: Supplementary Appendices

Connor T. Jerzak* Gary King† Anton Strezhnev‡

June 21, 2021

Abstract

Supplementary appendices for: Connor T. Jerzak, Gary King, and Anton Strezhnev. Working Paper. "An Improved Method of Automated Nonparametric Content Analysis for Social Science," Copy at http://j.mp/2DyLYxL.

*PhD Candidate and Carl J. Friedrich Fellow, 1737 Cambridge Street, Harvard University, Cambridge MA 02138, ConnorJerzak.com, cjerzak@g.harvard.edu.

†Albert J. Weatherhead III University Professor, Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; GaryKing.org, King@Harvard.edu, (617) 500-7570.

‡PhD Candidate, 1737 Cambridge Street, Harvard University, Cambridge MA 02138, antonstrezhnev.com, astrezhnev@fas.harvard.edu.

Contents

Appendix A Proofs of Analytical Results **2**

Appendix B Simulation Results **5**

Appendix C Objective Function Illustration **7**

Appendix D Applications to and Insights from Causal Inference **10**

Appendix A Proofs of Analytical Results

We now give proofs of propositions from Section 3.2.

Proof of Proposition 1

Start with the least-squares minimization problem

$$\widehat{\pi_1^{\mathrm{U}}} = \arg\min_{\pi_1^U} \sum_{w=1}^{W} \left(S_w^{\mathrm{U}} - \widehat{S_w^{\mathrm{U}}} \right)^2 .$$

Write $\widehat{S_w^{\mathrm{U}}}$ in terms of X^{L} and π_1^{U}

$$\widehat{\pi_1^{\mathrm{U}}} = \arg\min_{\pi_1^U} \sum_{w=1}^{W} \left(S_w^{\mathrm{U}} - X_{w2}^{\mathrm{L}} - \left(X_{w1}^{\mathrm{L}} - X_{w2}^{\mathrm{L}} \right) \pi_1^{\mathrm{U}} \right)^2 .$$

Take the derivative and set equal to 0

$$\begin{aligned}
0 &= \frac{\partial}{\partial \pi_1^{\mathrm{U}}} \sum_{w=1}^{W} \left(S_w^{\mathrm{U}} - X_{w2}^{\mathrm{L}} - \left(X_{w1}^{\mathrm{L}} - X_{w2}^{\mathrm{L}} \right) \pi_1^{\mathrm{U}} \right)^2 \\
&= \sum_{w=1}^{W} \left(S_w^{\mathrm{U}} - X_{w2}^{\mathrm{L}} \right) \left(X_{w1}^{\mathrm{L}} - X_{w2}^{\mathrm{L}} \right) - \left(X_{w1}^{\mathrm{L}} - X_{w2}^{\mathrm{L}} \right)^2 \pi_1^{\mathrm{U}} \\
\sum_{w=1}^{W} \left(S_w^{\mathrm{U}} - X_{w2}^{\mathrm{L}} \right) \left(X_{w1}^{\mathrm{L}} - X_{w2}^{\mathrm{L}} \right) &= \sum_{w=1}^{W} \left(X_{w1}^{\mathrm{L}} - X_{w2}^{\mathrm{L}} \right)^2 \pi_1^{\mathrm{U}} \\
\frac{\sum_{w=1}^{W} \left(S_w^{\mathrm{U}} - X_{w2}^{\mathrm{L}} \right) \left(X_{w1}^{\mathrm{L}} - X_{w2}^{\mathrm{L}} \right)}{\sum_{w=1}^{W} \left(X_{w1}^{\mathrm{L}} - X_{w2}^{\mathrm{L}} \right)^2} &= \pi_1^{\mathrm{U}} .
\end{aligned}$$

Since the expression being optimized is quadratic, this is a global optimum. Therefore the readme estimator in two categories has the closed-form expression:

$$\widehat{\pi_1^{\mathrm{U}}} = \frac{\sum_{w=1}^{W} \left(X_{w1}^{\mathrm{L}} - X_{w2}^{\mathrm{L}} \right) \left(S_w^{\mathrm{U}} - X_{w2}^{\mathrm{L}} \right)}{\sum_{w=1}^{W} \left(X_{w1}^{\mathrm{L}} - X_{w2}^{\mathrm{L}} \right)^2} .$$

Proof of Proposition 2

Start with the expression for $\widehat{\pi_1^{\mathrm{U}}}$:

$$\widehat{\pi_1^{\mathrm{U}}} = \frac{\sum_{w=1}^{W} \left(X_{w1}^{\mathrm{L}} - X_{w2}^{\mathrm{L}} \right) \left(S_w^{\mathrm{U}} - X_{w2}^{\mathrm{L}} \right)}{\sum_{w=1}^{W} \left(X_{w1}^{\mathrm{L}} - X_{w2}^{\mathrm{L}} \right)^2} .$$

Write X^{L}_{wc} in terms of X^{U}_{wc} and ϵ_{wc}.

$$\widehat{\pi^{\mathrm{U}}_1} = \frac{\sum_{w=1}^{W}\left(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2} + \epsilon_{w1} - \epsilon_{w2}\right)\left(S^{\mathrm{U}}_{w} - X^{\mathrm{U}}_{w2} - \epsilon_{w2}\right)}{\sum_{w=1}^{W}\left(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2} + \epsilon_{w1} - \epsilon_{w2}\right)^2}.$$

Substitute the accounting identity for S^{U}_{w}

$$\widehat{\pi^{\mathrm{U}}_1} = \frac{\sum_{w=1}^{W}\left(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2} + \epsilon_{w1} - \epsilon_{w2}\right)\left((X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2})\pi^{\mathrm{U}}_1 - \epsilon_{w2}\right)}{\sum_{w=1}^{W}\left(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2} + \epsilon_{w1} - \epsilon_{w2}\right)^2}.$$

Expand the numerator

$$\widehat{\pi^{\mathrm{U}}_1} = \frac{\sum_{w=1}^{W}\left(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2} + \epsilon_{w1} - \epsilon_{w2}\right)\left(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2}\right)}{\sum_{w=1}^{W}\left(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2} + \epsilon_{w1} - \epsilon_{w2}\right)^2}\pi^{\mathrm{U}}_1 - \frac{\sum_{w=1}^{W}\left(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2} + \epsilon_{w1} - \epsilon_{w2}\right)\left(\epsilon_{w2}\right)}{\sum_{w=1}^{W}\left(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2} + \epsilon_{w1} - \epsilon_{w2}\right)^2}.$$

and take expectation

$$E\left[\widehat{\pi^{\mathrm{U}}_1}\right] = E\left[\frac{\sum_{w=1}^{W}\left(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2} + \epsilon_{w1} - \epsilon_{w2}\right)\left(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2}\right)}{\sum_{w=1}^{W}\left(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2} + \epsilon_{w1} - \epsilon_{w2}\right)^2}\right]\pi^{\mathrm{U}}_1 - E\left[\frac{\sum_{w=1}^{W}\left(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2} + \epsilon_{w1} - \epsilon_{w2}\right)\left(\epsilon_{w2}\right)}{\sum_{w=1}^{W}\left(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2} + \epsilon_{w1} - \epsilon_{w2}\right)^2}\right].$$

Using the first-order Taylor approximation $E\left[\frac{X}{Y}\right] \approx \frac{E[X]}{E[Y]}$, we have

$$\begin{aligned}
E\left[\widehat{\pi^{\mathrm{U}}_1}\right] &\approx \frac{E\left[\sum_{w=1}^{W}\left(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2} + \epsilon_{w1} - \epsilon_{w2}\right)\left(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2}\right)\right]}{E\left[\sum_{w=1}^{W}\left(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2} + \epsilon_{w1} - \epsilon_{w2}\right)^2\right]}\pi^{\mathrm{U}}_1 - \frac{E\left[\sum_{w=1}^{W}\left(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2} + \epsilon_{w1} - \epsilon_{w2}\right)\left(\epsilon_{w2}\right)\right]}{E\left[\sum_{w=1}^{W}\left(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2} + \epsilon_{w1} - \epsilon_{w2}\right)^2\right]} \\
&= \frac{\sum_{w=1}^{W}\left(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2}\right)^2}{\sum_{w=1}^{W}(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2})^2 + E[(\epsilon_{w1} - \epsilon_{w2})^2]}\pi^{\mathrm{U}}_1 - \frac{\sum_{w=1}^{W} E[\epsilon_{w1}\epsilon_{w2}] - E[(\epsilon_{w2})^2]}{\sum_{w=1}^{W}(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2})^2 + E[(\epsilon_{w1} - \epsilon_{w2})^2]} \\
&= \frac{\sum_{w=1}^{W}\left(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2}\right)^2\pi^{\mathrm{U}}_1 - E[\epsilon_{w1}\epsilon_{w2}] + E[(\epsilon_{w2})^2]}{\sum_{w=1}^{W}(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2})^2 + E[(\epsilon_{w1} - \epsilon_{w2})^2]} \\
&= \frac{\sum_{w=1}^{W}\left(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2}\right)^2\pi^{\mathrm{U}}_1 + \mathrm{Var}(\epsilon_{w2}) - \mathrm{Cov}(\epsilon_{w1}, \epsilon_{w2})}{\sum_{w=1}^{W}(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2})^2 + \mathrm{Var}(\epsilon_{w1} - \epsilon_{w2})} \\
&= \frac{\sum_{w=1}^{W}\left(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2}\right)^2\pi^{\mathrm{U}}_1 + \mathrm{Var}(\epsilon_{w2}) - \mathrm{Cov}(\epsilon_{w1}, \epsilon_{w2})}{\sum_{w=1}^{W}(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2})^2 + \mathrm{Var}(\epsilon_{w1}) + \mathrm{Var}(\epsilon_{w2}) - 2\mathrm{Cov}(\epsilon_{w1}, \epsilon_{w2})},
\end{aligned}$$

where the last two lines follow from the definition of variance and the assumption that $E[\epsilon_{wc}] = 0$. Subtracting π^{U}_1 to get the bias:

$$\begin{aligned}
\mathrm{Bias}(\widehat{\pi^{\mathrm{U}}_1}) &\approx \frac{\sum_{w=1}^{W}\left(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2}\right)^2\pi^{\mathrm{U}}_1 + \mathrm{Var}(\epsilon_{w2}) - \mathrm{Cov}(\epsilon_{w1}, \epsilon_{w2})}{\sum_{w=1}^{W}(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2})^2 + \mathrm{Var}(\epsilon_{w1}) + \mathrm{Var}(\epsilon_{w2}) - 2\mathrm{Cov}(\epsilon_{w1}, \epsilon_{w2})} - \pi^{\mathrm{U}}_1 \\
&= \frac{\sum_{w=1}^{W}\mathrm{Var}(\epsilon_{w2}) - \mathrm{Cov}(\epsilon_{w1}, \epsilon_{w2}) - \mathrm{Var}(\epsilon_{w1})\pi^{\mathrm{U}}_1 - \mathrm{Var}(\epsilon_{w2})\pi^{\mathrm{U}}_1 + 2\mathrm{Cov}(\epsilon_{w1}, \epsilon_{w2})\pi^{\mathrm{U}}_1}{\sum_{w=1}^{W}(X^{\mathrm{U}}_{w1} - X^{\mathrm{U}}_{w2})^2 + \mathrm{Var}(\epsilon_{w1}) + \mathrm{Var}(\epsilon_{w2}) - 2\mathrm{Cov}(\epsilon_{w1}, \epsilon_{w2})}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\sum_{w=1}^{W} \text{Var}(\epsilon_{w2})(1-\pi_1^{\text{U}}) - \text{Var}(\epsilon_{w1})\pi_1^{\text{U}} - \text{Cov}(\epsilon_{w1},\epsilon_{w2}) + 2\text{Cov}(\epsilon_{w1},\epsilon_{w2})\pi_1^{\text{U}}}{\sum_{w=1}^{W}(X_{w1}^{\text{U}} - X_{w2}^{\text{U}})^2 + \text{Var}(\epsilon_{w1}) + \text{Var}(\epsilon_{w2}) - 2\text{Cov}(\epsilon_{w1},\epsilon_{w2})} \\
&= \frac{\sum_{w=1}^{W} \text{Var}(\epsilon_{w2})(1-\pi_1^{\text{U}}) - \text{Var}(\epsilon_{w1})\pi_1^{\text{U}} - \text{Cov}(\epsilon_{w1},\epsilon_{w2})(1-\pi_1^{\text{U}}) + \text{Cov}(\epsilon_{w1},\epsilon_{w2})\pi_1^{\text{U}}}{\sum_{w=1}^{W}(X_{w1}^{\text{U}} - X_{w2}^{\text{U}})^2 + \text{Var}(\epsilon_{w1}) + \text{Var}(\epsilon_{w2}) - 2\text{Cov}(\epsilon_{w1},\epsilon_{w2})} \\
&= \frac{\sum_{w=1}^{W} \left(\text{Var}(\epsilon_{w2}) - \text{Cov}(\epsilon_{w1},\epsilon_{w2})\right)(1-\pi_1^{\text{U}}) - \left(\text{Var}(\epsilon_{w1}) - \text{Cov}(\epsilon_{w1},\epsilon_{w2})\right)\pi_1^{\text{U}}}{\sum_{w=1}^{W}(X_{w1}^{\text{U}} - X_{w2}^{\text{U}})^2 + \text{Var}(\epsilon_{w1}) + \text{Var}(\epsilon_{w2}) - 2\text{Cov}(\epsilon_{w1},\epsilon_{w2})}.
\end{aligned}
$$

Proof of Proposition 3

Substituting in the known measurement error variances and assuming independence in measurement errors across categories yields:

$$
\text{Bias}(\widehat{\pi_1^{\text{U}}}) \approx \frac{\sum_{w=1}^{W}\left(\frac{\sigma_{w2}^2}{N_2^{\text{L}}}\right)(1-\pi_1^{\text{U}}) - \left(\frac{\sigma_{w1}^2}{N_1^{\text{L}}}\right)\pi_1^{\text{U}}}{\sum_{w=1}^{W}(X_{w1}^{\text{U}} - X_{w2}^{\text{U}})^2 + \left(\frac{\sigma_{w2}^2}{N_2^{\text{L}}}\right) + \left(\frac{\sigma_{w1}^2}{N_1^{\text{L}}}\right)}
$$

Using the fact that $N_c^{\text{L}} = N^{\text{L}}\pi_c^{\text{L}}$

$$
\begin{aligned}
\text{Bias}(\widehat{\pi_1^{\text{U}}}) &\approx \frac{(1-\pi_1^{\text{U}})\sum_{w=1}^{W}\left(\frac{\sigma_{w2}^2}{N^{\text{L}}\pi_2^{\text{L}}}\right) - \pi_1^{\text{U}}\sum_{w=1}^{W}\left(\frac{\sigma_{w1}^2}{N^{\text{L}}\pi_1^{\text{L}}}\right)}{\sum_{w=1}^{W}(X_{w1}^{\text{U}} - X_{w2}^{\text{U}})^2 + \left(\frac{\sigma_{w2}^2}{N^{\text{L}}\pi_2^{\text{L}}}\right) + \left(\frac{\sigma_{w1}^2}{N^{\text{L}}\pi_1^{\text{L}}}\right)} \\
&\approx \frac{\frac{1}{N^{\text{L}}}\left[\frac{(1-\pi_1^{\text{U}})}{(1-\pi_1^{\text{L}})}\sum_{w=1}^{W}\sigma_{w2}^2 - \frac{\pi_1^{\text{U}}}{\pi_1^{\text{L}}}\sum_{w=1}^{W}\sigma_{w1}^2\right]}{\sum_{w=1}^{W}(X_{w1}^{\text{U}} - X_{w2}^{\text{U}})^2 + \frac{1}{N^{L}(1-\pi_1^{\text{L}})}\sum_{w=1}^{W}\sigma_{w2}^2 + \frac{1}{N^{L}\pi_1^{\text{L}}}\sum_{w=1}^{W}\sigma_{w1}^2} \\
&\approx \frac{\frac{(1-\pi_1^{\text{U}})}{(1-\pi_1^{\text{L}})}\sum_{w=1}^{W}\sigma_{w2}^2 - \frac{\pi_1^{\text{U}}}{\pi_1^{\text{L}}}\sum_{w=1}^{W}\sigma_{w1}^2}{N^{\text{L}}\sum_{w=1}^{W}(X_{w1}^{\text{U}} - X_{w2}^{\text{U}})^2 + \frac{1}{(1-\pi_1^{\text{L}})}\sum_{w=1}^{W}\sigma_{w2}^2 + \frac{1}{\pi_1^{\text{L}}}\sum_{w=1}^{W}\sigma_{w1}^2}
\end{aligned}
$$

The denominator is strictly positive. Therefore, bias is minimized when the numerator is equal to 0. Solving for π_1^{U} in terms of π_1^{L} yields

$$
\begin{aligned}
0 &= \frac{(1-\pi_1^{\text{U}})}{(1-\pi_1^{\text{L}})}\sum_{w=1}^{W}\sigma_{w2}^2 - \frac{\pi_1^{\text{U}}}{\pi_1^{\text{L}}}\sum_{w=1}^{W}\sigma_{w1}^2 \\
\frac{\pi_1^{\text{U}}}{\pi_1^{\text{L}}}\sum_{w=1}^{W}\sigma_{w1}^2 &= \frac{(1-\pi_1^{\text{U}})}{(1-\pi_1^{\text{L}})}\sum_{w=1}^{W}\sigma_{w2}^2 \\
\pi_1^{\text{U}}(1-\pi_1^{\text{L}})\sum_{w=1}^{W}\sigma_{w1}^2 &= (1-\pi_1^{\text{U}})\pi_1^{\text{L}}\sum_{w=1}^{W}\sigma_{w2}^2 \\
\left(\pi_1^{\text{U}} - \pi_1^{\text{U}}\pi_1^{\text{L}}\right)\sum_{w=1}^{W}\sigma_{w1}^2 &= \left(\pi_1^{\text{L}} - \pi_1^{\text{U}}\pi_1^{\text{L}}\right)\sum_{w=1}^{W}\sigma_{w2}^2
\end{aligned}
$$

$$\pi_1^{\mathrm{U}} \sum_{w=1}^{W} \sigma_{w1}^2 = \pi_1^{\mathrm{L}} \sum_{w=1}^{W} \sigma_{w2}^2 + \pi_1^{\mathrm{U}} \pi_1^{\mathrm{L}} \left(\sum_{w=1}^{W} \sigma_{w1}^2 - \sigma_{w2}^2 \right)$$

$$\frac{\pi_1^{\mathrm{U}} \sum_{w=1}^{W} \sigma_{w1}^2}{\sum_{w=1}^{W} \sigma_{w2}^2 + \pi_1^{\mathrm{U}} \left(\sum_{w=1}^{W} \sigma_{w1}^2 - \sigma_{w2}^2 \right)} = \pi_1^{\mathrm{L}}$$

$$\frac{\pi_1^{\mathrm{U}} \sum_{w=1}^{W} \sigma_{w1}^2}{\pi_1^{\mathrm{U}} \sum_{w=1}^{W} \sigma_{w1}^2 + (1 - \pi_1^{\mathrm{U}}) \sum_{w=1}^{W} \sigma_{w2}^2} = \pi_1^{\mathrm{L}}$$

When the measurement error variances are generally constant across categories, the bias is zero when the labeled set proportions are equal to the unlabeled set proportions.

Appendix B Simulation Results

We illustrate our analytical results with simulation to build intuition for Section 3. We have found that simulations with large values of C and W generate more complexity without much additional insight, and so we set $C = 2$ and $W = 2$. We then evaluate readme performance while varying the degree of semantic change, textual discrimination, and proportion divergence. Our improvements to readme exploit the relationship between these three quantities.

For the purpose of our simulations, define category distinctiveness, or columns of X, as $(b_1 + b_2)/2$, where the absolute differences between categories is $b_w = |X_{wc} - X_{wc'}|$ for row $w = 1, 2$. Then, we define feature distinctiveness as $|b_1 - b_2|/2$, which is is the distinctiveness between rows of X. For our error metric, we use the sum of the absolute errors over categories (SAE), averaged over simulations.[1]

We draw our simulations as follows. First, we control proportion divergence by drawing π^{L} and π^{U} from i.i.d. Dirichlet distributions with concentration parameters set to 1. (In our figures, we measure average proportion divergence as $(|\pi_1^{\mathrm{L}} - \pi_1^{\mathrm{U}}| + |\pi_2^{\mathrm{L}} - \pi_2^{\mathrm{U}}|)/2$.) We sample X_{wc}^{U} from an i.i.d. Normal with mean 0 and variance $1/9$. Then, we generate $X_{wc}^{\mathrm{L}} = X_{wc}^{\mathrm{U}} + \epsilon$, where $\epsilon = 0$ or, to simulate semantic change, from a Normal with a mean at 0 and standard deviation proportional to $|X_{wc}^{\mathrm{U}}|$. We then treat these parameters as fixed

[1] We use the SAE to make our analysis consistent across data sets with different numbers of categories. We have found that simple attempts to normalize, such as dividing by the number of categories, tends to weaken this comparability, especially because in all cases the target quantity is located on the simplex.

and, to simulate measurement error in the calculation of X^L, generate 5,000 repeated sample data sets from each set of parameters, apply the readme estimator, and estimate the mean over simulations of SAE. To generate each of the 5,000 sampled data sets, we randomly generate document-level features from Normal densities by adding a draw from a standard Normal to each cell value of X^{L}_{wc} and X^{U}_{wc}.

Figure 1 illustrates how the SAE behaves as a function of category distinctiveness (vertically) and proportion divergence (horizontally). SAE is coded in colors from low (blue) to high (yellow). The bottom right of the figure is where readme performance is best: where proportion divergence is low and category distinctiveness is high. When the language is clearly distinguishable among categories, readme can overcome even large divergences between the labeled and unlabeled sets. Without high levels of textual discrimination, readme then becomes vulnerable to high levels of proportion divergence. Category distinctiveness and proportion divergence appear to have roughly the same relative importance, as the contour lines in Figure 1 are not far from $45°$ angles.

Figure 2 studies textual discrimination further by illustrating how category distinctiveness (horizontally) and feature distinctiveness (vertically) jointly impact SAE. If we hold feature distinctiveness fixed, increased category distinctiveness improves performance; if we hold category distinctiveness fixed, greater feature distinctiveness similarly leads to better performance over most of the range. Of the two, feature distinctiveness is somewhat more predictive of performance, especially for low category distinctiveness.

Finally, Figure 3 illustrates how the relationship between feature distinctiveness (three separate lines in each panel) and proportion divergence (horizontal axis) is mediated by the presence of semantic change (difference between the panels). Without semantic change (left panel), highly distinctive features greatly reduce SAE (which can be seen by the wide separation among the lines on the graph). In contrast, in the presence of semantic change (in this case we moved the mean of $E(X^{\mathrm{L}})$ by a quarter of a standard deviation from X^{U}), more distinctive features still tend to outperform less distinctive features, but the difference is less pronounced. With semantic change, distinctive features in the la-

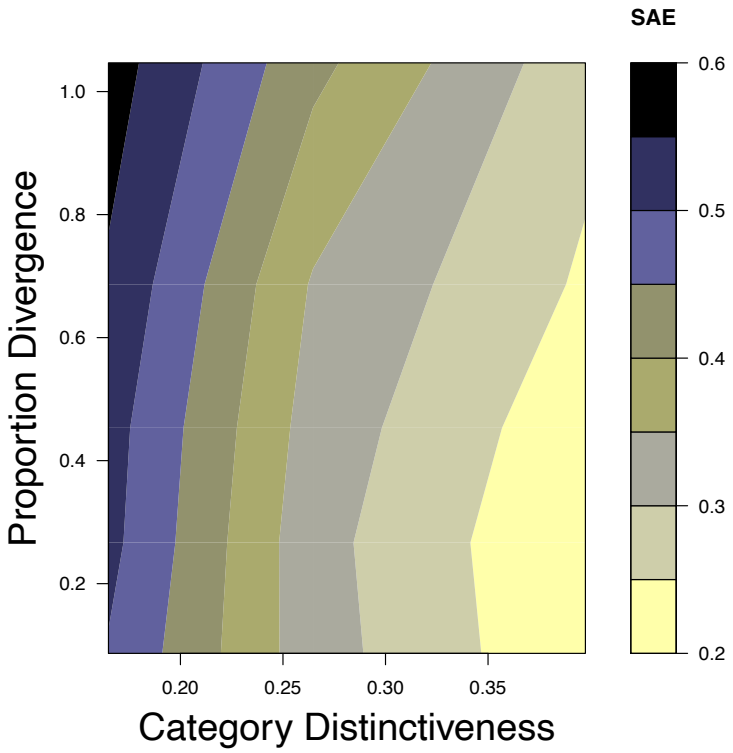

Figure 1: Category distinctiveness is plotted (horizontally) by proportion divergence between the labeled and unlabeled sets is plotted (vertically), with mean absolute sum of errors color coded (with yellow in the lower right corner best).

beled set may no longer be distinctive in the unlabeled set or may in fact be misleading about documents' category membership.

Appendix C Objective Function Illustration

For our application, it is crucial that our objective function incorporates both category and feature distinctiveness. Optimizing Γ for category distinctiveness alone would lead to high variance through high collinearity in $\underline{X}$, and optimizing Γ for feature distinctiveness alone would lead to higher bias and low category distinctiveness. Optimizing both together, as we do, reduces the overall error rate. We illustrate this point with the 1,426 emails drawn from the Enron corpora.

For expository clarity, we set $W' = 2$ and choose Γ by first maximizing the cat-

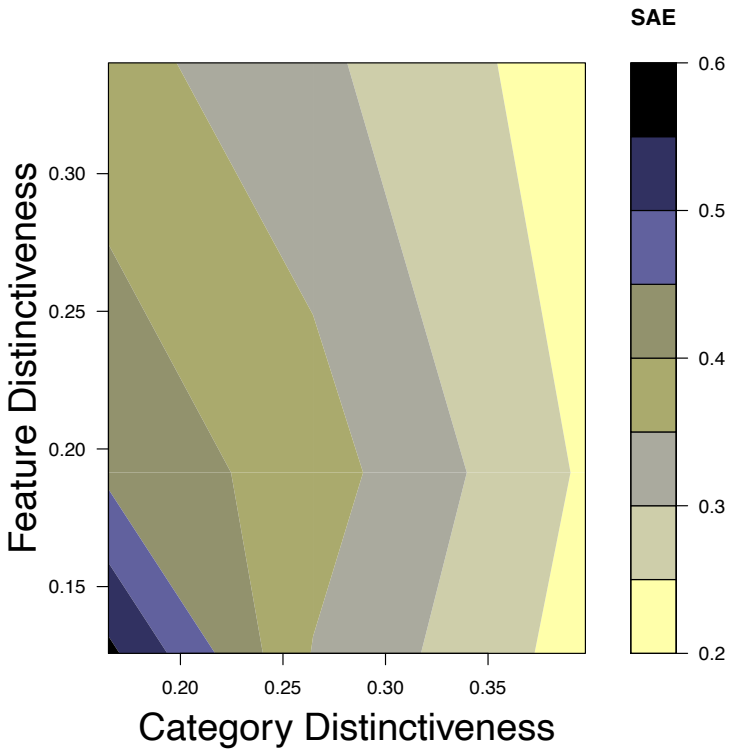

Figure 2: Category distinctiveness is plotted (horizontally) and feature distinctiveness (vertically), with mean absolute sum of errors color coded (with yellow indicating best performance).

egory distinctiveness metric alone. We offer a scatterplot of the resulting projections, $\underline{F}$, in the left panel of Figure 4, with different colors and symbols to represent the five categories. This panel reveals that these features do indeed discriminate between the categories (which can be seen by separation between the different colored symbols). However, as is also apparent, the two dimensions are highly correlated which, as in linear regression, would lead to higher variance estimates. In linear regression, given a fixed sample size, collinearity is an immutable fact of the fixed data set and specification; in contrast, in our application operating in the space of the features that we construct rather than the data we are given, we can change the projections, the space in which the regression is operating, and therefore the level of collinearity. As a second illustration, we again set $W' = 2$ but now optimize Γ by maximizing only feature distinctiveness. In this case, as can be seen in the middle panel of Figure 4, the columns of $\underline{X}^{\mathrm{L}}$ are uncorrelated but unfortunately do not

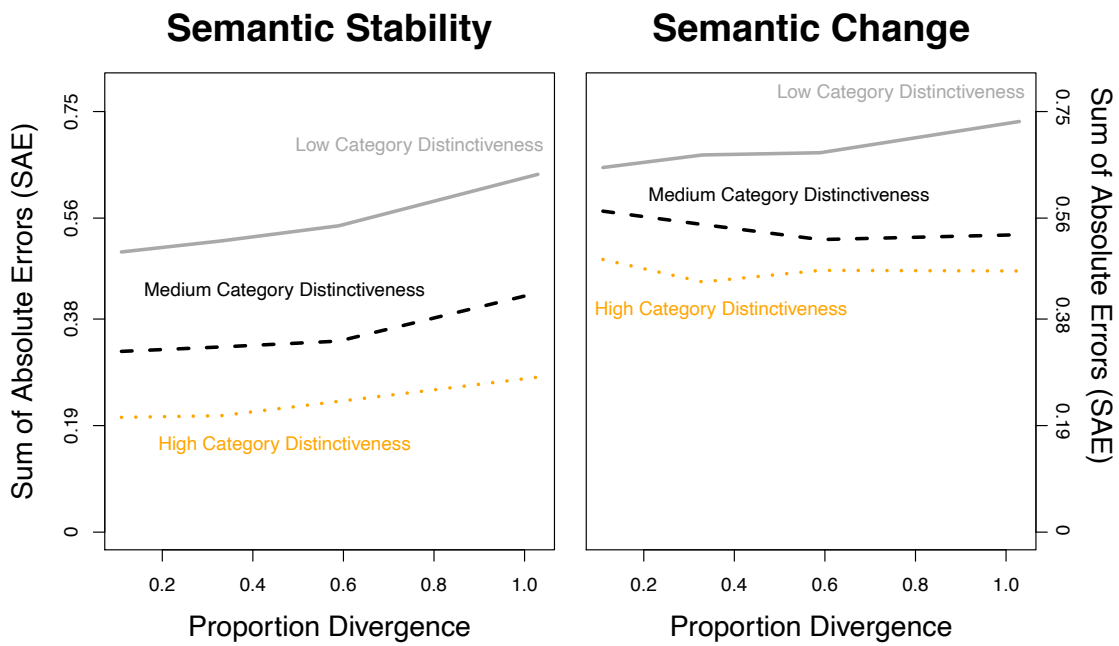

Figure 3: Proportion divergence is plotted (horizontally) by the mean of the sum of the absolute error (vertically) for different levels of category distinctiveness (separate lines) and for the absence (left panel) and presence (right panel) of semantic change.

discriminate between categories well (as can be seen by the points with different colors and symbols overlapping).

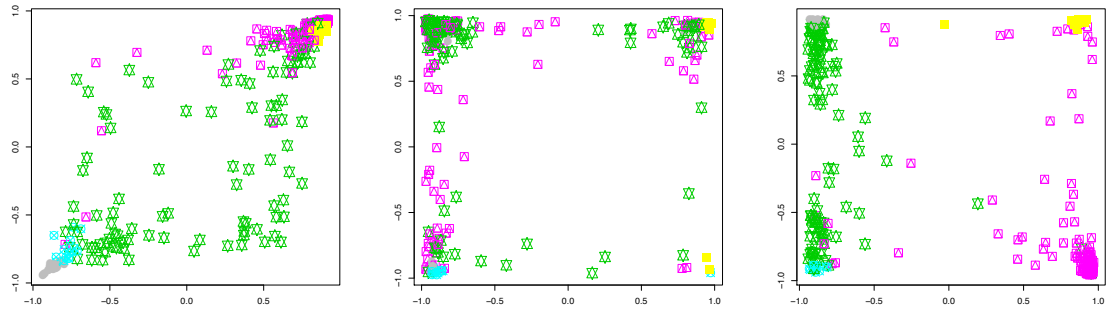

Figure 4: Optimizing Γ by category distinctiveness only (left panel), feature distinctivenss only (middle), and both (right panel). Each point is a document, with categories coded with a unique color and symbol. The axes (or components) are columns of tbe constructed $\phi(F \times \Gamma)$.

Thus, we implement our metric, optimizing the sum of both category and feature distinctiveness, which we present in the right panel of Figure 4. This result is well calibrated

for estimating category proportions: The dimensions are discriminatory and thus bias reducing, which can be seen by the color separation, but still uncorrelated, and thus variance reducing. After matching on these features and performing the least squares regression in these Enron data, we find the sum of the absolute errors (SAE) in estimating π^{U} of 0.28, compared to 0.50 for the original readme, a substantial improvement.

Empirically, also, we find that matching indeed has the desired effects: in the 73 real-world data sets we introduce in Section 4, matching alone reduces the divergence between $\underline{X}^{\mathrm{L}}$ and the true, unobserved $\underline{X}^{\mathrm{U}}$ in 99.6% of cases and on average by 19.8%. Proportion divergence, which is not observed in real applications but which we can measure because we have coded unlabeled sets for evaluation, is reduced in 83.2% of data sets, on average by 25%. We now turn to the details of these data, and the consequence of using all parts of our new methods, for mean square error in the quantities of interest.

Appendix D Applications to and Insights from Causal Inference

We discuss two areas where developments in this paper might profitably be applied to an area outside of automated text analysis.

First, our paper follows an implicit principle in defining Γ that is also implicit in several strains of the causal inference literature. We call this the "Tied Hands Principle" (THP), as it involves "tying ones hands" by ignoring certain types of certain information to avoid the possibility of biasing results in favor of one's favorite hypotheses, without regard to evidence in the data. In causal inference, the literature recommends matching treated and control observations without using the outcome variable (or at least not in both treatment regimes). Violating THP would enable the researcher to produce any estimate they wish, regardless of the evidence. For another example, in prospective designs, standard advice is to allow researchers to select observations conditional on the explanatory variables, but not the outcome variable. Similarly, in retrospective (case-control) designs,

researchers are advised to select observations based on the outcome variable but not on the explanatory variables. Violating these rules would also enable estimates without constraints from the data.

In our case, if we were to generate $\underline{\mathrm{X}}^{\mathrm{L}}$ in the labeled set by taking into account information about the resulting S^U in the unlabeled set, we could generate essentially any estimated unlabeled set category proportion $\hat{\pi}^U$ because we would be choosing both the left and right side of the readme regression. We therefore "tie our hands" by only including information from the labeled set in our optimization routine for Γ, which is reflected in the choice of objective function. Thus, the special case of the THP for readme2 prohibits finding Γ^* by minimizing $f(\Gamma, L, U)$.

We formalize a general version of THP, designed to apply to all these examples and others:

Tied Hands Principle (THP). *Let t denote a function that transforms data objects A and B, into $A^* = t(A, Z)$ and $B^* = t(B, Z)$, by matching or weighting data subsets (rows), or transforming or selecting features (columns), where Z denotes exogenous information such that $p(A, B|Z) = p(A, B)$. Denote as $T_{A,Z}$ the set of all functions of A and Z (but not B) and $T_{B,Z}$ the set of all functions of B and Z (but not A). Then define $g(j|k)$ as a function to choose an element from set j using information in k. Consider statistical procedures that are functions of both A^* and B^*. Then, for $D = A$ or $D = B$, THP requires that the transformation function t be chosen such that $t = g(T_{D,Z}|D, Z)$ but not $t = g(T_{D,Z}|A, B, Z)$.*

A second connection with the causal inference literature is our dimensionality reduction technique, which may have applications to matching studies, from which we originally borrowed inspiration for some of our ideas. An important issue in matching is the optimal feature space, such as in the space of predicted values for the outcome under the control intervention (to follow THP, as in "predictive mean matching"). One such feature space is the one derived here which might enable researchers to control dimensionality,

as well as the tradeoff between feature independence and informativeness. The resulting causal estimator could thus have attractive properties, since it would take into account both the relationship between the covariates and the outcome (leading to lower bias) while incorporating several independent sources of information (leading to lower variance).

To be more precise, consider the Average Treatment Effect on the Treated:

$$\tau = E[Y_i(1)|X_i, T_i = 1] - E[Y_i(0)|X_i, T_i = 1].$$

We could estimate $E[Y_i(0)|X_i, T_i = 1]$ directly using a regression model trained on the control units, predicting their outcomes. However, to avoid model dependence we could instead use non-parametric methods, matching each treated unit with the nearest (say) 3 control units and taking the average outcome value of those control units as an estimate of $E[Y_i(0)|X_i, T_i = 1]$. However, some variables on which we match may be unimportant, in that they are poorly predictive of the outcome. When outcomes are discrete, we can consider applying the technique developed above to generate feature projections for readme to instead create features for matching that are highly predictive of the outcome. We could do this by fitting a feature projection on units in the control group (to avoid violating the THP) that optimizes the degree of feature discrimination in Y_i. We can then apply this projection to all units and match on the new feature space.

In this arrangement, the background covariates function like the word vector summaries in our text analysis, and $Y_i(0)$ plays a role that category membership did before. Matching on features that generate a high-quality $E[\underline{\mathbf{X}}_i|Y_i(0)]$ matrix may seem strange. Yet, the exercise here is in many ways the natural one: either we make no assumptions about the covariate-outcome relationship (as in fully non-parametric matching), we condition on the background covariates and use this information to predict the outcome (as in the case of regression-adjusted inference), or we condition on the outcome data and re-weight the background covariates in a way that maximizes the distinctiveness of the resulting features (as in our proposed approach).

As a proof of concept, we conduct simulations beginning with only the control units from each of 9 prominent social science experiments (Bailey, Hopkins, and Rogers, 2016; Bolsen, Ferraro, and Miranda, 2013; Callen, De Mel, McIntosh, and Woodruff, 2014; Enos and Fowler, 2016; Finkelstein et al., 2012; Gerber and Green, 2000; Kugler, Kugler, Saavedra, and Prada, 2015; McClendon and Riedl, 2015; Taylor, Stein, Woods, and Mumford, 2011). In each replication data set, we assign half of the control units to receive a "treatment" of a size equal to a constant plus 0.1 times the standard deviation of the outcome (which we take to be the original dependent variable in the study). We assign this synthetic treatment probabilistically in a way that seeks to replicate the amount of confounding present in the original experiment. In particular, after fitting a non-parametric binary classifier to the original treatment/control status, we selected units into the synthetic treatment group with probabilities proportional to the resulting propensity scores.

We then compare the error of 5 different methods for extracting the causal effect between the synthetic treatment and control groups. These methods are each based on nearest neighbor matching and differ only in the feature space in which the matching takes place. We consider matching on estimated propensity scores, on random projections, and on the predicted values from a regression modeling the relationship between the data inputs and the outcome trained on the control data. We also considered matching on the original features, on random projections, and on the features generated from the readme2 algorithm. We repeated the synthetic experiment 100 times for each of the 9 replication data sets.

The results, which appear in Figure 5, suggest that the readme2 features perform well for the non-parametric estimation of causal effects. In 7 of the 9 data sets, estimation in the readme2 space yields the largest decrease in absolute error. In the other 2 cases, performance is still very good. It appears that the features from the readme2 algorithm can be profitably used in other contexts where non-parametric analyses will be done and where outcome data is available.

Finally, our dimensionality reduction technology may also be used for data visualiza-

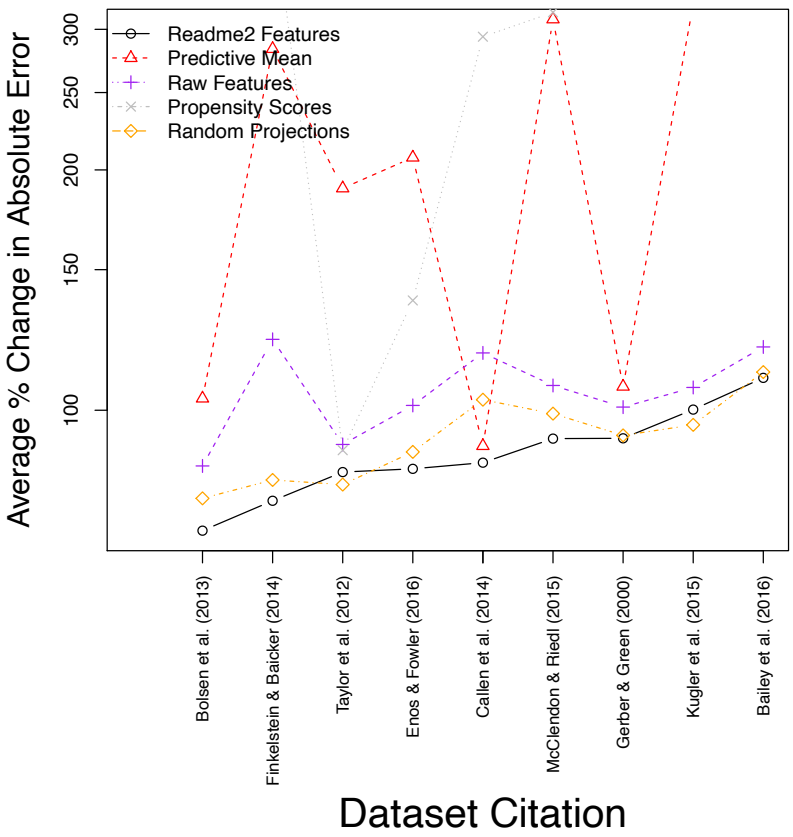

Figure 5: The change in absolute error in this graph is calculated relative to the error from the simple difference in means estimator. Points below "100" are those for which the average error of the matching estimator was lower than the naive difference in means estimator. Points above "100" are those for which the error was higher.

tion. For example, in visualizing data on partisanship, we could find the 2-dimensional projection that maximally discriminates between Democrats, Republicans, and Independents and simultaneously contains minimal redundancy. The relevant clusters would then become more visible, and could even be paired with a data clustering algorithm on the 2-dimensional projection for additional visualization or analysis purposes.

Additional References

Abadi, Martin et al. (2015): *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. URL: tensorflow.org.

Bailey, Michael, Daniel Hopkins, and Todd Rogers (2016): "Unresponsive and Unpersuaded: The Unintended Consequences of a Voter Persuasion Effort". In: *Political Behavior*, vol. 3, pp. 713–746.

Bolsen, Toby, Paul Ferraro, and Juan Jose Miranda (Aug. 2013): "Are Voters More Likely to Contribute to Other Public Goods? Evidence from a Large-Scale Randomized Policy Experiment". In: *ajps*, no. 1, vol. 58, pp. 17–30.

Callen, Michael, Suresh De Mel, Craig McIntosh, and Christopher Woodruff (2014): *What are the Headwaters of Formal Savings? Experimental Evidence from Sri Lanka*. NBER. URL: bit.ly/callen14.

Enos, Ryan and Anthony Fowler (Apr. 2016): "Aggregate Effects of Large-Scale Campaigns on Voter Turnout". In: *Political Science Research and Methods*, vol. 21, pp. 1–19.

Finkelstein, Amy et al. (Aug. 2012): "The Oregon Health Insurance Experiment: Evidence from the First Year". In: *The Quarterly Journal of Economics*, no. 3, vol. 127, pp. 1057–1106.

Gerber, Alan S. and Donald P. Green (Sept. 2000): "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment". In: *American Political Science Review*, no. 3, vol. 94, pp. 653–663.

Kugler, Adriana, Maurice Kugler, Juan Saavedra, and Luis Omar Herrera Prada (2015): *Long-term Direct and Spillover Effects of Job Training: Experimental Evidence from Colombia*. NBER. URL: bit.ly/KugKugS.

McClendon, Gwyneth and Rachel Beatty Riedl (Oct. 2015): "Religion as a stimulant of political participation: Experimental evidence from Nairobi, Kenya". In: *jop*, no. 4, vol. 77, pp. 1045–1057.

Pennington, Jeffrey, 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)*, pp. 1532–1543.

Taylor, Bruce, Nan Stein, Dan Woods, and Elizabeth Mumford (Oct. 2011): *Shifting Boundaries: Final Report on an Experimental Evaluation of a Youth Dating Violence Prevention Program in New York City Middle Schools*. Final Report. Police Executive Research Forum.