# Inverse Regression for Analysis of Sentiment in Text

Matthew A. Taddy
taddy@chicagobooth.edu
*The University of Chicago Booth School of Business*

ABSTRACT: Text data, including speeches, stories, and other document forms, is often composed with regard to *sentiment* variables that are of interest for research in marketing, economics, and other social research fields. It is also very high dimensional and difficult to incorporate into statistical analysis. This article introduces a straightforward framework of sentiment-preserving dimension reduction for text data. Our aim is to provide a general approach to text regression while avoiding the model complexity characterizing much of statistical learning for language. Using an inverse regression approach, we show that multinomial logistic regression of phrase counts onto document characteristics can be used to obtain low dimensional document representations that are rich in sentiment information. In addition to this text-specific work, the article introduces an estimation framework that should be generally applicable for high-dimensional logistic regression. In particular, we propose independent Laplace priors for each coefficient loading and advocate joint MAP estimation of coefficients and the associated prior scale. It is shown that this scheme yields a novel and stable approach to nonconcave penalized likelihood estimation for logistic regression. We also survey related approaches from the literature, connecting econometric methodology to partial least squares and linking both to inverse regression and related frameworks. Finally, the work is motivated through three detailed examples and we provide an out-of-sample prediction study to illustrate the effectiveness of our methods.

M. A. Taddy is Assistant Professor of Econometrics and Statistics and Neubauer Family Faculty Fellow at the University of Chicago Booth School of Business, and this work was partially supported by the IBM Corporation Faculty Research Fund at Chicago Booth. The author thanks Matthew Gentzkow, Jesse Shapiro, Robert Gramacy, and Nicholas Polson for much helpful discussion on this material.

# 1 Introduction

Text data, including stories, speeches, emails, and many other document forms, is a vast source of information that can often be connected to variables of interest for research in marketing, economics, and other social science fields. It thus has enormous potential as a kind of soft data – as a stand-in for related variables of primary interest, which we generically term document *sentiment*. On the other hand, text data is inherently of very high dimension and thus difficult to incorporate into statistical analysis. This article introduces a straightforward framework of sentiment-preserving dimension reduction for text data, which we hope will be useful for document visualization and as a component of more comprehensive models.

An overview of the field of *sentiment analysis*, from the perspective of information retrieval researchers, is provided by Pang and Lee (2008). There is a lack of consensus on terminology for this area, with *opinion mining* and *subjectivity analysis* both competing against *sentiment analysis* as field title and various definitions abound for specific jargon. We focus on the term *sentiment* which, having possible definition as "sensible quality", seems more general than alternative labels (despite being used by some authors to refer only to opinion polarity). In particular, *sentiment* is defined here as an observable characteristic linked to the variables motivating document language choice. For example, we consider restaurant reviews accompanied by ratings (1-5 stars) which are directly connected to a writer's motivation in composing their review. In a less straightforward example, we consider political speeches from members of the US congress and take voting behavior of their constituents as an associated sentiment.

Much of existing research on sentiment analysis revolves around application of generic regression and classification algorithms to word count data (see examples in Pang and Lee, 2008; Srivastava and Sahami, 2009; Joshi et al., 2010, as well as those mentioned in Section 2.3). Alternatively, the supervised latent Dirichlet allocation (SLDA) of Blei and McAuliffe (2010) is part of a recent push in machine learning to account for sentiment in latent-factor document representations. At the other end of the spectrum, this article was originally motivated by *slant* indexing for political ideology in text, an empirical technique from research by Gentzkow and Shapiro (2010) on the economics of media bias. Our goal is to, like SLDA and slant, present a regression framework that takes advantage of the specific nature of text data. However, unlike slant, we do so with foundation in a probability model and, unlike SLDA, avoid modeling for a high-dimensional latent variable expansion.

The framework proposed here is founded on the method of inverse regression: knowledge of the distribution for $\mathbf{X} \mid y$ is used to find a lower dimensional covariate representation which preserves information relevant to $y$. Providing this inverse distribution introduces powerful information, and hence efficiency, into the estimation. While the difficulty in applying inverse regression techniques usually rests in building the high-dimensional covariate distribution, we will show that text data presents a unique situation where there is an obvious fully specified inverse representation. In particular, multinomial logistic regression for phrase-counts can be used to map phrase-frequencies onto a sentiment-preserving document reduction (i.e., a small set of document-sentiment scores). This inverse regression framework is detailed in Section 3, with background on sufficient reduction ideas in 3.1, followed by a probabilistic characterization of text and sentiment data in 3.2, the resulting multinomial regression and sufficiency arguments in 3.3, and specification of sparsity inducing Laplace coefficient priors in 3.4.

After the modeling and sufficient reduction are established, we are left with an extremely high dimensional multinomial logistic regression problem, where response dimension is equal to the number of distinct terms across documents. Section 4 outlines estimation in this setting and, combined with the sparsity-inducing priors of 3.4, provides a complete framework for general applications of high-dimensional multinomial regression. First, 4.1 describes and adapts algorithms for maximum likelihood estimation in the setting of text regression. This serves as background for Section 4.2, which introduces a novel algorithm for joint *maximum a posteriori* (MAP) estimation of both coefficients and prior scale parameters. One innovation of this procedure is to replace each prior scale with its conditional solution during estimation, leading to an efficient lower-dimensional algorithm. The resulting regularization is an instance of nonconcave likelihood penalization with unbiasedness for large coefficients (e.g., Fan and Li, 2001), showing that such penalties can arrive naturally in MAP estimation of a hierarchical Bayesian model. In this context, Section 4.3 discusses prior specification and penalty behavior.

Section 2 details our motivating sentiment analysis applications: the political speeches and restaurant reviews mentioned above, as well as a third example involving business news and stock price. An out-of-sample prediction study for all examples, illustrating our methods and comparators from the literature, is in Section 5.1. The remainder of Section 5 describes full analysis results for each individual example. Finally, we close with a short discussion.

## 2 Motivating Datasets

Our analysis begins at text that has been parsed, stemmed, and generally cleaned into a list of informative phrases. This includes removing stop words (such as *and* or *but*) and stemming words with identical roots (e.g. *tax*, *taxation*, *taxing*, and *taxes*) into a single token. Although beyond the scope of this article, there is considerable flexibility in how this formatting is achieved (refer to, e.g., Jurafsky and Martin, 2009). The chosen number of words constituting a phrase (i.e, the $N$-gram, with $N$ the number of words) will be both goal and data dependent, and there is no NLP consensus on the optimal size for sentiment modeling; we encounter unigrams, bigrams, and a mix of bi and trigrams in our examples. Conditional on this parsing, each document is represented as a simple vector of the associated phrase-counts. Although this is very common in analysis of text data (see, e.g., Srivastava and Sahami, 2009), it may be possible to find alternative document representations that provide more useful information (e.g., accounting for order, or the semantic parsing of Poon and Domingos, 2009). Considering such alternatives, and adapting inverse regression for these richer text representations, is a future area of research.

### *2.1 Ideology in Political Speeches*

This example originally appears in Gentzkow and Shapiro (GS; 2010) and considers text of the 2005 Congressional Record, containing all speeches in that year for members of the United States House and Senate. GS record the number times each of 529 legislators used terms in a list of 1000 phrases. Each document corresponds to a single person, and associated sentiment is the two-party vote-share from each speaker's constituency (congressional district for representatives; state for senators) obtained by George W. Bush in the 2004 presidential election.

Following the political economy notion that there should be little discrepancy between voter and representative ideology, this vote-share provides a noisy measure of the specific type of ideology expressed in a Bush vote. Given that ideology can have more than one dimension (e.g., social and fiscal conservatism; see Poole and Rosenthal, 2007, for an overview of quantitative assessment of ideology), vote-share is a map from some unidentified space to the specific *Bush test* of ideology. For this reason, we also consider a speaker's first and second *common-score* values (from `voteview.com`) as higher-dimensional sentiment factors. Part of a widely used framework of spatial models for roll-call voting (Poole, 2005), common-score provides a unified measure over both House and Senate for the ideology of elected representatives. The first common-score has correlation of 0.77 with our Bush vote-share metric, and is generally taken to be the dominant dimension of ideology (Poole and Rosenthal, 2007).

The phrases in this example were chosen by pre-screening for partisanship the bigrams and trigrams in raw speeches. In particular, GS built contingency tables for phrase-usage by party membership, ranked the phrases according to Pearson $\chi^2$-test statistics, and used this ranking to obtain a set of the 1000 most partisan phrases. Full details are contained in the original article. This example is thus distinguished from our others in that phrases are chosen with reference to sentiment, whereas elsewhere phrases are selected based only on relative usage.

In their analysis, GS use the relationship between phrase-counts and vote-share to build a *slant* index for partisanship of text. This index is then fit for various US daily newspapers and applied to study the economics of bias and competition in media. Noting that our investigation was originally motivated in part by a desire to understand and generalize these slant indices, Appendix A.1 contains discussion on the technique. In particular, the slant index is equivalent to summation of phrase frequencies weighted by their correlation with the sentiment of interest. As such, slant is a first-order partial least-squares (PLS) procedure, and an improved slant arises from the input scaling usually applied in PLS. In turn, following from this appendix and the discussion in 3.1, PLS can be seen as a specific inverse regression procedure, providing a connection to our dimension reduction strategy in Section 3.

## 2.2 *On-Line Restaurant Reviews*

This application concerns 6260 restaurant reviews from the site `www.we8there.com`. The short user-submitted reviews (average of 90 words) are accompanied by a five-star rating on four specific aspects of restaurant quality – *food*, *service*, *value*, and *atmosphere* – as well as the *overall experience*. This example thus presents a classic and straightforward setting for sentiment analysis: the text is immediately relevant to an author-provided sentiment measure. We treat *overall experience* as the main sentiment of interest, but also consider the four aspect-ratings as potential expanded sentiment factors.

Mauá and Cozman (2009) use this data to illustrate their methodology for sorting words into topics under guidance of sentiment. We have processed the reviews by converting words to lowercase and removing punctuation, numbers, and a minimal list of common English stop-words. The Porter stemmer, as implemented in R's `Snowball` package, stems text into root terms that are then parsed into bigram phrases. After discarding phrases that appear less than ten times in the dataset and reviews which do not use any of the remaining phrases, we obtain for each of 6147 reviews a vector of counts for 2978 bigrams. Use of bigrams was based on a notion that modifiers (e.g., *very*, *high*, *tiny*) would be useful for evaluating restaurant reviews; however, a model based on unigrams leads to equivalent predictive performance.

## 2.3 *Business News and Stock Performance*

In our final example, we compare summary text for Wall Street Journal (WSJ) stories with *IBM* in the headline to performance for the stock of International Business Machines Corp. (IBM). In detail, headlines and one-sentence abstracts dating from August 1988 to August 2010 were retrieved from the ProQuest database. Under the same NLP steps as in Section 2.2, but with unigrams due to a more focused vocabulary, we obtain counts for 808 words in 2105 documents.

Our empirical sentiment measure for each article is two-day return-over-market for shares in IBM listed on the New York Stock Exchange, calculated from the opening of the previous day to market close on the day of publication (i.e., the S&P500 index return is subtracted from IBM's return for stock held over this two day period, and our sentiment is the difference standardized to have a mean of zero and variance of one). Since business news should be reflected in share-price, this return signals IBM sentiment information on that publication date (a similar logic is found in Koppel and Shtirmberg, 2004). Return-over-market isolates sentiment that is unique to IBM, and the two day window allows both for news already priced by the market (e.g., stories *about* price change) as well as news that affects market pricing. Saturday news is conflated with Monday news, and Friday's open is treated as the previous-day price for Monday articles.

Sorting news stories by sentiment is a prominent task in text mining, with business news receiving particular attention. In academic literature, Tetlock (2007) connects market activity to a WSJ column, Fortuna et al. (2009) discern media bias from the terminology in news reports, and Schumaker and Chen (2009) build a predictive model for stock returns from breaking financial news. This work all relies upon generic classification or regression tools, such as support vector machines, PCA, and neural networks. We do not here have the goal of predicting stock returns based upon news stories; rather, we are investigating whether dimension-reduction for text based on returns can help in automatic classification of articles by sentiment, and this presents a case-study of model performance under extremely weak and noisy sentiment signal.

## 3 Inverse Regression for Text Data

This section develops inverse logistic regression as a dimension reduction strategy for sentiment analysis. We show that with each document summarized through a length-$p$ vector of term counts, $\mathbf{X}$, and accompanied by $K$ observables related to sentiment, $\mathbf{V}$, inverse regression for text is just estimation of a multinomial logistic model for $\mathbf{X}$ regressed onto $\mathbf{V}$. Then, if $\mathbf{\Phi}$ is the $p \times K$ matrix of regression coefficients and $\mathbf{F}$ contains term frequency (counts divided by document totals), the reduced dimension document summary $\mathbf{F}'\mathbf{\Phi}$ is sufficient for sentiment – that is, it contains all information in $\mathbf{X}$ relevant to $\mathbf{V}$. Thus, for the researcher interested in sentiment analysis, these simple length-$K$ document scores (see $\mathbf{F}'\mathbf{\Phi}$ in figures 5, 7, and 9) can replace text-data in the inferential or prediction problem of interest. Before moving to our text-specific work, however, we first provide some background on the general idea of sufficient reduction through inverse regression.

## 3.1 *Sufficient Reduction*

In many contemporary regression problems, including text analysis, very high-dimensional covariates necessitate dimension reduction in the estimation procedure – it is not possible to efficiently estimate the conditional response distribution $y \mid \mathbf{X}$ without also thinking about how to simplify $\mathbf{X}$. Hence, the goal is to reduce dimension of $\mathbf{X}$ while retaining information relevant to $y \mid \mathbf{X}$. A *sufficient reduction* for $\mathbf{X}$ will achieve this without any loss of relevant information.

Cook's 2007 overview highlights the usefulness of inverse regression – inference for the multivariate distribution $\mathbf{X} \mid y$ – in obtaining a sufficient reduction. With $\mathbf{X}$ a vector of $p$ covariates, the generic linear inverse regression formulation can be written

$$\mathbf{X} = \mathbf{\Phi}\mathbf{V} + \mathbf{\Omega} \tag{1}$$

where $\mathbf{\Phi}$ is the $p \times K$ matrix of inverse regression coefficients, $\mathbf{\Omega}$ is an unspecified $p$-vector of error terms, and $\mathbf{V}$ is a $K$-vector of *response factors* which depend on $y$. Cook outlines conditions on $\mathbf{\Omega}$ such that $y$ is conditionally independent of $\mathbf{X}$ given $\mathbf{X}'\mathbf{\Phi}$, and thus that the *principal fitted components* in the $K$-vector $\mathbf{X}'\mathbf{\Phi}$ provide a sufficient reduction for $\mathbf{X}$. Several options are discussed for specification of $\mathbf{V}$, including a step-function expansion of $y$ which leads to the same reduction results as sliced inverse regression (Li, 1991).

Many dimension reduction approaches are related to models of the type in (1). Principal component regression (Massy, 1965) is shown by Cook to be the maximum likelihood solution when $\mathbf{V}$ is left unspecified. Following our discussion in Appendix A.1, PLS(1) is the inverse regression solution when $\mathbf{V} = y$ and the errors in $\mathbf{\Omega}$ are assumed independent, while higher-order PLS procedures provide stagewise OLS inverse regression. In a slight variation, single index models (e.g., Antoniadis et al., 2004) require joint estimation of both the inverse regression and a nonparametric forward regression model for $y \mid \mathbf{X}'\mathbf{\Phi}$.

As an alternative framework, both supervised principal component analysis (SPC; Blair et al., 2006) and Bayesian factor regression models (BFRM; West, 2003) augment (1) with a model for $\mathbf{y} \mid \mathbf{V}$. In the context of sentiment analysis, supervised latent Dirichlet allocation

(SLDA; Blei and McAuliffe, 2010) proposes generalized linear regression for $\mathbf{y} \mid \mathbf{V}$ where factors $\mathbf{V}$ are probability weights over topics (word distributions) in a document. Despite also incorporating the covariate distribution, members of this general class of *supervised factor models* lead to analysis that is distinct from inverse regression. Conceptually, supervised factor models are fitting jointly underlying latent factors while inverse regression is seeking a simplifying linear map for $\mathbf{X}$. On a practical level, BFRM, SPC, and SLDA infer both $\mathbf{V}$ and $\mathbf{\Phi}$, whereas $\mathbf{V}$ will typically be specified for inverse regression. When Cook's sufficiency conditions are satisfied, parsimonious inverse regression models should lead to more efficient inference. In particular, it is hypothesized that sufficient reduction techniques provide stronger dependence between $y$ and $\mathbf{X\Phi}$ than between $y$ and an estimated $\mathbf{V}$; this was empirically found to be true in Section 5.1, where we compare our inverse regression for text against SLDA.

Our methodology builds on the general idea of sufficient reduction, but is developed to account for specific demands of text regression. In a key point, much of the existing work on inverse regression requires covariates that are conditionally independent given response; for example, Cook and Li (2009) incorporate non-Gaussian predictor distributions through use of single-parameter exponential families. This conditional independence assumption is invalid for text data, with each covariate phrase-vector drawn from a single document, and we look to existing text models in building our inverse regression framework.

## *3.2 Document Representation*

Assuming the *bag-of-words* representation for each document, our text data is summarized by a *document-phrase* count matrix $\boldsymbol{\mathcal{X}}$, consisting of $n$ rows $\mathbf{X}_i$ containing document counts for each of $p+1$ phrases in the corpus. The counts map directly to a frequency matrix, $\boldsymbol{\mathcal{F}}$ with rows $\mathbf{F}_i$, after division by document term totals. Finally, the text data is accompanied by the $n \times K$ sentiment matrix $\boldsymbol{\mathcal{V}}$ with rows $\mathbf{V}_i$ containing observed *sentiment factors* for each document.

How one specifies these sentiment factors depends upon the analysis goals. In many problems, $\mathbf{V}$ is just the observed sentiment of interest; this can be a univariate response or a multidimensional vector of sentiment factors. The inverse regression resulting from such specification provides a document projection along each direction in $\mathbf{V}$, and will be useful for document visualization or for sentiment prediction. Alternatively, when focused on prediction for a single sentiment response $y$, $\mathbf{V}$ can consist of any variables for which $[y \mid \mathbf{V}, \mathbf{X}] \equiv [y \mid \mathbf{V}]$. Most simply, $K = 1$ and $\mathbf{V} = y$. More generally, one can seek expansions of $y$ which lead to multivariate document reductions; if regression onto $\mathbf{V}$ yields a more precise model for $\mathbf{X}$, and if $\mathbf{V}$ is truly sufficient for $y$, then this should provide a more informative document reduction.

Cook (2007) advocates a step-function expansion of $y$ among other options, while Mao et al. (2010) use mixture models to estimate $\mathbf{V}$. Both specifications are accompanied by a second estimated reduction (such that, with $\mathbf{U}$ an $r$-dimensional expansion of $y$, $\mathbf{V} = \mathbf{U}'\boldsymbol{\Psi}$ where $\boldsymbol{\Psi}$ is $r \times K$) and require matrix constraints for identification. Taking a different approach, we consider only the basic $\mathbf{V} = y$ specification or set $\mathbf{V}$ to other observable sentiment factors (e.g., ratings for food, service, atmosphere, and value when $y$ is an overall restaurant rating). While leading to more interpretable models and simpler estimation, this strategy is also motivated by the specific nature of text-sentiment data: there is often training information available that is very relevant to $y$ but will not be available when evaluating future documents. That said, our methodology is straightforward to combine with any specification for $\mathbf{V}$.

### *3.3 Multinomial Inverse Regression*

The bag-of-words representation is premised on exchangeability for document phrases, and thus implies that each $\mathbf{X}$ can be treated as generated from a multinomial distribution (as in the latent Dirichlet allocation framework of Blei et al., 2003, and its many extensions). For inverse regression, each document's multinomial probability vector can then be modeled as dependent upon the associated sentiment factors. In particular, assume a logistic multinomial regression framework such that, for each document $i$ containing $m_i$ phrase occurrences,

$$\mathbf{X}_i \sim \mathrm{MN}(m_i, \mathbf{q}_i) \;\; \text{with} \;\; \log\left(\frac{q_{ij}}{q_{i0}}\right) = \eta_{ij} = \alpha_j + \boldsymbol{\varphi}_j'\mathbf{V}_i, \;\; \text{for} \;\; j = 1, \ldots, p. \tag{2}$$

Note the existence of a null category, corresponding to probabilities $q_{i0}$: this can be set to any phrase in the corpus, although for interpretability it is convenient to choose a phrase with little impact on $y$. By assumption, we have $\eta_{i0} = 0$, $\alpha_0 = 0$, and $\varphi_{0k} = 0$ for $i = 1 \dots n$ and $k = 1 \dots K$, such that a corpus of size $p + 1$ requires estimation of $p \times K$ coefficients in $\mathbf{\Phi}$.

The choice of a logit link for $\mathbf{q}$, implying a linear mean for the natural exponential family parameter $\boldsymbol{\eta} = [\eta_0 \dots \eta_p]'$, is important for connecting forward and inverse regressions. Under the model in (2), with $\mathbf{\Phi} = [\boldsymbol{\varphi}_0, \cdots \boldsymbol{\varphi}_p]'$, it is easy to show that the $K$-vector $\mathbf{X}'\mathbf{\Phi}$ is a sufficient reduction. Adapting Cook's argument to multinomial $\mathrm{MN}(m, \mathbf{q})$ predictors, with $f(\mathbf{X} \mid \mathbf{V}) = \binom{m}{\mathbf{X}} \exp(\mathbf{X}'\boldsymbol{\eta} - A(\boldsymbol{\eta}))$ where $\boldsymbol{\eta} = \boldsymbol{\alpha} + \mathbf{\Phi}\mathbf{V}$ and $A(\boldsymbol{\eta}) = m \log\left(1 + \sum_{j=1}^{p} \eta_j\right)$, then

$$f(\mathbf{X} \mid y) = \binom{m}{\mathbf{X}} e^{\mathbf{X}'\boldsymbol{\alpha}} \exp((\mathbf{X}'\mathbf{\Phi})\mathbf{V} - A(\boldsymbol{\eta})) = h(\mathbf{X})g(\mathbf{X}'\mathbf{\Phi}, \mathbf{V}). \tag{3}$$

Hence, the usual Fisher sufficiency factorization implies that $[\mathbf{X} \mid \mathbf{X}'\mathbf{\Phi}, \mathbf{V}]$ is equal to $[\mathbf{X} \mid \mathbf{X}'\mathbf{\Phi}]$ and, finally, that $\mathbf{V}$ is independent of $\mathbf{X}$ given $\mathbf{X}'\mathbf{\Phi}$. As a corollary, independence between $\mathbf{X}$ and $y$ conditional on $\mathbf{V}$ implies that $\mathbf{X}'\mathbf{\Phi}$ is a sufficient reduction for $y$.

The argument in (3) is implicitly conditioning on $m$, the total phrase-count for each document, such that our full sufficient reduction result is $y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{X}'\mathbf{\Phi}, m$. Hence the corresponding forward regression model *must* condition on both $m$ and $\mathbf{X}$ for this to be a valid sufficient reduction. The most obvious way to achieve this is to regress $y$ onto phrase frequencies, $\mathbf{F}$, rather than raw counts. Fortunately, our sufficiency argument survives this transformation: if $y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{X}'\mathbf{\Phi}, m$ and $[y \mid \mathbf{X}, m] \equiv [y \mid \mathbf{F}]$, then since $\mathbf{F}'\mathbf{\Phi} = \mathbf{X}'\mathbf{\Phi}/m$ we have $y \perp\!\!\!\perp \mathbf{F} \mid \mathbf{F}'\mathbf{\Phi}$. Thus, whatever link function is used to model $y \mid \mathbf{F}$ should be translated to $y \mid \mathbf{F}'\mathbf{\Phi}$ without any loss of information; for example, if $y = \alpha + \mathbf{F}'\boldsymbol{\beta}^p + \varepsilon$, then we can write $y = \alpha + (\mathbf{F}'\mathbf{\Phi})\boldsymbol{\beta}^K + \varepsilon$ as a lower dimensional ($K \ll p$) forward regression.

## *3.4 Laplace Regularized Dimension Reduction*

Document dimension reduction is thus resolved as regression of multinomial phrase-counts onto sentiment factors through a logit link. While maximum likelihood estimation for multinomial logistic regression is well-studied (e.g., Krishnapuram et al., 2005), estimation in this inherently high-dimensional setting should benefit from prior regularization for $\mathbf{\Phi}$.

We assume improper flat priors for each phrase intercept, $\alpha_j$, and independent Laplace priors for each $\varphi_{jk}$, with coefficient-specific scale $\lambda_{jk}$, such that $\pi(\varphi_{jk}) = \lambda_{jk}/2\exp(-\lambda_{jk}|\varphi_{jk}|)$ for $j = 1 \dots p$ and $k = 1 \dots K$. This specification encourages sparsity in $\mathbf{\Phi}$ through a sharp density spike at $\varphi_{jk} = 0$, and MAP inference with fixed $\lambda_{jk}$ is equivalent to likelihood maximization under $L_1$-regularization (see, e.g., the lasso procedures of Tibshrani, 1996; Park and Casella, 2008; Hans, 2009). Hence, the MAP solution for $\mathbf{\Phi}$ will set some elements to zero, with degree of sparsity governed by both data and a penalty induced by the prior scale.

While it is common to rely on cross-validation for choosing $\boldsymbol{\lambda} = \{\lambda_{jk}\}_{j=1,k=1}^{p,K}$, we prefer to infer these parameters at almost no additional cost. To do so, we introduce conjugate gamma hyperpriors $\mathrm{Ga}(\lambda_{jk}; s, r) = r^s/\Gamma(s)\lambda_{jk}^{s-1}e^{-r\lambda_{jk}}$ for $\lambda_{jk}$; specification for $[s, r]$ is discussed in Section 4.3. Our choice to use independent priors for each coefficient, rather than a single shared $\lambda$, is a further departure from standard practice in the literature. However, we feel that it provides a more plausible representation of prior belief and, in Section 4.3, we describe how this yields a connection between joint MAP estimation and the nonconcave penalized likelihood literature (beginning from Fan and Li, 2001). It is straightforward to adapt our methods for any pooling of coefficient subsets around shared prior scale terms.

Laplace priors are a well-understood and widely used standard in the regression literature. They have the heaviest possible tails in a log-concave density and have long been recognized as providing, as a scale-mixture of normal densities, a robust alternative to the conjugate normal prior (e.g., Carlin et al., 1992). At the same time, there are many other Bayesian frameworks for regularized regression, including the spike-and-slab mixture priors of George and McCulloch (1997) and the horseshoe estimator from Carvalho et al. (2010). Our general strategy of inverse logistic regression for text-data can be combined with any coefficient prior scheme.

## 4 Estimation

Following our model specification in Section 3, the full posterior distribution of interest is

$$\mathrm{p}(\boldsymbol{\Phi}, \boldsymbol{\alpha}, \boldsymbol{\lambda} \mid \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{V}}) \propto \prod_{i=1}^{n} \left[ \binom{m_i}{\mathbf{X}_i} \prod_{j=0}^{p} q_{ij}^{x_{ij}} \right] \\ \prod_{j=1}^{p} \prod_{k=1}^{K} \frac{\lambda_{jk}}{2} \exp\left(-\lambda_{jk}|\varphi_{jk}|\right) \frac{r^s}{\Gamma(s)} \lambda_{jk}^{s-1} \exp(-r\lambda_{jk}), \tag{4}$$

where $q_{ij} = \exp(\alpha_j + \boldsymbol{\varphi}_j' \mathbf{V}_i) / \left[\sum_{h=0}^{p} \exp\left(\alpha_h + \boldsymbol{\varphi}_h' \mathbf{V}_i\right)\right]$. There is a considerable literature on posterior simulation for logistic regression, generally based around latent variable representations of the likelihood (e.g., Holmes and Held, 2006; Albert and Chib, 1993). In recent work, Gramacy and Polson (2010) describe a routine that requires $O(np)$ fewer latent variables than previous Gibbs samplers and make specific reference to the type of multinomial problems considered herein. However, even this improved algorithm requires cycling through conditional binomial likelihoods associated with each individual phrase in the corpora and requires approximations that become poor for documents with more than 20 total terms. In practice, this leads to poor mixing, slow convergence, and prohibitively expensive inference. Hence, any sampling will be limited to exploration of the local posterior around $\hat{\boldsymbol{\Phi}}$, the mode of (4). The remainder of this section focuses on finding that mode.

There are two main approaches in the literature for estimation under the lasso penalty corresponding to Laplace priors with fixed shared scale (i.e., with $\lambda_{jk} = \lambda$ for $j = 1 \ldots p$ and $k = 1 \ldots K$). Krishnapuram et al. (2005) and Madigan et al. (2005) present local convex minimization, while the predictor-corrector methods of Park and Hastie (2007) and Friedman et al. (2010) provide entire regularization paths (parameter estimates corresponding to a grid of values for shared $\lambda$). These latter procedures initialize $\lambda$ at the lowest value such that optimal $\boldsymbol{\Phi}$ is all zeros and repeatedly optimize new parameter values, starting from predicted solutions based on previous values, for decreasing $\lambda$. Under either approach, full parameter-set updates are prohibitively expensive in high-dimension due to Hessian matrix storage requirements. Hence, feasible algorithms make use of cyclic coordinate descent (CCD), wherein the joint parameter move is replaced by component-wise updates for each argument.

Section 4.1 provides background on and adapts convex minimization algorithms, of the type in Madigan et al. (2005), to account for the specific demands of text regression. We focus on these algorithms, rather than predictor-corrector methods, due to the unique accounting of algorithm cost for large-response multinomial regression: since likelihood evaluation requires normalization of $n$ probability vectors ($\mathbf{q}_i$) of length $p+1$, regardless of sparsity in $\mathbf{\Phi}$, the overly comprehensive approach of finding entire regularization paths becomes impractical. More significantly, Section 4.2 outlines a novel algorithm for estimation under the full posterior in (4), with independent unknown Laplace scale for each coefficient. In particular, we characterize a joint MAP estimation procedure for both regression coefficients and the prior scale terms, providing a very efficient method for sufficient reduction of text data.

## *4.1 Maximum Likelihood and Lasso Estimation*

Estimation procedures for logistic regression usually make use of bound-optimization (also known as majorization; Lange et al., 2000). In this framework, each update $\Theta^{t-1} \to \Theta^t$ in minimization of some function $L(\Theta)$ is the minimizing argument for a bound function $B(\Theta; \Theta^{t-1})$, where $B$ is such that setting $\Theta = \Theta^{t-1}$ will minimize the distance $B(\Theta; \Theta^{t-1}) - L(\Theta)$. This property, along with $\Theta^t$ being a minimizing argument for $B$, implies algorithm monotonicity through the inequality $L(\Theta^t) = B(\Theta^t; \Theta^{t-1}) + L(\Theta^t) - B(\Theta^t; \Theta^{t-1}) \leq B(\Theta^{t-1}; \Theta^{t-1}) - [B(\Theta^{t-1}; \Theta^{t-1}) - L(\Theta^{t-1})] = L(\Theta^{t-1})$. In CCD, easily optimized bound functions are found for each individual parameter with the remaining parameters held fixed in evaluating $L$ and $B$.

Taking the negative log likelihood and removing constant factors, our likelihood maximization problem equates with minimization of the objective function

$$l(\mathbf{\Phi}, \boldsymbol{\alpha}) = -\sum_{i=1}^{n} \left[ \mathbf{X}_i'(\boldsymbol{\alpha} + \mathbf{\Phi}\mathbf{V}_i) - m_i \log\left( \sum_{j=0}^{p} \exp(\alpha_j + \boldsymbol{\varphi_j}'\mathbf{V}_i) \right) \right], \tag{5}$$

where coordinate-wise gradient and curvature are

$$g_l(\varphi_{jk}) = \frac{\partial l}{\partial \varphi_{jk}} = -\sum_{i=1}^{n} v_{ik}(x_{ij} - m_i q_{ij}) \text{ and } h_l(\varphi_{jk}) = \frac{\partial^2 l}{\partial \varphi_{jk}^2} = \sum_{i=1}^{n} m_i v_{ik}^2 q_{ij}(1 - q_{ij}) \quad (6)$$

Here, and in the following development, we describe estimation for $\mathbf{\Phi}$; the steps for $\boldsymbol{\alpha}$ follow by recognizing that the model can be re-written by augmenting each $\mathbf{V}$ with $v_{i0} = 1$ and adding $\varphi_{j0} = \alpha_j$ as an extra element in each $\boldsymbol{\varphi}_j$ that is excluded from any regularization.

Since $l$ is strictly convex, it is possible to build a bounding function through quadratic Taylor expansion. Omitting $\boldsymbol{\alpha}$ and with $\mathbf{\Phi}$ set to the current iteration values, the CCD bound is

$$b(\varphi_{jk}^{\star}) = l(\mathbf{\Phi}) + g_l(\varphi_{jk})(\varphi_{jk}^{\star} - \varphi_{jk}) + \frac{1}{2}(\varphi_{jk}^{\star} - \varphi_{jk})^2 H_{jk}, \quad (7)$$

where $H_{jk} \geq h_l(\varphi_{jk}^{\star})$ is an upper bound on the curvature in a *trust-region* around the current $\varphi_{jk}$, defined as $\{\varphi_{jk}^{\star} \in \varphi_{jk} \pm \delta_{jk}\}$ for a specified $\delta_{jk} > 0$. This bound is optimized at $\varphi_{jk}^{\star} = \varphi_{jk} - g_l(\varphi_{jk})/H_{jk}$ and, accounting for trust region boundaries, the parameter update is $\varphi_{jk}^{\star} = \varphi_{jk} - \Delta\varphi_{jk}$, where $\Delta\varphi_{jk} = g_l(\varphi_{jk})/H_{jk}$ if $|g_l(\varphi_{jk})/H_{jk}| < \delta_{jk}$ and $\text{sgn}(g_l(\varphi_{jk}))\delta_{jk}$ otherwise.

Quadratic bounding is the approach used in both Krishnapuram et al. (2005) and Madigan et al. (2005). The former makes use of a pre-computed (but very loose) bound on $h_l$ that is valid for all $\varphi_{jk}$, while the latter advocates updating $H_{jk}$ after each iteration to obtain tighter bounding in a constrained trust region. Both approaches are straightforward to extend for the setting of grouped-response data ($m_i \geq 1$), and we provide full details in the Appendix. As could be expected, trust-region bounding leads to significantly more efficient optimization than the loose static bound (requiring an order-of-magnitude fewer iterations). Together, these steps yield a CCD algorithm for maximum likelihood estimation and, with replacements $v_{i0} = 1$, $\varphi_{j0} = \alpha_j$, provide our updates for the unpenalized intercept parameters.

The joint objective remains convex (although not strictly so) under lasso penalization with fixed $\lambda$ penalty, and component updates are the minimizing argument to $b(\varphi_{jk}^{\star}) + \lambda|\varphi_{jk}^{\star}|$. Despite lack of a derivative at $\varphi_{jk}^{\star} = 0$, parameter updates are available for $\varphi_{jk} \neq 0$ through $\Delta\varphi_{jk} = (g_l(\varphi_{jk}) + \text{sgn}(\varphi_{jk})\lambda)/H_{jk}$ if this move would not lead to $\text{sgn}(\varphi_{jk}^{\star}) \neq \text{sgn}(\varphi_{jk})$, and $\Delta\varphi_{jk} = -\varphi_{jk}$ otherwise (i.e., the update map cannot cross zero). When $\varphi_{jk} = 0$, the parameter update is found by calculating both $\Delta^{+}\varphi_{jk} = (g_l(\varphi_{jk}) + \lambda)/H_{jk}$ and $\Delta^{-}\varphi_{jk} = (g_l(\varphi_{jk}) - \lambda)/H_{jk}$, and only moving in the direction (at most one) which leads to decreased objective. Crucially, the test for decreased objective in each direction can performed on the bound function, instead of the full objective, without affecting algorithm monotonicity. This removes any need for full likelihood evaluation in our optimization. Finally, as before, the update change is bounded by $\pm\delta_{jk}$ if a trust region approach was used in calculating $H_{jk}$.

### *4.2 Joint MAP Estimation*

We now consider estimation under the full posterior in (4). In characterizing the optimization program, we note that the conditional MAP for each $\lambda_{jk}$ given $\varphi_{jk}$ is available as $\lambda(\varphi_{jk}) = s/(r + |\varphi_{jk}|)$ where $s$ and $r$ are shape and rate in the gamma prior on $\lambda_{jk}$. Hence, any joint maximizing solution $[\hat{\boldsymbol{\Phi}}, \hat{\boldsymbol{\lambda}}]$ for (4) will have $\hat{\lambda}_{jk} = \lambda(\varphi_{jk})$ for each scale parameter; otherwise, it is always possible to increase the posterior by replacing $\hat{\lambda}_{jk}$ with the conditional MAP. Fixing these prior scales and removing constant factors, the joint objective function becomes

$$L(\boldsymbol{\Phi}, \boldsymbol{\alpha}) \; = \; l(\boldsymbol{\Phi}, \boldsymbol{\alpha}) + \sum_{j=1}^{p}\sum_{k=1}^{K} \frac{s}{r + |\varphi_{jk}|}|\varphi_{jk}| \; = \; l(\boldsymbol{\Phi}, \boldsymbol{\alpha}) + c(\boldsymbol{\Phi}) \tag{8}$$

with $l(\cdot)$ from (5) proportional to the negative log likelihood and $c(\cdot)$ a penalty term based on the conditional MAP estimate for $\boldsymbol{\lambda}$.

To build a bound function for each CCD step, we can adopt the quadratic bound on $l$ from (7) and add it to the actual cost function, such that our component-wise objectives are

$$B(\varphi^{\star}_{jk}) = b(\varphi^{\star}_{jk}) + \frac{s}{r + |\varphi^{\star}_{jk}|}|\varphi^{\star}_{jk}|. \tag{9}$$

An exact minimum for (9) is obtained based upon roots to the cubic equation $B'(\varphi^{\star}_{jk}) = 0 = g(\varphi_{jk}) + (\varphi^{\star}_{jk} - \varphi_{jk})H_{jk} + \mathrm{sgn}(\varphi^{\star}_{jk}) rs/\left(r + |\varphi^{\star}_{jk}|\right)^2$. As was the case for fixed $\lambda$, solutions are not allowed to cross zero, where this derivative is undefined, and the minimization when current

$\varphi_{jk} = 0$ involves checking the solution for both positive and negative $\text{sgn}(\varphi_{jk})$ directions and moving only if the bound function is decreased. Full details are in Appendix A.2.

Although slightly more complicated than the updates for fixed $\lambda$, our joint coefficient-scale optimization steps remain exact and do not lead to any noticeable increase in computation time. In particular, we are guaranteed a single minimum solution for $B$ for all values of current $\mathbf{\Phi}$ and it is inexpensive to obtain and characterize roots for the cubic $B'(\varphi^\star_{jk}) = 0$. Taking data from Section 2.1, Figure 1 illustrates objective and bound functions around the converged solution for three phrase loadings from regression of congressional speeches onto Republican vote-share. With $\delta = 0.05$, $B$ provides tight bounding throughout this neighborhood. Behavior around the origin is most interesting: the solution for *education cut*, a low-loading democratic term, is a local minimum at $B'(\varphi^\star_{jk}) = 0$ just left of the undefined derivative at zero, while the excluded term *deep sea coral* falls in the sharp point of the objective function at zero. The neighborhood around *death tax*, a high-loading republican term, is everywhere smooth.

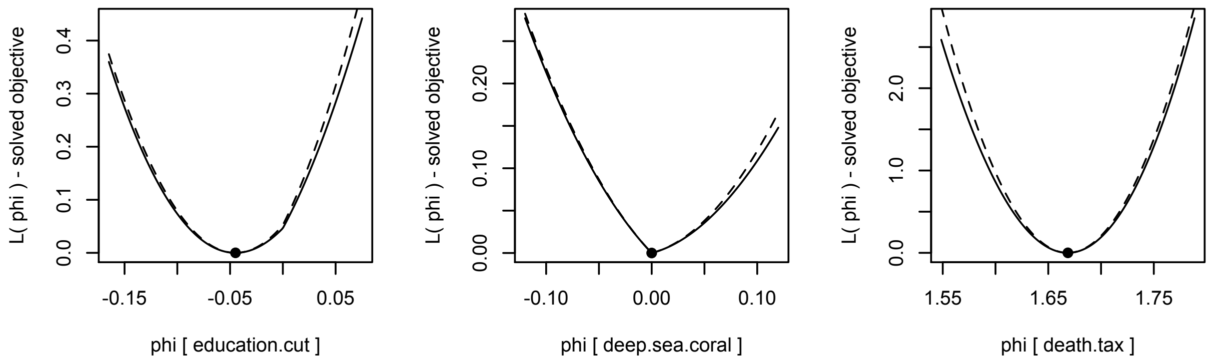


Figure 1: Political speech data. Objective functions at convergence for loadings of three phrases onto republican vote-share. In each, the solid line is full objective $L$ and the dashed line is our component-wise bound $B$ with $\delta = 0.05$. Both are shown as a function of $\varphi^\star_j$ and in terms of difference over the solved objective, $L(\mathbf{\Phi})$. Converged solutions are marked at $[\varphi_j, 0]$.

Note that the objective in (8) is no longer uniformly convex; in fact, $c(\mathbf{\Phi})$ is everywhere concave. However, from (4), the full minimization objective is $l(\mathbf{\Phi}) - \sum_{j=1}^p \sum_{k=1}^K s \log(\lambda_{jk}) + \sum_{j=1}^p \sum_{k=1}^K (r + |\varphi_{jk}|)\lambda_{jk}$. The first two terms are clearly convex in $\mathbf{\Phi}$ and $\boldsymbol{\lambda}$, respectively, while the third term is jointly convex in both $\mathbf{\Phi}$ and $\boldsymbol{\lambda}$. Hence, in contrast with the corresponding setting in linear regression (e.g., Park and Casella, 2008), our joint MAP estimation problem is unimodal. Any minor solutions will occur when the objective in (8) is concave on either side of $\varphi_{jk} = 0$; from likelihood and penalty curvature, this is only possible when

$$\sum_{i=1}^{n} m_i v_{ik}^2 q_{ij}(1 - q_{ij}) < \frac{2s}{r^2} . \tag{10}$$

To guarantee global convergence, simply assess (10) and, in the rare case that it holds, evaluate $B$ over a grid of potential $\varphi^{\star}_{jk}$ to check for an alternative mode.

There are a variety of details which will be useful for readers who wish to implement this type of algorithm for sentiment analysis. Whereas previous authors exploit sparsity in $\boldsymbol{\mathcal{V}}$, we need to take advantage of a sparse representation for the count matrix, $\boldsymbol{\mathcal{X}}$, and evaluate functionals over only non-zero elements whenever possible. As mentioned above, almost all likelihood evaluations are replaced by a bound-function calculation. We do not store multinomial probability vectors but rather the easily updated $\eta_{jk}$ values and each document's probability denominator, $\exp\left(\sum_{h=0}^{p}\left[\alpha_h + \boldsymbol{\varphi}_h'\mathbf{V}_i\right]\right)$. Furthermore, we adopt an *active set* updating strategy: elements of $\mathbf{\Phi}$ currently set at zero (which tend to stay there) are updated only $10\%$ of the time, except for full-sweep updates at initialization and at convergence. PLS coefficients can provide a basis for initialization. Finally, the algorithm terminates when the objective decrease after a full-sweep update is less than a pre-determined threshold ($10^{-4}$ throughout this article).

## *4.3 Prior Specification and Penalty Behavior*

The only prior parameters to be specified are the shape, $s$, and rate, $r$, for our gamma priors on each $\lambda_{jk}$. There are two points of intuition for this choice: prior expectation for $\lambda_{jk}$ is $s/r$, and the conditional posterior is $\mathrm{p}(\lambda_{jk} \mid \varphi_{jk}) = \mathrm{Ga}(\lambda_{jk}; s, r + |\varphi_{jk}|)$, such that $s$ provides the strength of prior information. We specify $r = 1/5, s = 1$ for all examples in Section 5, and find that this yields considerable posterior flexibility while encouraging sparsity. However, we have also found results to be very robust to orders of magnitude change in this specification; we encourage interested readers to experiment by changing the default prior parameterization in `textir`'s implementation of the examples in Section 5.

Before moving to examples, we note an important alternative perspective on the model and its estimation. Instead of viewing this as a subjective Bayesian framework, we can consider the estimation as nonconcave penalized likelihood maximization. In particular, our collapsing each $\lambda_{jk}$ into its conditional mode yields the convex parameter penalty, $-c(\varphi) = -s|\varphi|/(r + |\varphi|)$. To our knowledge, this penalty has not previously been proposed in the literature. Behavior is governed by the competing roles of $s$ and $r$: when $r$ is very small, the ratio becomes $|\varphi|/|\varphi| = 1$ and we approach an MLE solution; when $s$ is large, estimation approaches that from a lasso procedure with $\lambda$ fixed at $s/r$. Like for the lasso, our penalty is singular at the origin and smooth away from zero. However, unlike the lasso, the penalized likelihood objective can be convex near the origin whenever (10) holds, leading to possible discontinuities in the associated coefficient solution paths; Figure 2 illustrates estimated coefficients under changes in data and parameterization. For this reason, general performance of the penalty implied by joint estimation will be more similar to that of the bridge regression proposed by Frank and Friedman (1993), which is based on $L_\alpha$ regularization for $\alpha < 1$.

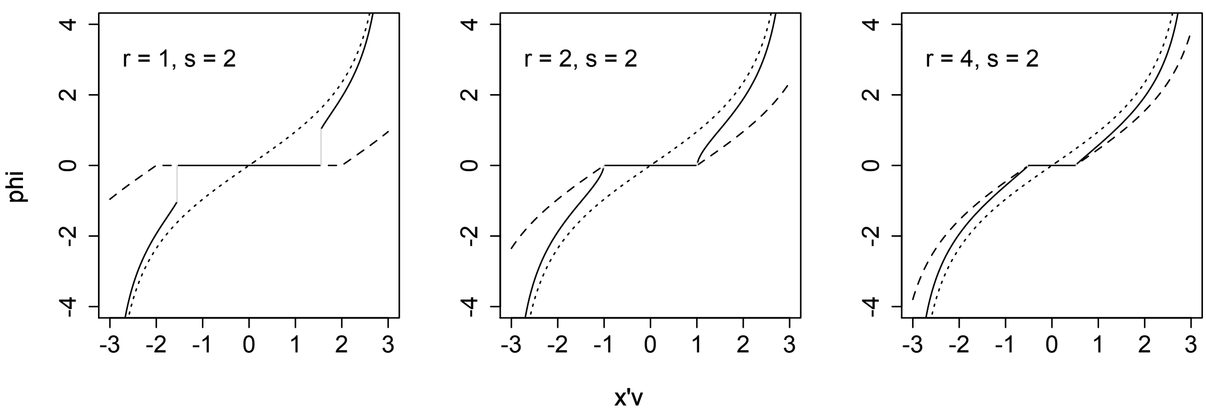


Figure 2: Thresholding functions. The plotted paths correspond to maximizing solutions for the penalized univariate logistic regression $L(\varphi) = \mathbf{x}'\mathbf{v}\varphi - \sum_i \log\left[1 + e^{\varphi v_i}\right] - z(\varphi)$, given $\mathbf{v} = -1, -0.8, \ldots 0.8, 1$. The dotted line is the MLE, with $z(\varphi) = 0$, the dashed line is lasso, with $z(\varphi) = s|\varphi|/r$, and the solid line is joint MAP penalization, with $z(\varphi) = s|\varphi|/(r + |\varphi|)$.

It is well known (Section 3.4) that fixed-$\lambda$ MAP estimation is equivalent to the lasso, another nonconcave penalization scheme, and we have already established that both fixed-$\lambda$ and joint MAP estimation procedures yield sparse parameter estimates. However, we note that $-c(\varphi)$

results in the additional property of *unbiasedness for large coefficients* listed by Fan and Li (2001) in their unified theory for nonconcave penalization. This observation connects two seemingly disparate approaches to estimation: Fan and Li remark that only a penalty associated with improper Bayesian priors could satisfy their requirement of a gradient that goes to zero as $\varphi \to \infty$. But since this clearly holds for the derivative of $-c(\varphi)$, we have that under joint prior-parameter MAP estimation it is indeed possible to recover this particular notion of unbiasedness in estimation of a Bayesian model with proper priors (more generally, see Polson and Scott, 2011, who describe a class of priors which yields a large set of popular penalty schemes). Furthermore, with this diminishing gradient, penalty smoothness away from a singularity at zero, and curvature $-c''(\varphi)$ that also goes to zero, $-c(\varphi)$ satisfies the penalty regularity conditions of Fan and Peng (2004). Hence, given the likelihood conditions of that paper and appropriately chosen $s$ and $r$, it is possible to obtain the oracle property under a diverging number of parameters for our joint MAP estimation.

## 5 Application and Results

We now apply our framework to the datasets of Section 2. The implemented software is available as the `textir` package for R, with these examples included as demos. Section 5.1 examines out-of-sample predictive performance, and is followed by individual data analyses.

### *5.1 A Comparison of Text Regression Methods*

A set of models were fit 100 times to random subsets of the data, and we report predictive performance on left-out data. At each repetition we estimate $\mathbf{\Phi}$ given 100, 1000, and 2000 randomly sampled observations from the political speech, we8there, and WSJ-IBM datasets respectively. The reduction $\mathbf{F}'\mathbf{\Phi}$ is calculated for each left-out observation, and we record correlation between true left-out sentiment and fitted values for the least-squares regression of response onto the reduction (e.g., just $\mathrm{cor}(y, \mathbf{F}'\mathbf{\Phi})$ when $\mathbf{V} = y$). We consider various sentiment factors for inverse regression: each example includes $\mathbf{V} = y$, the political text example also has $\mathbf{V}$ set to first and second common-scores, and the we8there example includes $\mathbf{V}$ set to all four aspect ratings as well as to only *Food* and *Service* scores (refer to Section 2 for details). In each case, sentiment factors are standardized to have mean of zero and variance of one.

For comparison, we also consider a set of alternative text regression techniques trained on the random subsample, and report correlation between predicted and true sentiment response in the left-out sample. These include our implementation of PLS(1) as defined in Appendix A.1, and three methods from add-on packages for R: linear lasso regression from the `glmnet` package (with cross-validated penalty term), mean-inference Laplace regularized linear regression (the *Bayesian lasso*) from the `monomvn` package (with penalty inferred under the default prior), and Blei and McAuliffe's supervised latent Dirichlet allocation from the `lda` package. PLS and the lasso procedures condition on phrase frequencies standardized over each training sample, while SLDA works directly with the document-term counts. Mean inference for the Bayesian lasso relies on 1000 MCMC samples, after a burn-in of 300; for the we8there data, with almost 3000 phrases, we use a reversible-jump model selection option for faster inference. In contrast with these other methods, the approximate expectation-maximization SLDA implementation has no default specification and requires detailed parameter choices: we fix the number of SLDA topics at 10 with prior Dirichlet precisions of $1/10$ for topic probabilities and $1/p$ for word probabilities. Moreover, additive error variance for sentiment response must be pre-specified; we set conditional variance at 25% of the marginal variance for Bush vote-share and overall restaurant ratings, and at 75% of the marginal variance for IBM return-over-market.

Results are shown in Figure 3. In all cases, multinomial inverse regression reductions provide the highest out-of-sample predictive correlation (with the exception of PLS, our MAP estimation was also the fastest procedure). In the political speech example, for both $\mathbf{V} = y$ and $\mathbf{V} = \mathbf{cs}$ (the two common-scores), inverse regression yields significantly higher mean and lower variance predictive correlation. On both measures, the common-score model performs slightly better. In the we8there example, inverse regression with $\mathbf{V} = y$ is a clear winner, indicating that overall experience cannot be represented as a function of aspect ratings ($F$ for food, $S$ for service, $A$ for atmosphere, and $V$ for value). Finally, the WSJ-IBM example's very weak sentiment signal leads to low predictive correlation. Although the advantage of inverse regression is not as dramatic here, it again leads to higher mean and lower variance correlation. Moreover, ours is also the only scheme showing strictly positive predictive correlation.

Among comparators, PLS(1) leads to the best results. This suggests that some characteristic of text regression is amenable to the PLS procedure, and provides evidence to recommend it as a very fast tool for exploratory sentiment analysis. The two lasso approaches trade advantage between each other: mean-inference Bayesian lasso wins in political speech, but the MCMC did not converge for some of our we8there repetitions. SLDA is an especially interesting case, as a state-of-the-art method that nonetheless showed poor predictive performance except in for we8there data. This could be due to limited vocabularies in the other two examples – political text pre-screened for partisanship and very short news article summaries – that do not yield the rich latent topics upon which SLDA depends for predictive power. In addition, the SLDA model has many free parameters (including the number of topics) and without taking the time for repeated cross-validations we were only able to consider a single specification.

## *5.2 Application: Ideology in Political Speeches*

To illustrate the political speech example of Section 2.1, we consider the two common-scores as sentiment factors. Results are shown in Figures 4 and 5. The former graphic contains two *Wordles*, each of which show the 75 phrases with highest absolute value coefficients for that common-score factor. Phrases are sized proportional to loading magnitude and colored by loading sign (i.e. size $\propto |\varphi_{jk}|$ and color from $\mathrm{sgn}(\varphi_{kj})$ for phrase $j$ in the $k^{th}$ common-score plot). It is easy to see which terms are most closely tied to each common-score dimension (recall that the first common-score is the dominant measure of ideology, while the second is tied to local and social issues). The latter plot shows each dimension of the reductive subspace $\mathbf{F}'\mathbf{\Phi}$ against the 2004 Bush vote-share. The first dimension exhibits clear cleavage between political parties, while the second dimension appears to cross party lines. In particular, notice the cluster of Democrats in Republican majority districts who score high on the second reductive dimension. In forward regression, least-squares fit of $y$ regressed onto $\mathbf{z}$ yields $R^2 = 0.43$.

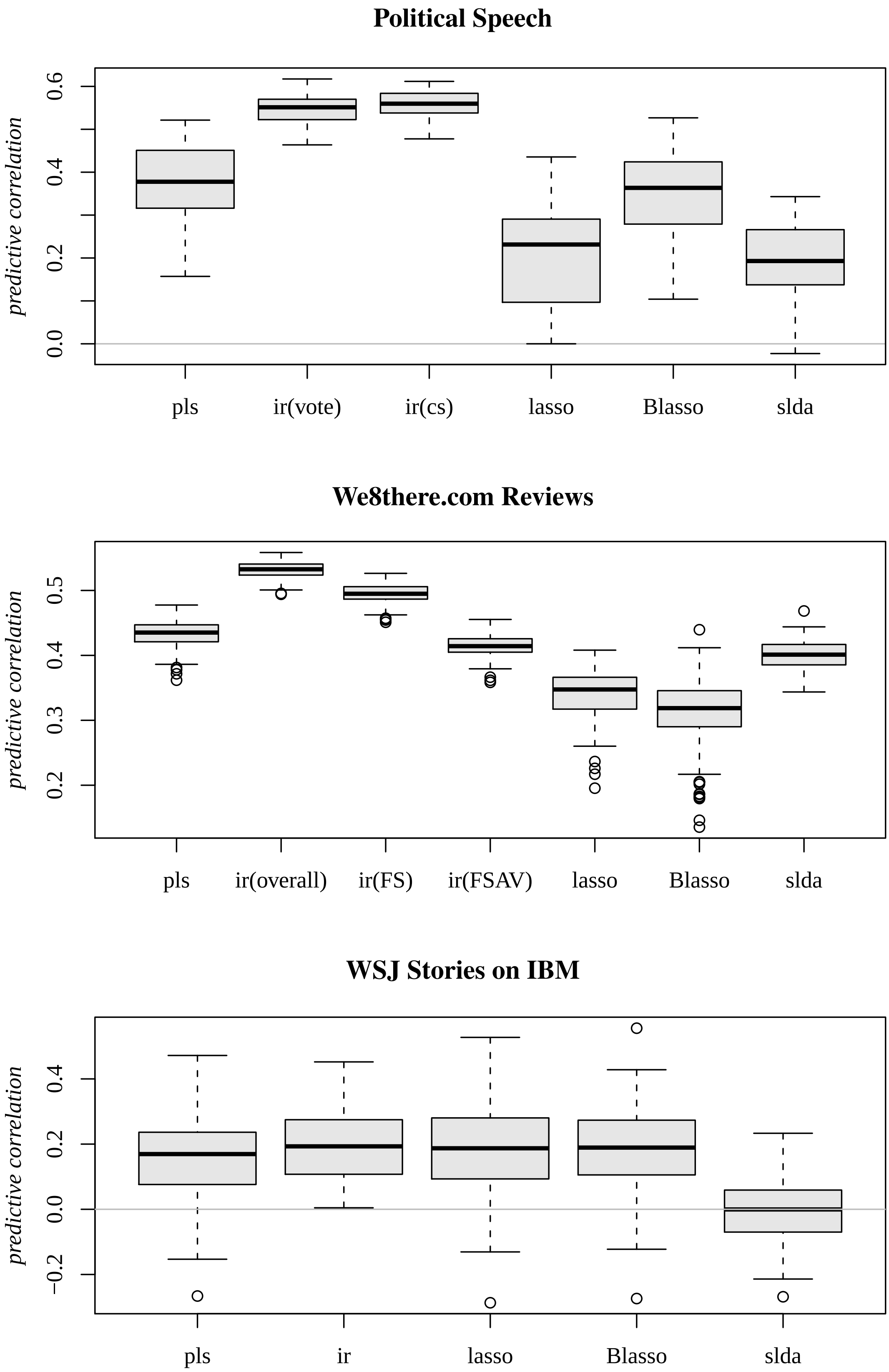


Figure 3: Predictive correlation for the example datasets. In each of 100 repetitions, models fit to subsamples of size (from top) 100, 1000, and 2000 were used to predict left-out observations.


Figure 4: High-loading phrases in the two-factor common score model. Phrase size is proportional to coefficient absolute value and color indicates the sign (red is positive, blue is negative).

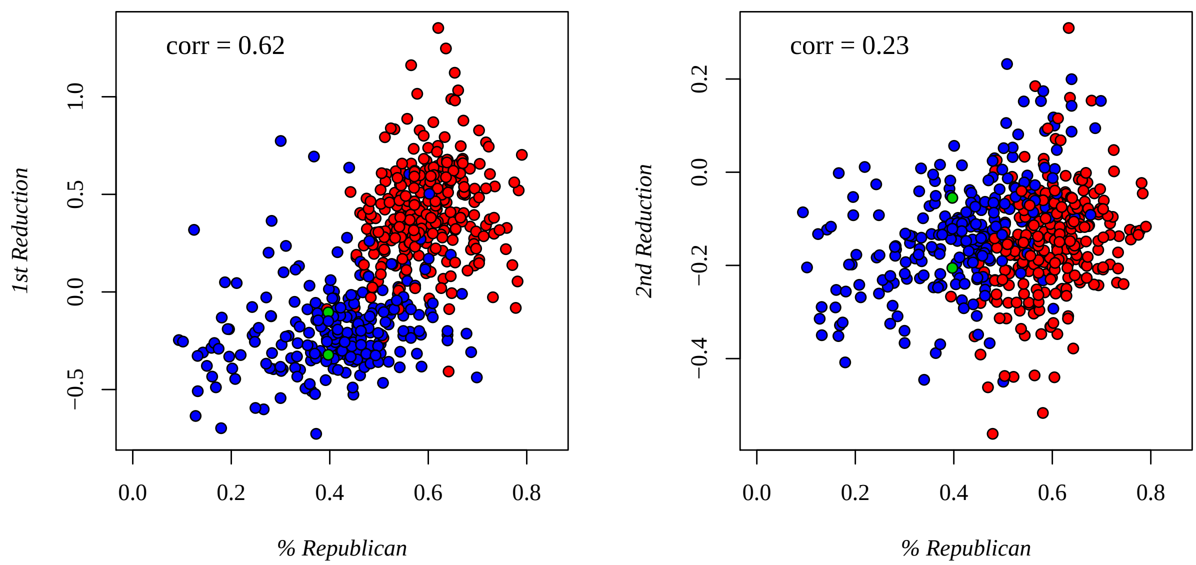


Figure 5: Republican vote-share and reductive directions for the Common Score inverse regression (i.e., each column of $\mathbf{F}'\mathbf{\Phi}$), with blue for Democrat, red for Republican, and green independents.

## 5.3 Application: On-line Restaurant Reviews

Results for the we8there example from Section 2.2 are illustrated in Figures 6 and 7. In this case, regression onto the *overall experience* rating is used to obtain a single reductive dimension. As in 5.2, Wordle phrase size is proportional to loading magnitude and indicates sign through color. Influential terms on either side of the sentiment spectrum can be easily connected with common elements of a good or bad meal. The left panel of Figure 7 has the observed rating plotted against our fitted reduction for each review (within-sample correlation is 0.7). Illustrating the utility of this document representation, the right panel shows the receiver operating characteristic (ROC) curve for discrimination of positive reviews ($> 3$ overall experience) based on this univariate reduction. The minimum misclassification rate is 13%.

## 5.4 Application: Business News and Stock Returns

As outlined in Section 2.3, our final example concerns the connection between WSJ stories about IBM and the concurrent two-day return-over-market for IBM. Factor regression coefficients are shown as a Wordle in Figure 8. Many influential words correspond to common terminology in describing market movement (e.g., *gain*, *plunge*), although others are capturing elements that seem unique to this dataset (e.g., the large loading on *firstquarter* may be related to systematic positive first quarter results from IBM, or to one-off events within our time frame). Figure 9 shows the fitted reduction plotted against two-day return-over-market, values which have a within-sample correlation of 0.43. As mentioned previously, this example involves an extremely weak sentiment response; we see in this figure that the correlation is driven by news articles occurring on days of very high or very low relative IBM returns. Finally, Figure 10 shows headlines from each of the five most positive (highest reduction) and most negative (lowest reduction) stories in the out-of-sample prediction study of Section 5.1.

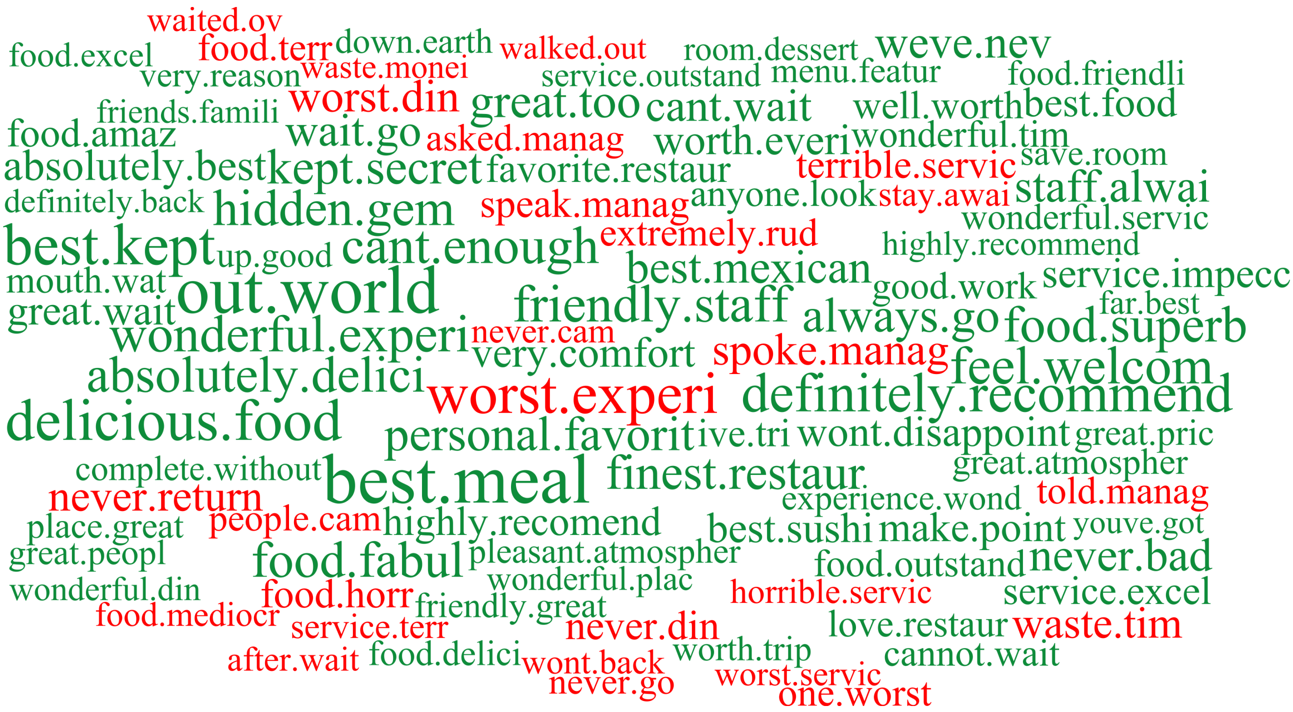


Figure 6: High-loading phrases in the restaurant review model. Phrase size is proportional to coefficient absolute value and color indicates the sign (green is positive, red is negative).

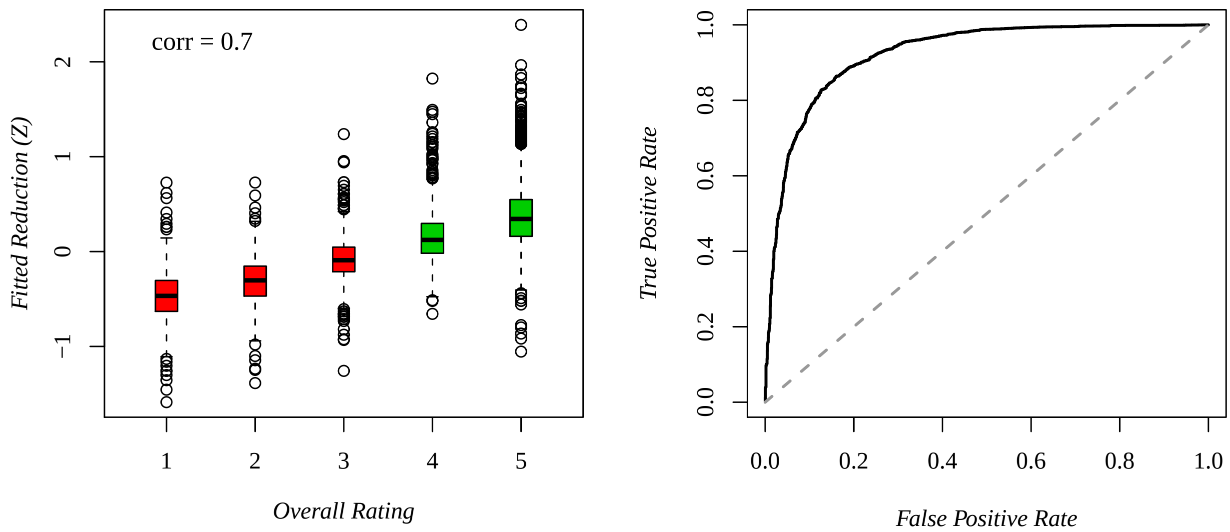


Figure 7: Fitted reduction of the we8there.com reviews. The left plot shows sample mappings $z = \mathbf{F}'\mathbf{\Phi}$ for each true overall rating, and the right plot shows the ROC curve for predictive classification of "positive" ratings ($> 3$) based on discrimination in our fitted 1D $z$-space.


Figure 8: High-loading phrases in Wall Street Journal stories on IBM. Size is proportional to coefficient absolute value and color indicates the sign (black is positive, red is negative).

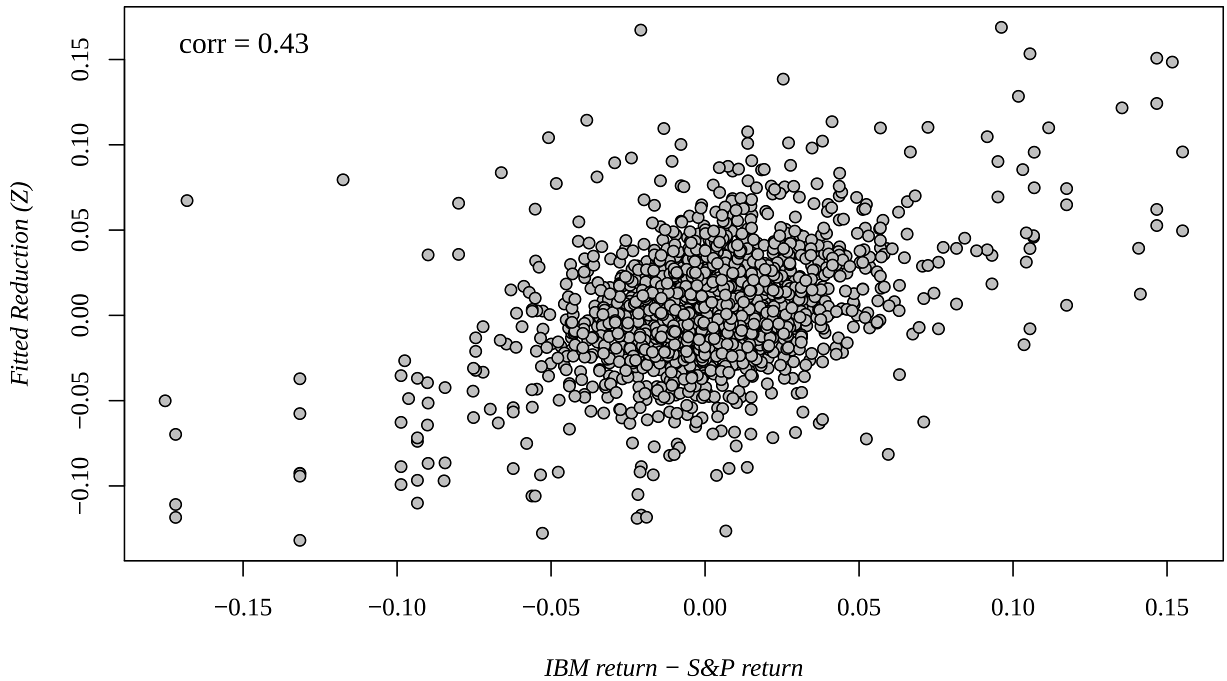


Figure 9: IBM two-day return over the S&P composite index and our fitted reduction, $z = \mathbf{F}'\mathbf{\Phi}$.

**Good News:**

CALL OPTIONS ON IBM, DELL, AND INTEL ARE LIFTED AMID AGGRESSIVE BUYING IN TECH
IBM RECEIVES SONY CONTRACT FOR NEW CHIPS
IBM NET RISES 15% AS SALES STAY STRONG
IBM'S PROFIT RISE BEATS FORECASTS
LINTAS RALLIES FOR IBM

**Bad News:**

IBM'S TROUBLES DRAG DOWN STOCKS AS INDUSTRIALS FALL 784 TO 328,436
IBM'S PROFIT WARNING PROMPTS BRISK TRADING DUE TO TECH FEARS
FUND MANAGERS HOPE IBM WONT BE DOWN LONG
IBM'S LAYOFF PLANS FOR 1,000 WILL TARGET GLOBAL-SERVICES UNIT
FIRM CUTS PRICES IN RESPONSE TO MOVES BY COMPAQ

Figure 10: Headlines from Wall Street Journal stories corresponding to the five highest (*good news*) and lowest (*bad news*) fitted reduction values during the out-of-sample prediction study of Section 5.1.

# 6 Discussion

This article describes methodology for analysis of text data into lower dimensional representations that preserve sentiment information. We show that the multinomial representation of text phrase-counts allows for use of inverse logistic regression as a general strategy whenever sentiment is regressed onto text. Furthermore, in methodology that is applicable to any multinomial regression problem, we propose efficient joint MAP estimation for both covariate coefficients and the scale of corresponding independent Laplace priors. These innovations – joint estimation and independent priors – yield a novel departure from common practice in the literature on regularized regression. We establish stability of the algorithm, as optimization of a unimodal posterior, and show that it presents an novel instance of nonconcave penalized likelihood estimation which satisfies notions of unbiasedness and consistency from that literature.

In considering the success of our framework, we emphasize that specifying an inverse distribution introduces information into the inference problem and leads to more efficient estimation. In particular, it is very beneficial that both means and covariances of the multinomial distribution are fully specified through regression coefficients. While the strategy of not modeling a corresponding forward regression falls short of full Bayesian analysis, joint inference would significantly complicate estimation and detract from our goal of providing a default method for document reduction. Hence, we are happy to take advantage of efficient parametric hierarchical Bayesian inference in obtaining sentiment-rich reductions and suggest that appropriate techniques for low-dimensional forward regression should be readily available in application.

Selection of optimal $\mathbf{V}$ remains an open question. In many settings, document visualization is a goal in its own right and $\mathbf{V}$ will include all sentiment dimensions of interest. However, when viewing document reduction as input for forward regression, we are faced with a classic variable selection problem. The cross-validation of Section 5.1 presents one approach. Cook and Li (2009) report success in model selection based on AIC and BIC for the inverse regression model. Making similar use of information criteria and staying close to our Bayesian foundation, we can consider BIC values for forward regression. It seems, due to the lack of joint inverse-forward estimation, that the correct "number of parameters" for the BIC in this setting is a naive $K + 1$ rather than the true $K(p + 1) + 1$. This leads to values which roughly mirror results from Section 5.1: with least-squares forward regression and after converting to approximate model probabilities, evidence is 65% vs 35% for $\mathbf{V} = y$ against $\mathbf{V} = \mathbf{cs}$ in the political speech example, and 100% for $\mathbf{V} = y$ in the we8there data. Our limited experience suggests that, for purely predictive purposes, the basic $\mathbf{V} = y$ specification will usually be most successful.

## Appendix

### *A.1 Slant and Partial Least Squares*

The GS slant index measures sentiment in text through a weighted sum of mean-adjusted term frequencies. Given $n \times p$ covariate matrix $\boldsymbol{\mathcal{F}}$, with elements $f_{ij}$ for the frequency of term $j$ in document $i$, and sentiment $\mathbf{y} = [y_1, \ldots, y_n]$, slant parameters are obtained through the ordinary least-squares (OLS) minimizations $[a_j, b_j] = \arg\min_{a,b} \sum_{i=1}^{n} [f_{ij} - (a + by_i)]^2$ for $j = 0 \ldots p$. The slant value for document $i$ is then $z_i^{\text{slant}} = \sum_{j=1}^{p} b_j(f_{ij} - a_j) / \sum_{j=1}^{p} b_j^2$.

Since $b_j = \text{cov}(f_j, y)/\text{var}(y)$, slant is equivalent (up to a uniform shift and scale for all index values) to a weighted sum of term frequencies loaded by their covariance with $y$. This reveals a connection to first-order partial least-squares (PLS; Wold, 1975), an iterative regression technique that builds successive factor scores for each observation as sums of predictors weighted by covariance with the response. In the first step, these predictors are simply the original covariates; each additional step derives new predictors that are orthogonal to the factor scores from the previous step. See Frank and Friedman (1993) for statistical properties of PLS and its relationship to OLS, and Hastie et al. (2009) for a common version of the algorithm.

This relationship can be used to gain insight about slant. For example, PLS is not scale-invariant and standard implementations transform covariates to have unit variance. Hence, an improved slant measure is given by $z_i^{\text{slant}} = \sum_{j=1}^{p} f_{ij}\text{cor}(f_j, y)$ and, for the congressional speech data of Section 2.1, this change increases within-sample $R^2$ from $0.37$ to $0.57$. We can also gain intuition about PLS from GS's original slant development as inverse regression for each phrase frequency $f_{ij}$ onto $y_i$ – that is, starting from the notion that language is *caused* by ideology. Given $\hat{\boldsymbol{\mathcal{F}}}$ as a standardized frequency matrix with mean-zero and variance-one columns, we propose a PLS algorithm to highlight inverse regression structure:

1. Set the initial response factor $\mathbf{v}_0 = \mathbf{y}$, and for $k = 1, \ldots, K$:
   - Loadings are $\boldsymbol{\varphi}_k = \text{cor}(\hat{\boldsymbol{\mathcal{F}}}, \mathbf{v}_{k-1}) = [\text{cor}(\hat{\mathbf{f}}_1, \mathbf{v}_{k-1}) \ldots \text{cor}(\hat{\mathbf{f}}_p, \mathbf{v}_{k-1})]'$.
   - The $k^{th}$ PLS direction is $\mathbf{z}_k = \hat{\boldsymbol{\mathcal{F}}}\boldsymbol{\varphi}_k$.
   - The new response factors are $\mathbf{v}_k = \mathbf{v}_{k-1} - [\mathbf{z}_k'\mathbf{v}_{k-1}/(\mathbf{z}_k'\mathbf{z}_k)]\mathbf{z}_k$.
2. Set $\hat{\mathbf{y}}$ as OLS fitted values for regression of $\mathbf{y}$ onto $\boldsymbol{\mathcal{Z}}$, where $\boldsymbol{\mathcal{Z}} = [\mathbf{z}_1 \cdots \mathbf{z}_K]$.

Orthogonalization of $\mathbf{v}_k$ with respect to $\mathbf{z}_k$, independent of the forward regression, is algorithmically equivalent to predictor orthogonalization in the more common PLS procedure of Hastie et al. (2009). Moreover, loading calculations replaced by $\varphi_{kj} = \arg\min_{\varphi} \sum_{i=1}^{n} [f_{ij} - (a + \varphi v_{ki})]^2$ will only scale $\mathbf{z}_k$ by the variance of $\mathbf{v}_k$ and lead to the same forward fit. Thus PLS can be viewed as stagewise inverse regression, making a connection to the material in 3.1.

## A.2 Parameter Updates

Each component update in our joint parameter-penalty estimation involves minimization of

$$B(\varphi^\star_{jk}) \propto g_l(\varphi_{jk})(\varphi^\star_{jk} - \varphi_{jk}) + \frac{1}{2}(\varphi^\star_{jk} - \varphi_{jk})^2 H_{jk} + \frac{s|\varphi^\star_{jk}|}{r + |\varphi^\star_{jk}|},$$

where $g_l(\varphi_{jk})$ is the negative log likelihood gradient and $H_{jk}$ is an upper-bound on coordinate-wise curvature $h_l(\varphi^\star_{jk})$ within $\left\{\varphi^\star_{jk} \in \varphi_{jk} \pm \delta_{jk} : \mathrm{sgn}(\varphi^\star_{jk}) = \mathrm{sgn}(\varphi_{jk})\right\}$. As mentioned in 4.1, one can either set a static bound valid for all $\delta_{jk}$ or bound in a trust region around current $\varphi_{jk}$. For grouped-response data, with $m_i \geq 1$, both options are written $H_{jk} = \sum_{i=1}^n v_{ik}^2 m_i / F_{ij}$, where the static bound has constant $F_{ij} = 2(p+1)/p$ and the trust region approach uses

$$F_{ij} = \frac{\Delta e_{ij}}{E_{ij}} + \frac{E_{ij}}{\Delta e_{ij}} + 2 \text{ where } E_{ij} = \left(\sum_{h=0}^{p} e^{\eta_{ih}}\right) - e^{\eta_{ij}} \text{ and } \Delta e_{ij} = \begin{cases} e^{\eta_{ij}-\delta_{jk}} & \text{if } E_{ij} < e^{\eta_{ij}-\delta_{jk}} \\ e^{\eta_{ij}+\delta_{jk}} & \text{if } E_{ij} > e^{\eta_{ij}+\delta_{jk}} \\ E_{ij} & \text{otherwise.} \end{cases}$$

We follow Genkin et al. (2007) in updating the trust regions at each iteration with new bound $\delta^\star_{jk} = \max\{\delta_{jk}/2, 2|\Delta\varphi_{jk}|\}$. $B(\varphi^\star_{jk})$ is then minimized by finding the roots of $B'(\varphi^\star_{jk}) = 0$ and, when necessary, comparing to the bound evaluated at zero where $B'$ is undefined. Setting $B'(\varphi^\star_{jk}) = 0$ yields the cubic function $0 = \varphi^{\star 3}_{jk} + a\varphi^{\star 2}_{jk} + b\varphi^\star_{jk} + c$, with

$$a = G + \mathrm{sgn}(\varphi_{jk})2r, \quad b = \mathrm{sgn}(\varphi_{jk})2rG + r^2, \quad \text{and} \quad c = r^2 G + \mathrm{sgn}(\varphi_{jk})\frac{sr}{H_{jk}}$$

where $G = g_l(\varphi_{jk})/H_{jk} - \varphi_{jk}$. Using standard methods for solving cubic equations, we see that for $\left\{\varphi^\star_{jk} : \mathrm{sgn}(\varphi_{jk}) = \mathrm{sgn}(\varphi^\star_{jk})\right\}$ this function will have at most one real minimizing root – that is, with $H_{jk} > 2sr/\left(r + |\varphi^\star_{jk}|\right)^3$. Hence, the update step for each coordinate is to find this root (if it exists) and compare $B(\varphi^\star_{jk})$ to $B(0)$. The minimizing value (0 or possible root $\varphi^\star_{jk}$) dictates our parameter move $\Delta\varphi_{jk}$, and this move is truncated at $\mathrm{sgn}(\Delta\varphi_{jk})\delta_{jk}$ if it exceeds the trust region boundaries. Finally, when $\varphi_{jk} = 0$, repeat this procedure for both $\mathrm{sgn}(\varphi_{jk}) = \pm 1$; at most one direction will lead to a nonzero solution.

## References


Albert, J. H. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association 88*, 669–679.

Antoniadis, A., G. Grégorie, and I. McKeague (2004). Bayesian estimation of single-index models. *Statistica Sinica 14*, 1147–1164.

Blair, E., T. Hastie, D. Paul, and R. Tibshrani (2006). Prediction by supervised principle components. *Journal of the American Statistical Association 101*, 119–137.

Blei, D. M. and J. D. McAuliffe (2010). Supervised topic models. arXiv:1003.0783v1.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research 3*, 993–1022.

Carlin, B. P., N. G. Polson, and D. S. Stoffer (1992). A Monte-Carlo approach to nonnormal and nonlinear state-space modeling. *Journal of the American Statistical Association 87*, 493–500.

Carvalho, C. M., N. G. Polson, and J. G. Scott (2010). The horseshoe estimator for sparse signals. *Biometrika 97*, 465–480.

Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statistical Science 22*, 1–26.

Cook, R. D. and L. Li (2009). Dimension reduction in regressions with exponential family predictors. *Journal of Computational and Graphical Statistics 18*, 774–791.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*, 1348–1360.

Fan, J. and H. Peng (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics 32*, 928–961.

Fortuna, B., C. Galleguillos, and N. Cristianini (2009). Detection of bias in media outlets with statistical learning. In *Text Mining: Classification, Clustering, and Applications*. CRC Press.

Frank, I. E. and J. H. Friedman (1993). A statistical view of some chemometrics regression tools. *Technometrics 35*, 109–135.

Friedman, J. H., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1).

Genkin, A., D. D. Lewis, and D. Madigan (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics 49*, 291–304.

Gentzkow, M. and J. Shapiro (2010). What drives media slant? Evidence from U.S. daily newspapers. *Econometrica 78*, 35–72.

George, E. I. and R. E. McCulloch (1997). Approaches for Bayesian variable selection. *Statistica Sinica 7*, 339–373.

Gramacy, R. B. and N. G. Polson (2010). Simulation-based regularized logistic regression. arXiv:1005.3430v1.

Hans, C. (2009). Bayesian lasso regression. *Biometrika 96*, 835–845.

Hastie, T., R. Tibshrani, and J. H. Friedman (2009). *The Elements of Statistical Learning*. Springer.

Holmes, C. C. and L. Held (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis 1*, 145–168.

Joshi, M., D. Das, K. Gimpel, and N. A. Smith (2010). Movie reviews and revenues: An experiment in text regression. In *North American Conference for Computational Linguistics*.

Jurafsky, D. and J. H. Martin (2009). *Speech and Language Processing* (2nd ed.). New Jersey: Prentice Hall.

Koppel, M. and I. Shtirmberg (2004). Good news or bad news? let the market decide. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text*.

Krishnapuram, B., L. Carin, M. A. T. Figueiredo, and A. J. Hartemink (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27*, 957–969.

Lange, K., D. R. Hunter, and I. Yang (2000). Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics 9*, 1–20.

Li, K. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association 86*, 316–327.

Madigan, D., A. Genkin, D. D. Lewis, and D. Fradkin (2005). Bayesian multinomial logistic regression for author identification. In *AIP Conference Proceedings*, Volume 803.

Mao, K., F. Liang, and S. Mukherjee (2010). Supervised dimension reduction using Bayesian mixture modeling. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Massy, W. F. (1965). Principal components regression in exploratory statistical research. *Journal of the American Statistical Association 60*, 234–256.

Mauá, D. D. and F. G. Cozman (2009). Representing and classifying user reviews. In *ENIA '09: VIII Enconro Nacional de Inteligência Artificial*, Brazil.

Pang, B. and L. Lee (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval 1-2*, 1–135.

Park, M. Y. and T. Hastie (2007). L1 regularization path algorithms for generalized linear models. *Journal of the Royal Statistical Society, Series B 69*, 659–677.

Park, T. and G. Casella (2008). The Bayesian lasso. *Journal of the American Statistical Association 103*, 681–686.

Polson, N. G. and J. G. Scott (2011). Sparse bayes estimation in non-Gaussian models via data augmentation. arXiv:1103.5407v2.

Poole, K. T. (2005). *Spatial Models of Parliamentary Voting*. New York: Cambridge University Press.

Poole, K. T. and H. Rosenthal (2007). *Ideology and Congress*. New Jersey: Transaction.

Poon, H. and P. Domingos (2009). Unsupervised semantic parsing. In *Proceedings of the Conference on EMNLP*.

Schumaker, R. P. and H. Chen (2009). Textual analysis of stock market prediction using breaking financial news: The AZFinText system. *ACM Transactions on Information Systems 27*, 1–19.

Srivastava, A. N. and M. Sahami (Eds.) (2009). *Text Mining: Classification, Clustering, and Applications*. CRC Press.

Tetlock, P. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance 62*, 1139–1168.

Tibshrani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B 58*, 267–288.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics*, Volume 7. Oxford Univeresity Press.

Wold, H. (1975). Soft modeling by latent variables: The nonlinear iterative partial least squares approach. In *Perspectives in Probability and Statistics, Papers in Honour of M.S. Bartlett*. Academic Press.