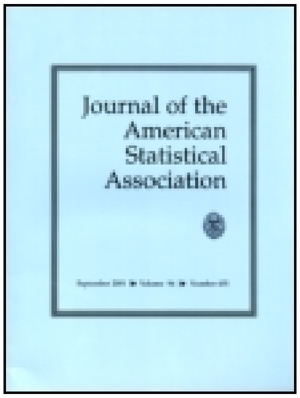

Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Multinomial Inverse Regression for Text Analysis

Matt Taddy [a]

[a] The University of Chicago Booth School of Business

Accepted author version posted online: 08 Oct 2012.

Multinomial Inverse Regression for Text Analysis

Matt Taddy
taddy@chicagobooth.edu
The University of Chicago Booth School of Business

Abstract

Text data, including speeches, stories, and other document forms, are often connected to *sentiment* variables that are of interest for research in marketing, economics, and elsewhere. It is also very high dimensional and difficult to incorporate into statistical analyses. This article introduces a straightforward framework of sentiment-sufficient dimension reduction for text data. Multinomial inverse regression is introduced as a general tool for simplifying predictor sets that can be represented as draws from a multinomial distribution, and we show that logistic regression of phrase counts onto document annotations can be used to obtain low dimension document representations that are rich in sentiment information. To facilitate this modeling, a novel estimation technique is developed for multinomial logistic regression with very high-dimension response. In particular, independent Laplace priors with unknown variance are assigned to each regression coefficient, and we detail an efficient routine for maximization of the joint posterior over coefficients and their prior scale. This 'gamma-lasso' scheme yields stable and effective estimation for general high-dimension logistic regression, and we argue that it will be superior to current methods in many settings. Guidelines for prior specification are provided, algorithm convergence is detailed, and estimator properties are outlined from the perspective of the literature on non-concave likelihood penalization. Related work on sentiment analysis from statistics, econometrics, and machine learning is surveyed and connected. Finally, the methods are applied in two detailed examples and we provide out-of-sample prediction studies to illustrate their effectiveness.

Taddy is an Associate 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. The author thanks Jesse Shapiro, Matthew Gentzkow, David Blei, Che-Lin Su, Christian Hansen, Robert Gramacy, Nicholas Polson, and anonymous reviewers for much helpful discussion.

ACCEPTED MANUSCRIPT

1 Introduction

This article investigates the relationship between text data – product reviews, political speech, financial news, or a personal blog post – and variables that are believed to influence its composition – product quality ratings, political affiliation, stock price, or mood polarity. Such language-motivating observable variables, generically termed *sentiment* in the context of this article, are often the main object of interest for text mining applications. When, as is typical, large amounts of text are available but only a small subset of documents are annotated with known sentiment, this relationship yields the powerful potential for text to act as a stand-in for related quantities of primary interest. On the other hand, language data dimension (i.e., vocabulary size) is both very large and tends to increase with the amount of observed text, making the data difficult to incorporate into statistical analyses. Our goal is to introduce a straightforward framework of sentiment-preserving dimension reduction for text data.

As detailed in Section 2.1, a common statistical treatment of text views each document as an exchangeable collection of phrase tokens. In machine learning, these tokens are usually just words (e.g., *tax*, *pizza*) obtained after stemming for related roots (e.g., *taxation*, *taxing*, and *taxes* all become *tax*), but richer tokenizations are also possible: for example, we find it useful to track common n-gram word combinations (e.g. bigrams *pay tax* or *cheese pizza* and trigrams such as *too much tax*). Under a given tokenization each document is represented as $\mathbf{x}_i = [x_{i1}, \ldots, x_{ip}]'$, a sparse vector of counts for each of p tokens in the vocabulary. These token counts, and the associated frequencies $\mathbf{f}_i = \mathbf{x}_i/m_i$ where $m_i = \sum_{j=1}^{p} x_{ij}$, are then the basic data units for statistical text analysis. In particular, the multinomial distribution for $\mathbf{x}_i$ implied by an assumption of token-exchangeability can serve as the basis for efficient dimension reduction.

Consider n documents that are each annotated with a single sentiment variable, y_i (e.g., restaurant reviews accompanied by a one to five star rating). A naive approach to text-sentiment prediction would be to fit a generic regression for $y_i|\mathbf{x}_i$. However, given the very high dimension of

ACCEPTED MANUSCRIPT

text-counts (with p in the thousands or tens of thousands), one cannot efficiently estimate this conditional distribution without also taking steps to simplify $\mathbf{x}_i$. We propose an inverse regression (IR) approach, wherein the *inverse conditional distribution* for text given sentiment is used to obtain low dimensional document scores that preserve information relevant to y_i.

As an introductory example, consider the text-sentiment contingency table built by collapsing token counts as $\mathbf{x}_y = \sum_{i:y_i=y} \mathbf{x}_i$ for each $y \in \mathcal{Y}$, the support of an ordered discrete sentiment variable. A basic multinomial inverse regression (MNIR) model is then

$$\mathbf{x}_y \sim \mathrm{MN}(\mathbf{q}_y, m_y) \;\text{ with }\; q_{yj} = \frac{\exp[\alpha_j + y\varphi_j]}{\sum_{l=1}^{p} \exp[\alpha_l + y\varphi_l]}, \;\text{ for }\; j = 1, \ldots, p, \;\; y \in \mathcal{Y} \tag{1}$$

where each MN is a p-dimensional multinomial distribution with size $m_y = \sum_{i:y_i=y} m_i$ and probabilities $\mathbf{q}_y = [q_{y1}, \ldots, q_{yp}]'$ that are a linear function of y through a logistic link. Under conditions detailed in Section 3, the *sufficient reduction* (SR) score for $\mathbf{f}_i = \mathbf{x}_i/m_i$ is then

$$z_i = \boldsymbol{\varphi}'\mathbf{f}_i \;\Rightarrow\; y_i \perp\!\!\!\perp \mathbf{x}_i, m_i \mid z_i. \tag{2}$$

Hence, given this SR projection, full $\mathbf{x}_i$ is ignored and modeling the text-sentiment relationship becomes a univariate regression problem. This article's examples include linear, $\mathbb{E}[y_i] = \beta_0 + \beta_1 z_i$, quadratic, $\mathbb{E}[y_i] = \beta_0 + \beta_1 z_i + \beta_2 z_i^2$, and logistic, $\mathrm{p}(y_i < a) = (1 + \exp[\beta_0 + \beta_1 z_i])^{-1}$, forms for this *forward regression*, and SR scores should be straightforward to incorporate into alternative regression models or structural equation systems. The procedure rests upon assumptions that allow for summary tables wherein the text-sentiment relationship of interest can be modeled as a logistic multinomial, but when such assumptions are plausible, as we find common in text analysis, they introduce information that should yield significant efficiency gains.

In estimating models of the type in (1), which involve many thousands of parameters, we propose use of fat-tailed and sparsity-inducing independent Laplace priors for each coefficient φ_j. To account for uncertainty about the appropriate level of variable-specific regularization, each Laplace rate parameter λ_j is left unknown with a gamma hyperprior. Thus, for example,

$$\pi(\varphi_j, \lambda_j) = \frac{\lambda_j}{2} e^{-\lambda_j|\varphi_j|} \frac{r^s}{\Gamma(s)} \lambda_j^{s-1} e^{-r\lambda_j}, \quad s, r, \lambda_j > 0, \tag{3}$$

ACCEPTED MANUSCRIPT

independent for each j under a $\text{Ga}(s, r)$ hyperprior specification. This departure from the usual shared-λ model is motivated in Section 3.3.

Fitting MNIR models is tough for reasons beyond the usual difficulties of high dimension regression – simply evaluating the large-response likelihood is expensive due to the normalization in calculating each $\mathbf{q}_i$. As surveyed in Section 4, available cross-validation (e.g., via solution paths) and fully Bayesian (i.e., through Monte-Carlo marginalization) methods for estimating φ_j under unknown λ_j are prohibitively expensive. A novel algorithm is proposed for finding the joint posterior maximum (MAP) estimate of both coefficients and their prior scale. The problem is reduced to log likelihood maximization for $\boldsymbol{\varphi}$ with a non-concave penalty, and it can be solved relatively quickly through coordinate descent. For example, given the prior in (3), the log likelihood implied by (1) is maximized subject to (i.e., minus) cost constraints

$$c(\varphi_j) = s \log(1 + |\varphi_j|/r) \tag{4}$$

for each coefficient. This provides a powerful new estimation framework, which we term the *gamma-lasso*. The approach is very computationally efficient, yielding robust SR scores in less than a second for documents with thousands of unique tokens. Indeed, although a full comparison is beyond the scope of this paper, we find that the proposed algorithm can also be far superior to current techniques for high-dimensional logistic regression in the more common large-predictor (rather than large-response) setting.

This article thus includes two main methodological contributions. First, Section 3 introduces multinomial inverse regression as an IR procedure for predictor sets that can be represented as draws from a multinomial, and details its application to text-sentiment analysis. This includes full model specification and general sufficiency results, guidelines on how text data should be handled to satisfy the MNIR model assumptions, and our independent gamma-Laplace prior specification. Second, Section 4 develops a novel approach to estimation in very high dimensional logistic regression. This includes details of coordinate descent for joint MAP estimation of coefficients and their unknown variance, conditions for global convergence, and an outline of estimator properties from the perspective of the literature on non-concave likelihood penalization. As background, Section 2 briefly surveys the literature on text mining and sentiment analysis, and on dimension reduction and inverse regression.

The following section describes language pre-processing and introduces two datasets that are used throughout to motivate and illustrate our methods. Performance comparison and detailed results for these examples are then presented in Section 5. Both example datasets, along with all implemented methodology, are available in the textir package for R.

1.1 Data processing and examples

Text is usually initially cleaned according to some standard information retrieval criteria, and we refer the reader to Jurafsky and Martin (2009) for an overview. In this article, we simply remove a limited set of stop words (e.g., *and* or *but*) and punctuation, convert to lowercase, and strip suffixes from roots according to the Porter stemmer (Porter, 1980). The main data preparation step is then to parse clean text into informative language tokens; as mentioned in the introduction, counts for these tokens are the starting point for statistical analysis. Most commonly (see, e.g., Srivastava and Sahami, 2009) the tokens are just words, such that each document is treated as a vector of word-counts. This is referred to as the *bag-of-words* representation, since these counts are summary statistics for language generated by exchangeable draws from a multinomial ‘bag’ of word options.

Despite its apparent limitations, the token-count framework can be made quite flexible through more sophisticated tokenization. For example, in the *N-gram* language model words are drawn from a Markov chain of order N (see, e.g., Jurafsky and Martin, 2009). A document is then summarized by its length-N word sequences, or N-gram tokens, as these are sufficient for the underlying Markov transition probabilities. Our general practice is to count common unigram, bigram, and trigram tokens (i.e., words and 2-3 word phrases). Another powerful technique is to use domain-specific knowledge to parse for phrases that are meaningful in the context of a specific field. Talley and O'Kane (2011) present one such approach for tokenization of legal agreements; for example, they use any conjugation of the word *act* in proximity of *God* to identify a common *Act of God* class of carve-out provisions. Finally, work such as that of Poon and Domingos (2009) seeks to parse language according to semantic equivalence.

Thus while we focus on token-count data, different language models are able to influence analysis through tokenization rules. And although separation of parsing from statistical modeling limits our ability to quantify uncertainty, it has the appealing effect of allowing text data from various sources and formats to all be analyzed within a multinomial likelihood framework.

1.1.1 Ideology in political speeches

This example originally appears in Gentzkow and Shapiro (GS; 2010) and considers text of the 109th (2005-2006) Congressional Record. For each of the 529 members of the United States House and Senate, GS record usage of phrases in a list of 1000 bigrams and trigrams. Each document corresponds to a single person. The sentiment of interest is political partisanship, where party affiliation (Republican, Democrat, or Independent) provides a simple indicator and a higher-fidelity measure is calculated as 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. Note that token vocabulary in this example is influenced by sentiment: GS built contingency tables for bigram and trigram usage by party, and kept the top 1000 'most partisan' phrases according to ranking of their Pearson χ^2-test statistic.

Define phrase frequency *lift* for a given group as $\bar{f}_{j\mathrm{G}}/\bar{f}_j$, where $\bar{f}_{j\mathrm{G}}$ is mean frequency for phrase j in group G and $\bar{f}_j = \sum_{i=1}^{n} f_{ij}/n$ is the average across all documents. The following tables show top-five lift phrases used at least once by each party.

ACCEPTED MANUSCRIPT

Democratic Frequency Lift	
congressional.hispanic.caucu	2.163
medicaid.cut	2.154
clean.drinking.water	2.154
earth.day	2.152
tax.cut.benefit	2.149

Republican Frequency Lift	
ayman.al.zawahiri	1.850
america.blood.cent	1.849
million.budget.request	1.847
million.illegal.alien	1.846
temporary.worker.program	1.845

1.1.2 On-line restaurant reviews

This dataset, which originally appears in the topic analysis of Mauá and Cozman (2009), contains 6260 user-submitted restaurant reviews (90 word average) from `www.we8there.com`. The reviews are accompanied by a five-star rating on four specific aspects of quality – *food*, *service*, *value*, and *atmosphere* – as well as the *overall experience*. After tokenizing the text into bigrams (based on a belief that modifiers such as *very* or *small* would be useful here), we discard phrases that appear in less than ten reviews and documents which do not use any of the remaining phrases. This leaves a dataset of 6147 review counts for a token vocabulary of 2978 bigrams. Top-five lift phrases that occur at least once in both positive (*overall experience* > 3) and negative (*overall experience* < 3) reviews are below.

Negative Frequency Lift	
food poison	5.402
food terribl	5.354
one worst	5.339
spoke manag	5.318
after left	5.285

Positive Frequency Lift	
worth trip	1.393
everi week	1.390
melt mouth	1.389
alway go	1.389
onc week	1.389

2 Background

This section briefly reviews the relevant literatures on sentiment analysis and inverse regression. Additional background is in the appendices and material specific to estimation is in Section 4.

2.1 Analysis of sentiment in text

As already outlined, we use *sentiment* to refer to any variables related to document composition. Although broader than its common ‘opinion polarity’ usage, this definition as ‘sensible quality’ fits our need to refer to the variety of quantities that may be correlated with text.

Much of existing work on sentiment analysis uses word frequencies as predictors in generic regression and classification algorithms, including support vector machines, principle components (PC) regression, neural networks, and penalized least-squares. Examples from this machine learning literature can be found in the survey by Pang and Lee (2008) and in the collection from Srivastava and Sahami (2009). In the social sciences, research on ideology in political text includes both generic classifiers (e.g., Yu et al., 2008) and analysis of contingency tables for individual terms (e.g., Laver et al., 2003) (machine learning researchers, such as Thomas et al., 2006, have also made contributions in this area). In economics, particularly finance, it is more common to rely on weighted counts for pre-defined lists of terms with positive or negative *tone*; examples of this approach include Tetlock (2007) and Loughran and McDonald (2011) (again, machine learners such as Bollen et al., 2011, have also studied prediction for finance).

These approaches all have drawbacks: generic regression does nothing to leverage the particulars of text data, independent analysis of many contingency tables leads to multiple-testing issues, and pre-defined word lists are subjective and unreliable. A more promising strategy is to use text-specific dimension reduction based upon the multinomial implied by exchangeability of token-counts. For example, a *topic model* treats documents as drawn from a multinomial distribution with probabilities arising as a weighted combination of 'topic' factors. Thus $\mathbf{x}_i \sim \mathrm{MN}(\omega_{i1}\boldsymbol{\theta}_1 + \ldots + \omega_{iK}\boldsymbol{\theta}_K, m_i)$, where topics $\boldsymbol{\theta}_k = [\theta_{k1} \cdots \theta_{kp}]'$ and weights $\boldsymbol{\omega}_i$ are probability vectors. This framework, also known as *latent Dirichlet allocation* (LDA), has been widely used in text analysis since its introduction by Blei et al. (2003).

The low dimensional topic-weight representation (i.e., $\boldsymbol{\omega}_i$) serves as a basis for sentiment analysis in the original Blei et al. article, and has been used in this way by many since. The approach is especially popular in political science, where work such as that of Grimmer (2010) and Quinn et al. (2010) investigates political interpretation of latent topics (these authors restrict $\omega_{ik} \in \{0, 1\}$ such that each document is drawn from a single topic). Recently, Blei and McAuliffe (2007) have introduced supervised LDA (sLDA) for joint modeling of text and sentiment. In particular, they augment topic model with a forward regression $y_i = f(\omega_i)$, such that token counts and sentiment are connected through shared topic-weight factors.

Finally, our investigation was originally motivated by a desire to build a model-based version of the specific *slant* indices proposed by Gentzkow and Shapiro (2010), which are part of a general political science literature on quantifying partisanship through weighted-term indices (e.g., Laver et al., 2003). Appendix A.1 shows that the GS indices can be written as summation of phrase frequencies loaded by their correlation with measured partisanship (e.g., vote-share), such that slant is equivalent to first-order partial least-squares (PLS; Wold, 1975).

2.2 Inverse regression and sufficient reduction

This article is based on a notion that, given the high dimension of text data, it is not possible to efficiently estimate conditional response $y|\mathbf{x}$ without finding a way to simplify $\mathbf{x}$. The same idea motivates many of the techniques surveyed above, including LDA and sLDA, PLS/slant, and PC regression. A framework to unify techniques for dimension reduction in regression can be found in Cook's 2007 overview of *inverse regression*, wherein inference about the multivariate conditional distribution $\mathbf{x}|y$ is used to build low dimension summaries for $\mathbf{x}$.

Suppose that $\mathbf{v}_i$ is a K-vector of *response factors* through which $\mathbf{x}_i$ depends on y_i (i.e., $\mathbf{v}_i$ is a possibly random function of y_i). Then Cook's linear IR formulation has $\mathbf{x}_i = \boldsymbol{\Phi}\mathbf{v}_i + \boldsymbol{\epsilon}_i$, where $\boldsymbol{\Phi} = [\boldsymbol{\varphi}_1 \cdots \boldsymbol{\varphi}_K]$ is a $p \times K$ matrix of inverse regression coefficients and $\boldsymbol{\epsilon}_i$ is p-vector of error terms. Under certain conditions on $\mathrm{var}(\boldsymbol{\epsilon}_i)$, detailed by Cook, the projection $\mathbf{z}_i = \boldsymbol{\Phi}'\mathbf{x}_i$ provides a *sufficient reduction* (SR) such that y_i is independent of $\mathbf{x}_i$ given $\mathbf{z}_i$. As this implies $\mathrm{p}(\mathbf{x}_i|\boldsymbol{\Phi}'\mathbf{x}_i, y_i) = \mathrm{p}(\mathbf{x}_i|\boldsymbol{\Phi}'\mathbf{x}_i)$, SR corresponds to the classical definition of sufficiency for 'data' $\mathbf{x}_i$ and 'parameter' y_i, but is conditional on unknown $\boldsymbol{\Phi}$ that must be estimated in practice. When such estimation is feasible, the reduction of dimension from p to K should make these SR *projections* easier to work with than the original predictors.

Many approaches to dimension reduction can be understood in context of this linear IR model: PC directions arise as SR projections for the maximum likelihood solution when $\mathbf{v}_i$ is unspecified (see, e.g., Cook, 2007) and, following our discussion in A.1, the first PLS direction is the SR projection for least-squares fit when $\mathbf{v}_i = y_i$. A closely related framework is that of factor analysis, wherein one seeks to estimate $\mathbf{v}_i$ directly rather than project $\mathbf{x}_i$ into its lower dimensional space. By augmenting estimation with a forward model for $y_i|\mathbf{v}_i$ researchers are able to build *supervised factor models*; see, e.g., West (2003).

The innovation of inverse regression, from Cook's 2007 paper and in earlier work including Li (1991) and Bura and Cook (2001), is to investigate the SR projections that result from explicit specification for $\mathbf{v}_i$ as a function of y_i. Cook's *principle fitted components* are derived for a variety of functional expansions of y_i, Li et al. (2007) interprets PLS within an IR framework, and the *sliced inverse regression* of Li (1991) defines $\mathbf{v}_i$ as a step-function expansion of y_i. Since in each case the $\mathbf{v}_i$ are conditioned upon, these IR models are more restrictive than the random joint forward-inverse specification of supervised factor models. But if the IR model assumptions are satisfied then its parsimony should lead to more efficient inference.

Instead of a linear equation, dimension reduction for text data is based on multinomial models. Following the topic model factor specification, LDA is akin to PC analysis for multinomials and sLDA is the corresponding supervised factor model. However, existing work on non-Gaussian inverse regression relies on conditional independence; for example, Cook and Li (2009) use single-parameter exponential families to model each $x_{ij}|\mathbf{v}_i$. To our knowledge, no-one has investigated SR projections based on the multinomial predictor distributions that arise naturally for text data. Hence, we seek to build a multinomial inverse regression framework.

3 Modeling

The subject-specific multinomial inverse regression model has, for $i = 1, \ldots, n$:

$$\mathbf{x}_i \sim \mathrm{MN}(\mathbf{q}_i, m_i) \text{ with } q_{ij} = \frac{e^{\eta_{ij}}}{\sum_{l=1}^{p} e^{\eta_{il}}}, \quad j = 1, \ldots, p, \quad \text{where } \eta_{ij} = \alpha_j + u_{ij} + \mathbf{v}_i' \boldsymbol{\varphi}_j. \tag{5}$$

This generalizes (1) with the introduction of K-dimensional response factors $\mathbf{v}_i$ and subject effects $\mathbf{u}_i = [u_{i1} \cdots u_{ip}]'$. Section 3.1 derives sufficient reduction results for projections $\mathbf{z}_i m_i = \boldsymbol{\Phi}' \mathbf{x}_i$, where $\boldsymbol{\Phi}' = [\boldsymbol{\varphi}_1, \cdots \boldsymbol{\varphi}_p]$. Section 3.2 then describes application of these results in text analysis and outlines situations where (5) can be replaced with a collapsed model as in (1). Finally, 3.3 presents prior specification for these very high dimensional regressions.

3.1 Sufficient reduction in multinomial inverse regression

This section establishes classical sufficiency-for-y (conditional on IR parameters) for projections derived from the model in (5). The main result follows, due to use of a logit link on $\boldsymbol{\eta}_i = [\eta_{i1} \cdots \eta_{ip}]'$, from factorization of the multinomial's natural exponential family parametrization.

Proposition 3.1. *Under the model in (5), conditional on m_i and $\mathbf{u}_i$*
$y_i \perp\!\!\!\perp \mathbf{x}_i \mid \mathbf{v}_i \Rightarrow y_i \perp\!\!\!\perp \mathbf{x}_i \mid \boldsymbol{\Phi}' \mathbf{x}_i$.

Proof. Setting $\alpha_{ij} = \alpha_j + u_{ij}$ and suppressing i, the likelihood is $\binom{m}{\mathbf{x}} \exp\left[\mathbf{x}'\boldsymbol{\eta} - A(\boldsymbol{\eta})\right] = \binom{m}{\mathbf{x}} e^{\mathbf{x}'\boldsymbol{\alpha}} \exp\left[(\mathbf{x}'\boldsymbol{\Phi})\mathbf{v} - A(\boldsymbol{\eta})\right] =$ $h(\mathbf{x}) g(\boldsymbol{\Phi}'\mathbf{x}, \mathbf{v})$, where $A(\boldsymbol{\eta}) = m \log\left[\sum_{j=1}^{p} e^{\eta_j}\right]$. Hence, the usual sufficiency factorization (e.g., Schervish, 1995, 2.21) implies $\mathrm{p}(\mathbf{x}|\boldsymbol{\Phi}'\mathbf{x}, \mathbf{v}) = \mathrm{p}(\mathbf{x}|\boldsymbol{\Phi}'\mathbf{x})$, and $\mathbf{v}$ is independent of $\mathbf{x}$ given $\boldsymbol{\Phi}'\mathbf{x}$. Finally, $\mathrm{p}(y|\mathbf{x}, \boldsymbol{\Phi}'\mathbf{x}) = \int_{\mathbf{v}} \mathrm{p}(y|\mathbf{v}) d\mathrm{P}(\mathbf{v}|\boldsymbol{\Phi}'\mathbf{x}) = \mathrm{p}(y|\boldsymbol{\Phi}'\mathbf{x})$.

Second, it is standard in text analysis to control for document size by regressing y_i onto frequencies rather than counts. Fortunately, our sufficient reductions survive this transformation.

Proposition 3.2. *If $y_i \perp\!\!\!\perp \mathbf{x}_i \mid \boldsymbol{\Phi}'\mathbf{x}_i, m_i$ and $\mathrm{p}(y \mid \mathbf{x}_i) = \mathrm{p}(y_i \mid \mathbf{f}_i)$, then $y_i \perp\!\!\!\perp \mathbf{x}_i \mid \mathbf{z}_i = \boldsymbol{\Phi}'\mathbf{f}_i$.*

Proof. We have that each of $\mathbf{f}$ and $[\boldsymbol{\Phi}'\mathbf{f}, m]$ are sufficient for y in $\mathrm{p}(\mathbf{x}|y) = \mathrm{MN}(\mathbf{q}, m)\mathrm{p}(m|y)$. Under conditions of Lehmann and Sheffé (1950, 6.3), there exists a minimal sufficient statistic $T(\mathbf{x})$ and functions g and $\tilde{g}$ such that $g(\mathbf{f}) = T(\mathbf{x}) = \tilde{g}(\mathbf{\Phi}'\mathbf{f}, m)$. Having $\tilde{g}$ vary with m, while $g(\mathbf{f})$ does not, implies that the map $\mathbf{\Phi}'\mathbf{f}$ has introduced such dependence. But since m cannot be recovered from $\mathbf{f}$, this must be false. Thus $\tilde{g} = \tilde{g}(\mathbf{\Phi}'\mathbf{f})$, and $\mathbf{z} = \mathbf{\Phi}'\mathbf{f}$ is sufficient for y.

3.2 MNIR for sentiment in text: collapsibility and random effects

For text-sentiment response factor specification, we focus on untransformed $v_i = y_i$ and discretized $v_i = \text{step}(y_i)$ along with their analogues for multivariate sentiment. The former is appropriate for categorical sentiment (e.g., political party, or 1-5 star rating) and, for reasons discussed below, the latter is used with continuous sentiment (e.g., vote-share is rounded to the nearest whole percentage, and in general one can bin and average y by quantiles). Regardless, our methods apply under generic $\mathbf{v}(y_i)$ including, e.g., the expansions of Cook (2007).

Given this setting of discrete $\mathbf{v}_i$, MNIR estimation can often be based on the *collapsed* counts that arise by aggregating within factor level combinations. For example, since sums of multinomials with equal probabilities are also multinomial, given shared intercepts (i.e., $u_{ij} = 0$) and writing the support of $\mathbf{v}_i$ as $\mathcal{V}$, the likelihood for the model in (5) is exactly the same as that from, for $\mathbf{v} \in \mathcal{V}$ with $\mathbf{x}_\mathbf{v} = \sum_{i:\mathbf{v}_i=\mathbf{v}} \mathbf{x}_i$ and $m_\mathbf{v} = \sum_{i:\mathbf{v}_i=\mathbf{v}} m_i$,

$$\mathbf{x}_\mathbf{v} \sim \text{MN}(\mathbf{q}_\mathbf{v}, m_\mathbf{v}), \quad \text{where } q_{\mathbf{v}j} = \frac{e^{\eta_{\mathbf{v}j}}}{\sum_{l=1}^{p} e^{\eta_{\mathbf{v}l}}} \text{ and } \eta_{\mathbf{v}j} = \alpha_j + \mathbf{v}\boldsymbol{\varphi}_j. \tag{6}$$

Since pooling documents in this way leaves only as many 'observations' as there are levels in the support of $\mathbf{v}_i$, it can lead to dramatically less expensive estimation.

Under the marginal model of (6), $\mathbf{\Phi}$ is the *population average* effect of $\mathbf{v}$ on $\mathbf{x}$. One needs to be careful in when and how estimates from this model are used in SR projection, since conditional document-level validity of these results is subject to the usual collapsibility requirements for analysis of categorical data (e.g., Bishop et al., 1975). In particular, omitted variables must be conditionally independent of $\mathbf{x}_i$ given $\mathbf{v}_i$; this can usually be assumed for sentiment-related variables (e.g., a congress person's voting record is ignored given their party and vote-share). Covariates that act on $\mathbf{x}_i$ independent of $\mathbf{v}_i$ should be included in MNIR, as part of the equation for subject effects $\mathbf{u}_i$ (e.g., although it is not considered in this article, it might be best to condition on geography when regressing political speech onto partisanship). The sufficient reduction result of (3.1) is then conditional on these sentiment-independent variables, such that they (or their SR projection) *may* need to be used as inputs in forward regression.

It is often unreasonable to assume that known factors account for all variation across documents, and treating the $\mathbf{u}_i$ of (5) as random effects independent of $\mathbf{v}_i$ provides a mechanism for explaining such heterogeneity and understanding its effect on estimation. Omitting $\mathbf{u}_i \perp\!\!\!\perp \mathbf{v}_i$ tends to yield estimated $\mathbf{\Phi}$ that is attenuated from its correct document-specific analogue (Gail et al., 1984), although the population-average estimators can be reliable in some settings; for example, Zeger et al. (1985) show consistency for the stationary distribution effect of covariates when the $\mathbf{u}_i$ encode temporal dependence (such as that between consecutive tokens in an N-gram text model). When their influence is considered negligible, it is common to simply ignore the random effects in estimation. In this article we also consider modeling $e^{u_{ij}}$ as independent gamma random variables, and use this to motivate a prior in 3.3 for the marginal random effects in a collapsed table. Another option would be to incorporate latent topics into MNIR and parametrize $\mathbf{u}_i$ through a linear factor model; this is especially appealing since SR projections onto estimated factor scores could then be used in forward regression.

This last point – on random effects and forward regression – is important: when $\mathbf{\Phi}$ is estimated with random effects, Section 3.1 only establishes sufficiency of $\mathbf{z}_i$ conditional on $\mathbf{u}_i$. Marginal sufficiency would follow from $\mathrm{p}(\mathbf{v}_i|\mathbf{u}_i, \mathbf{\Phi}'\mathbf{x}_i) = \mathrm{p}(\mathbf{v}_i|\mathbf{\Phi}'\mathbf{x}_i)$, which for $\mathbf{u}_i \perp\!\!\!\perp \mathbf{v}_i$ requires $\mathbf{u}_i \perp\!\!\!\perp \mathbf{\Phi}'\mathbf{x}_i$. Thus, information about $\mathbf{v}_i$ from this marginal dependence is lost when (as is usually necessary) $\mathbf{u}_i$ is omitted in regression of $\mathbf{v}_i$ onto $\mathbf{z}_i$. Section 5 shows that random effects in MNIR can be beneficial even if they are then ignored in forward regression. However, SR projection onto parametric representations of $\mathbf{u}_i$ is an open research interest.

It is clear that there are many relevant issues to consider when assessing an MNIR model, and it is helpful to have our sentiment regression problem placed within the well studied framework of contingency table analysis (e.g., Agresti, 2002, is a general reference). Ongoing work centers on inference according to specific dependence structures or random effect parametrizations. However, as illustrated in Section 5, even very simple MNIR models – measuring population average effects – allow SR projections that are powerful tools for forward prediction.

3.3 Prior specification

To complete the MNIR model, we provide prior distributions for the intercepts $\boldsymbol{\alpha}$, loadings $\boldsymbol{\Phi}$, and possible random effects $\mathbf{U} = [\mathbf{u}_{\mathbf{v}_1} \cdots \mathbf{u}_{\mathbf{v}_d}]'$, where d is the number of points in $\mathcal{V}$.

First, each phrase intercept is assigned an independent standard normal prior, $\alpha_j \sim \mathrm{N}(0, 1)$. This serves to identify the logistic multinomial model, such that there is no need to specify a *null* category, and we have found it diffuse enough to accomodate category frequencies in a variety of text and non-text examples. Second, we propose independent Laplace priors for each φ_{jk}, with coefficient-specific precision (or 'penalty') parameters λ_{jk}, such that $\pi(\varphi_{jk}) = \lambda_{jk}/2 \exp(-\lambda_{jk}|\varphi_{jk}|)$ for $j = 1 \ldots p$ and $k = 1 \ldots K$. The implied prior standard deviation for φ_{jk} is $\sqrt{2}/\lambda_{jk}$. Each λ_{jk} is then assigned a conjugate gamma hyperprior $\mathrm{Ga}(\lambda_{jk}; s, r) = r^s/\Gamma(s)\lambda_{jk}^{s-1}e^{-r\lambda_{jk}}$, yielding the joint gamma-Laplace prior introduced in (3). Hyperprior shape, s, and rate, r, imply expectation s/r and variance s/r^2 for each λ_{jk}.

As an example specification, consider variation in empirical token probabilities by level of the logical variables 'party = republican' for congressional speech and 'rating > 3' for we8there reviews. Standard deviation of finite $\log(\hat{q}_{\mathtt{true},j}/\hat{q}_{\mathtt{false},j})$ across tokens is 1.9 and 1.4 respectively, and given variables normalized to have $\mathrm{var}(v) = 1$ these deviations in log-odds correspond to a jump of two in v (from approximately -1 to 1). Hence, a coefficient standard deviation of around 0.7, implying $\mathbb{E}[\lambda_{jk}] = 2$, is at the conservative (heavy penalization) end of the range indicated by informal data exploration, recommending the exponential $\mathrm{Ga}(1, 1/2)$ as a penalty prior specification. In Section

5 we also consider shapes of 1/10 and 1/100, thus decreasing $\mathbb{E}[\lambda_{jk}]$ by two orders of magnitude, and find performance robust to these changes.

The above models have, with $s \leq 1$, hyperprior densities for φ_{jk} that are increasing as the penalty approaches zero (i.e., at MLE estimation). This strategy has performed well in many applications, both for text analysis and otherwise, when dimension is not much larger than 10^3. However, in examples with vocabulary sizes reaching 10^5 and higher, it is useful to increase both shape and rate for fast convergence and to keep the number of non-zero term loadings manageably small. As an informal practical recipe, if estimated $\mathbf{\Phi}$ is less sparse than desired and you suspect overfit, increase s. Following the discussion in 4.3 on hyperprior variance and algorithm convergence, if the optimization is taking too long or getting stuck in a minor mode, multiply both s and r by a constant to keep $\mathbb{E}[\lambda_{jk}]$ unchanged while decreasing $\mathrm{var}[\lambda_{jk}]$.

Finally, we use $\exp[u_{ij}] \sim \mathrm{Ga}(1,1)$ independent for each i and j as an illustrative random effect model. Considering $e^{u_{ij}}$ as a multiplier on relative odds, its mode at zero assumes some tokens are inappropriate for a given document, the mean of one centers the model on a shared intercept, and the fat right tail allows for occasional large counts of otherwise rare tokens. Counts are not immediately collapsable in the presence of random effects, but assumptions on the generating process for $\mathbf{x}_i$ unconditional on m_i can be used to build a prior model for their effect on aggregated counts: if each x_{ij} is drawn independent from a Poisson $\mathrm{Po}(e^{\alpha_j+u_{ij}+\mathbf{v}_i\varphi_j})$ with $\exp[u_{ij}] \sim \mathrm{Ga}(1,1)$, and $n_{\mathbf{v}} = \sum_i \mathbb{1}_{[\mathbf{v}_i=\mathbf{v}]}$, then $x_{\mathbf{v}j} \sim \mathrm{Po}(e^{\alpha_j+u_{\mathbf{v}j}+\mathbf{v}\varphi_j})$ with $\exp[u_{\mathbf{v}j}] \overset{ind}{\sim} \mathrm{Ga}(n_{\mathbf{v}},1)$. For convenience, we use a log-Normal approximation to the gamma and specify $u_{\mathbf{v},j} \sim \mathrm{N}(\log(n_{\mathbf{v}}) - 0.5\sigma^2_{\mathbf{v}}, \sigma^2_{\mathbf{v}})$ with $\sigma^2_{\mathbf{v}} = \log(n_{\mathbf{v}}+1) - \log(n_{\mathbf{v}})$. Note that $\sigma^2_{\mathbf{v}} \to 0$ as $n_{\mathbf{v}}$ grows, leading to static $u_{\mathbf{v},j}$ whose effect is equivalent to multiplying both numerator and denominator of $\exp[\eta_{\mathbf{v},j}]/\sum_l \exp[\eta_{\mathbf{v},l}]$ by a constant. Thus modeling random effects is unnecessary *under our assumed model* after aggregating large numbers of observations.

3.3.1 Motivation for independent gamma-Laplace priors

One unique aspect of this article's approach is the use of independent gamma-Laplace priors for each regression coefficient φ_{jk}. Part of the specification should not be surprising: the Laplace provides, as a scale-mixture of normal densities, a widely used robust alternative to the conjugate normal prior (e.g., Carlin et al., 1992). It also encourages sparsity in $\mathbf{\Phi}$ through a sharp density spike at $\varphi_{jk} = 0$, and MAP inference with fixed λ_{jk} is equivalent to likelihood maximization under an L_1-penalty in the *lasso* estimation and selection procedure of Tibshrani (1996). Similarly, conjugate gamma hyperpriors are a common choice in Bayesian inference for lasso regression (e.g., Park and Casella, 2008).

However, our use of independent precision for each coefficient, rather than a single shared λ, is a departure from standard practice. We feel that this provides a better representation of prior utility, and it avoids the overpenalization that can occur when inferring a single coefficient precision on data with a large proportion of spurious regressors. In their recent work on the Horseshoe prior, Carvalho et al. (2010) illustrate general practical and theoretical advantages of an independent parameter variance specification. As detailed in Section 4, our model also yields an estimation procedure, labeled the *gamma-lasso*, that corresponds to likelihood maximization under a specific nonconcave penalty; the estimators thus inherit properties deemed desirable by authors in that literature (beginning from Fan and Li, 2001).

Finally, given the common reliance on cross-validation (CV) for lasso penalty selection, it is worth discussing why we choose to do otherwise. First, our independent λ_{jk} penalties would require a CV search of impossibly massive dimension. Moreover, CV is just an estimation technique and, like any other, is sensitive to the data sample on which it is applied. As an illustration, Section 5.1 contains an example of CV-selected penalty performing far worse in out-of-sample prediction than those inferred under a wide range of gamma hyperpriors. CV is also not scaleable: repeated training and validation is infeasible on truly large applications (i.e., when estimating the model once is expensive). That said, one may wish to use CV to choose s or r in the hyperprior; since results are less sensitive to these parameters than they are to a fixed penalty, a small grid of search locations should suffice.

4 Estimation

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

$$\mathrm{p}(\boldsymbol{\Phi}, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \mathbf{U} \mid \mathbf{X}, \mathbf{V}) \propto \prod_{i=1}^{n} \prod_{j=0}^{p} q_{ij}^{x_{ij}} \pi(u_{ij}) \mathrm{N}(\alpha_j; 0, \sigma_\alpha^2) \prod_{k=1}^{K} \mathrm{GL}(\varphi_{jk}, \lambda_{jk}) \tag{7}$$

where $q_{ij} = \exp[\eta_{ij}] / \sum_{l=1}^{p} \exp[\eta_{il}]$ with $\eta_{ij} = \alpha_j + u_{ij} + \sum_{k=1}^{K} v_{ik}\varphi_{jk}$ and GL is our gamma-Laplace joint coefficient-penalty prior Laplace$(\varphi_{jk}; \lambda_{jk})Ga(\lambda_{jk}; r, s)$. We only consider here $u_{ij} = 0$ or $u_{ij} \stackrel{ind}{\sim} \mathrm{N}(0, \sigma_i^2)$ for $\pi(u_{ij})$, although sentiment-independent covariates can also be included trivially as additional dimensions of $\mathbf{v}_i$. Note that 'i' denotes an observation, but that in MNIR this will often be a combination of documents after the aggregation of Section 3.2.

Bayesian analysis of logistic regression typically involves posterior simulation, e.g. through Gibbs sampling with latent variables (Holmes and Held, 2006) or Metropolis sampling with posterior-approximating proposals (Rossi et al., 2005). Despite recent work on larger datasets and sparse signals (e.g., Gramacy and Polson, 2012), our experience is that these methods are too slow for text analysis applications. Even the more modest goal of posterior maximization presents considerable difficulty: unlike the usual high-dimension logistic regression examples, where K is big and p is small, our large response leads to a likelihood that is expensive to evaluate (due to normalization of each $\mathbf{q}_i$) and has a dense information matrix (from 4.2, $\partial^2 \log \mathrm{LHD}/\partial\varphi_{jk} = \sum_{i=1}^{n} m_i v_{ik}^2 q_{ij}(1 - q_{ij})$, which will not be zero unless v_{ik} is). As a result, commonly used path algorithms that solve over a grid of shared λ values (e.g., Friedman et al., 2010, as implemented in glmnet for R) do not work even for the small examples of this article.

We are thus motivated to develop super efficient estimation for sparse logistic regression. The independent gamma-Laplace priors of Section 3.3 are the first crucial aspect of our approach: it remains necessary to choose hyperprior s and r, but results are robust enough to misspecification that basic defaults can be applied. Section 4.1 derives the gamma-lasso (GL) non-concave penalty that results from MAP estimation under this prior. Second, Section 4.2 describes a coordinate descent algorithm for fast negative log posterior minimization wherein the GL penalties are incorporated at no extra cost over standard lasso regression. Lastly, 4.3 considers conditions for posterior log concavity and provides a check for *global* convergence.

4.1 Gamma-lasso penalized regression

Our estimation framework relies upon recognition that optimal λ_{jk} can always be written as a function of φ_{jk}, and thus does not need to be explicitly solved for in the joint objective.

Proposition 4.1. *MAP estimation for $\mathbf{\Phi}$ and $\boldsymbol{\lambda}$ under the independent gamma-Laplace prior model in (7) is equivalent to minimization of the negative log likelihood for $\mathbf{\Phi}$ subject to costs*

$$c(\mathbf{\Phi}) = \sum_{j=1}^{p}\sum_{k=1}^{K} c(\varphi_{jk}), \quad \textit{where} \;\; c(\varphi_{jk}) = s\log(1 + |\varphi_{jk}|/r) \tag{8}$$

Proof. Under conjugate gamma priors, the conditional posterior mode for each λ_{jk} given φ_{jk} is available as $\lambda(\varphi_{jk}) = s/(r + |\varphi_{jk}|)$. Any joint maximizing solution $[\hat{\mathbf{\Phi}}, \hat{\boldsymbol{\lambda}}]$ for (7) will thus consist of $\hat{\lambda}_{jk} = \lambda(\hat{\varphi}_{jk})$; otherwise, it is always possible to increase the posterior by replacing $\hat{\lambda}_{jk}$. Taking the negative log of (3) and removing constant terms, the influence of a $\mathrm{GL}(\lambda_{jk}, \varphi_{jk})$ prior on the negative log posterior is $-s\log(\lambda_{jk}) + (r + |\varphi_{jk}|)\lambda_{jk}$, which becomes $-s\log\left[(s/r)/(1 + |\varphi_{jk}|/r)\right] + s \propto s\log(1 + |\varphi_{jk}|/r)$ after replacing λ_{jk} with $\lambda(\varphi_{jk})$.

The implied penalty function is drawn in the left panel of Figure 2. Given its shape – everywhere concave with a sharp spike at zero – our gamma-lasso estimation fits within the general framework of nonconcave penalized likelihood maximization as outlined in Fan and Li (2001) and studied in many papers since. In particular, $c(\varphi_{jk})$ can be seen as a reparametrization of the 'log-penalty' described in Mazunder et al. (2011, eq. 10), which is itself introduced in Friedman (2008) as a generalization of the elastic net. Viewing estimation from the perspective of this literature is informative. Like the standard lasso, singularity at zero in $c(\varphi_{jk})$ causes some coefficients to be set to zero. However, unlike the lasso, the gamma-lasso has gradient $c'(\varphi_{jk}) = \text{sign}(\varphi_{jk})s/(r + |\varphi_{jk}|)$ which disappears as $|\varphi_{jk}| \to \infty$, leading to the property of *unbiasedness for large coefficients* listed by Fan and Li (2001) and referred to as *Bayesian robustness* by Carvalho et al. (2010). Other results from this literature apply directly; for example, in most problems it should be possible to choose s and r to satisfy requirements for the strong oracle property of Fan and Peng (2004) conditional on their various likelihood conditions.

It is important to emphasize that, despite sharing properties with cost functions that are purpose-built to satisfy particular notions of optimality, $c(\varphi_{jk})$ occurs simply as a consequence of proper priors in a principled Bayesian model specification. To illustrate the effect of this penalty, Figure 1 shows MAP coefficients for a simple logistic regression under changes to data and parameterization. In each case, gamma-lasso estimates threshold to zero before jumping to solution paths that converge to the MLE with increasing evidence. Figure 2 illustrates how these solution discontinuities arise due to concavity in the minimization objective, an issue that is discussed in detail in Section 4.3. Note that although the univariate lasso thresholds at larger values than the gamma-lasso, in practice we often observe greater sparsity under GL penalties since large signals are less biased and single coefficients are allowed to account for the effect of multiple correlated inputs. In contrast, standard lasso estimates also fix some estimates at zero but lead to continuous solution paths that never converge to the MLE.

4.2 Negative log posterior minimization by coordinate descent

Taking negative log and removing constant factors, maximization equates with minimization of $l(\boldsymbol{\Phi}, \boldsymbol{\alpha}, \mathbf{U}) + \sum_{j=1}^{p} (\alpha_j/\sigma_\alpha)^2 - \log \pi(\mathbf{U}) + c(\boldsymbol{\Phi})$, where l is the strictly convex

$$l(\boldsymbol{\Phi}, \boldsymbol{\alpha}, \mathbf{U}) = -\sum_{i=1}^{n} \left[\mathbf{x}_i'(\boldsymbol{\alpha} + \boldsymbol{\Phi}'\mathbf{v}_i + \mathbf{u}_i) - m_i \log \left(\sum_{j=1}^{p} \exp(\alpha_j + \boldsymbol{\varphi_j}'\mathbf{v}_i + u_{ij}) \right) \right]. \tag{9}$$

Full parameter-set moves for this problem are prohibitively expensive in high-dimension due to (typically dense) Hessian storage requirements. Hence, feasible algorithms make use of coordinate descent (CD), wherein the optimization cycles through updates for each parameter conditional on current estimates for all other parameters (e.g., Luenberger and Ye, 2008). Although conditional minima for logistic regression are not available in closed-form, one can bound the CD objectives with an easily solvable function and optimize that instead. In such bound-optimization (also known as majorization; Lange et al., 2000) for, say, $l(\theta)$, each move $\theta^{t-1} \rightarrow \theta^t$ proceeds by setting new θ^t as the minimizing argument to bound $b(\theta)$, where b is such that previous estimate θ^{t-1} minimizes $b(\theta) - l(\theta)$. Algorithm monotonicity is then guaranteed through the inequality $l(\theta^t) = b(\theta^t) + l(\theta^t) - b(\theta^t) \leq b(\theta^{t-1}) - \left[b(\theta^{t-1}) - l(\theta^{t-1}) \right] = l(\theta^{t-1})$.

Using $\theta^\star$ to denote a new value for a parameter currently estimated at θ, a quadratic bound for each element of (9) conditional on all others is available through Taylor expansion as

$$b(\theta^\star) = l(\boldsymbol{\Phi}, \boldsymbol{\alpha}, \mathbf{U}) + g_l(\theta)(\theta^\star - \theta) + \frac{1}{2}(\theta^\star - \theta)^2 H_\theta \tag{10}$$

where $g_l(\theta) = \partial l/\partial\theta$ is the current coordinate gradient and H_θ is an upper bound on curvature at the updated estimate, $h_l(\theta^\star) = \partial^2 l/\partial\theta^{\star 2}$. Quadratic bounding is also used in the logistic regression CD algorithms of Krishnapuram et al. (2005) and Madigan et al. (2005): the former makes use of a loose static bound on h_l, while the latter updates H_θ after each iteration to obtain tighter bounding in a constrained *trust-region* $\{\theta^\star \in \theta \pm \delta\}$ for specified $\delta > 0$. We have found that dynamic trust region bounding can lead to an order-of-magnitude fewer iterations, and Appendix A.2 derives H_θ as the least upper bound on $h_l(\theta^\star)$ for $\theta^\star$ within δ of θ.

In implementing this approach, coordinate-wise gradient and curvature for φ_{jk} are

$$g_l(\varphi_{jk}) = \frac{\partial l}{\partial \varphi_{jk}} = -\sum_{i=1}^{n} v_{ik}(x_{ij} - m_i q_{ij}) \;\text{ and }\; h_l(\varphi_{jk}) = \frac{\partial^2 l}{\partial \varphi_{jk}^2} = \sum_{i=1}^{n} m_i v_{ik}^2 q_{ij}(1 - q_{ij}), \tag{11}$$

and similar functions hold for random effects and intercepts but with covariates of one and without summing over i for random effects. Then under normal, say $\mathrm{N}(\mu_\theta, \sigma_\theta^2)$, priors for $\theta = u_{ij}$ or α_i, the negative log posterior bound is $B(\theta^\star) = b(\theta^\star) + 0.5(\theta - \mu_\theta)^2/\sigma_\theta^2$ which is minimized in $\{\theta \pm \delta\}$ at $\theta^\star = \theta - \mathrm{sgn}(\Delta\theta)\min\{|\Delta\theta|, \delta\}$ with $\Delta\theta = \left[g_l(\theta) + (\theta - \mu_\theta)/\sigma_\theta^2\right] / \left[H_\theta + 1/\sigma_\theta^2\right]$.

Although the GL penalty on φ_{jk} is concave and lacks a derivative at zero, coordinate-wise updates are still available in closed form. Suppressing the jk subscript, each coefficient update under GL penalty requires minimization of $B(\varphi^\star) = g_l(\varphi)(\varphi^\star - \varphi) + \frac{1}{2}(\varphi^\star - \varphi)^2 H_\varphi + s\log(1 + |\varphi^\star|/r)$ within the trust region $\{\varphi^\star \in \varphi \pm \delta : \mathrm{sgn}(\varphi^\star) = \mathrm{sgn}(\varphi)\}$. This is achieved by finding the roots of $B'(\varphi^\star) = 0$ and, when necessary, comparing to the bound evaluated at zero where B' is undefined. Setting $B'(\varphi^\star) = 0$ yields the quadratic equation

$$\varphi^{\star 2} + (\mathrm{sgn}(\varphi) r - \tilde{\varphi})\,\varphi^\star + \frac{s}{H_\varphi} - \mathrm{sgn}(\varphi) r\tilde{\varphi} = 0 \tag{12}$$

with characteristic $(\mathrm{sgn}(\varphi) r + \tilde{\varphi})^2 - 4s/H_\varphi$, where $\tilde{\varphi} = \varphi - g_l(\varphi)/H_\varphi$ would be the updated coordinate for an MLE estimator. From standard techniques, for $\{\varphi^\star : \mathrm{sgn}(\varphi) = \mathrm{sgn}(\varphi^\star)\}$ this function will have at most one real minimizing root – that is, with $H_\varphi > s/\left(r + |\varphi^\star|\right)^2$. Hence, each coordinate update is to find this root (if it exists) and compare $B(\varphi^\star)$ to $B(0)$. The minimizing value (0 or possible root $\varphi^\star$) dictates our parameter move $\Delta\varphi$, and this move is truncated at $\mathrm{sgn}(\Delta\varphi)\delta$ if it exceeds the trust region. Finally, when $\varphi = 0$, repeat this procedure for both $\mathrm{sgn}(\varphi) = \pm 1$; at most one direction will lead to a nonzero solution.

As it is inexpensive to characterize roots for $B'(\varphi^\star)$, the gamma-lasso does not lead to any noticeable increase in computation time over standard lasso algorithms (e.g., Madigan et al., 2005). Crucially, tests for decreased objective can performed on the bound function, instead of the full negative log posterior. Figure 3 shows objective and bound functions around the converged solution for three phrase loadings from regression of we8there reviews onto overall rating. With

$\delta = 0.1$, B provides tight bounding throughout this neighborhood. Behavior around the origin is most interesting: the solution for *chicken wing*, a low-loading negative term, is at $B'(\varphi^\star) = 0$ just left of the singularity at zero, while *ate here* falls in the sharp point at zero. The neighborhood around *first date*, a high-loading term, is everywhere smooth.

4.3 Posterior log concavity and algorithm convergence

Since the gamma-lasso penalty is everywhere concave, our minimization objective is not guaranteed to be convex. This is illustrated by the right two plots of Figure 2, where a very low-information likelihood (four observations) can be combined with a relatively diffuse prior on λ ($s = 1$, $r = 1/2$) to yield concavity near zero. The effect of this is benign when the gradient is the same direction on either side of the origin (as in the right panel of 2), but in other cases it will lead to local minima at zero away from the true global solution (as in the center panel). Such non-convexity is the cause of the discontinuities in the solution paths of Figure 1.

From the second derivative of $l(\varphi_{jk}) + c(\varphi_{jk})$, the conditional objective for φ_{jk} will be concave only if $h_l(\varphi_{jk} = 0) < s/r^2$ – that is, if prior variance on λ_{jk} is greater than the negative log likelihood curvature at $\varphi_{jk} = 0$. In our experience, this problem is rare: the likelihood typically overwhelms penalty concavity and real examples behave like those shown in Figure 3. Moreover, although it is possible to show stationary limit points for CD on such nonconvex functions (e.g. Mazunder et al., 2011), we advocate avoiding the issue through prior specification. In particular, hyperprior shape and rate can be raised to decrease $\mathrm{var}(\lambda_{jk})$ while keeping $\mathbb{E}[\lambda_{jk}]$ unchanged. Although this may require more prior information than desired, it is the amount necessary to have both fast MAP estimation and estimator stability. If you want to use more diffuse priors, you should pay the computational price of marginalization and mean inference (as in, e.g., Gramacy and Polson, 2012).

Even convexity in the coordinate updates is no guarantee of full objective convexity. However, we close this section by showing that the joint problem of optimizing both $\boldsymbol{\lambda}$ and $\boldsymbol{\Phi}$ *is* convex and has a single global minimum. Hence, we can derive gradient conditions on this expanded objective, and one is always able to check the estimation for global convergence.

Proposition 4.2. $[\boldsymbol{\Phi}, \boldsymbol{\alpha}, \mathbf{U}]$ *estimated following Section 4.2 will correspond to the global MAP of (7) if and only if* $G(\varphi_{jk}) = \mathrm{sgn}(\varphi_{jk})s/(r + |\varphi_{jk}|) - \sum_{i=1}^{n} v_{ik}(x_{ij} - m_i q_{ij})$ *is zero for* $\varphi_{jk} \neq 0$*, and is negative and positive in its left and right limits respectively around* $\varphi_{jk} = 0$.

Proof. Since the objective for $[\boldsymbol{\alpha}, \mathbf{U}]$ given $\boldsymbol{\Phi}$ is strictly convex, these will always be global conditional solutions. We can thus focus on the negative log conditional posterior for $[\boldsymbol{\Phi}, \boldsymbol{\lambda}]$, written $l(\boldsymbol{\Phi}) - \sum_{j=1}^{p} \sum_{k=1}^{K} s \log(\lambda_{jk}) + (r + |\varphi_{jk}|)\lambda_{jk}$. With $\lambda_{jk} > 0$, the first two terms are convex in $\boldsymbol{\Phi}$ and $\boldsymbol{\lambda}$, respectively, and the third term is jointly convex (but not strictly so) in $\boldsymbol{\Phi}$ and $\boldsymbol{\lambda}$, such that this function has a single minimum with each component at either the origin or a point of zero gradient. Taking derivatives and replacing $\lambda_{jk} = s/(r + |\varphi_{jk}|)$ yields $G(\varphi_{jk})$.

This simple result removes any uncertainty about global convergence, a standard issue with nonconcave penalization routines. Our fitted examples of Section 5 all satisfy the test in (4.2).

5 Examples

We now apply our framework to the datasets of Section 1.1. 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

Our prediction performance study considers three text analyses: both constituent percentage voteshare for G.W. Bush (bushvote) and Republican party membership (gop) regressed onto speech for a member of the 109$^{\text{th}}$ US congress, and a user's overall rating (overall) regressed onto the

ACCEPTED MANUSCRIPT

content of their we8there restaurant review. In each case, we report root mean square error or misclassification rate over 100 training and validation iterations. Full results and study details are provided in Appendix A.3, and performance for a subset of models is plotted in Figure 4. Here, we focus on some main comparisons that can be drawn from the study.

MNIR is considered under three different hyperprior specifications, with rate $r = 1/2$ and shapes of $s = 1/100$, $1/10$, and 1. Response factors are $v_i = y_i$ for gop and overall, and v_i is set as y_i rounded by whole number for bushvote (note that instead setting $v_i = y_i$ here leads to no discernable improvement). In each case, MNIR is fit for observations binned by factor level. We consider models both with and without independent random effects. As predicted, performance is unaffected by random effects for discrete y_i, where we are collapsing together hundreds of observations. However, they do improve out-of-sample performance by approximately 1.5% for bushvote, where only a small number of speakers are binned at each whole percentage point. Hence, detailed MNIR results are reported with random effects included only for bushvote. Finally, resulting SR scores $z_i = \boldsymbol{\varphi}'\mathbf{f}_i$ are incorporated into a variety of forward regression models: linear $\mathbb{E}[y_i] = \alpha + \beta z_i$ and quadratic $\mathbb{E}[y_i] = \alpha + \beta_1 z_i + \beta_2 z_i^2$ for bushvote, logistic $\mathbb{E}[y_i] = \exp[\alpha + \beta z_i]/(1 + \exp[\alpha + \beta z_i])$ for gop, and linear and proportional-odds logistic $\mathrm{p}(y_i \leq c) = \exp[\alpha_c - \beta z_i]/(1 + \exp[\alpha_c - \beta z_i])$, $c = 1 \ldots 5$, for overall.

Performance is very robust to changes in the MNIR hyperprior. Figure 4 shows little difference between otherwise equivalent models using the conservative default $s = 1$ and the lowest expected penalty $s = 1/100$; results for $s = 1/10$ are squeezed in-between. In congressional speech examples $s = 1/100$ has a slight edge; phrases here have already been pre-selected for partisanship and are thus largely relevant to the sentiment. On the other hand, $s = 1$ is the best performing shape for the we8there example, where phrases were only filtered by a minimum document threshold. Looking at forward regressions, the problem specific quadratic bushvote (see Section 5.2 for justification) and proportional odds overall (accounting for ordinal response) forward regressions provide lower average out-of-sample error rates at the price of slightly higher variability across

ACCEPTED MANUSCRIPT

iterations, when compared to simple linear forward regression.

As comparators, we consider text-specific LDA (both supervised and standard topic models) as well as an assortment of generic regression techniques: lasso penalized linear (bushvote and overall) and binary logistic (gop) regression, with penalty either optimized under our gamma hyperpriors (gop), marginalized in MCMC (bushvote), or tuned through CV (all examples); first-direction PLS (bushvote and overall); and support vector machines (gop). In *every* comparison, gamma-lasso MNIR provides higher quality predictions with lower run-times. The only similar predictive performance was for LDA with 25 and 50 topics in the bushvote example, at 15-50 times higher computational cost. Note that, given the size of real text analysis applications, we view the speed and scaleability of MNIR as a primary strength and only considered feasible alternatives, with short Gibbs runs for 50 topic sLDA and the Bayesian lasso (7-9 min) at the very high end of our runtimes. Moreover, both sLDA and CV lasso occasionally fail to converge (these runs were excluded); this never happened for MNIR.

Among comparators, the multinomial topic models outperform generic alternatives. Interestingly, LDA combined with simple regression outperforms sLDA in both congress examples. Again, this is probably due to pre-selection of phrases: topics are relevant to ideology regardless of supervision, and the extra parameters in sLDA are not worth their cost in degrees of freedom. Moreover, the simpler LDA models can be fit with the MAP estimation of Taddy (2012b), whereas sLDA is applied here through a slow-to-converge Gibbs sampler (we note that the original sLDA paper uses a variational EM algorithm). However, in the we8there data, the extra machinery of sLDA offers a clear improvement over unsupervised LDA, as should be the case in many text applications. Finally, in an important side comparison, binary logistic regressions were fit for gop regressed onto phrase frequencies using both CV and independent gamma hyperpriors for the lasso penalty. The scaleable, low-cost, gamma-lasso yields large performance improvements over a CV optimized model, regardless of hyperprior specification.

5.2 Application: partisanship and ideology in political speeches

For the data of Section 1.1.1, we have two sentiment metrics of interest: an indicator for party membership, and each speaker's constituent vote-share for Bush in 2004. Since the two independents caucused with Democrats, the former metric can be summarized in gop as a two-party *partisanship*. Following the political economy notion that there should be little discrepancy between voter and representative beliefs, bushvote provides a measure of *ideology* as expressed in support for G.W. Bush (and lack of support for John Kerry) in the context of that election.

Figure 5 shows MNIR fit in separate models for each of gop and bushvote, as studied in Section 5.1. For partisanship, fit with $s = 1/100$ and $r = 1/2$, a simple univariate logistic forward regression yields clear discrimination between parties; 8.5% (45 speakers) are misclassified under a maximum probability rule. In the bushvote MNIR, fit under the same hyperprior but with inclusion of random effects, the resulting SR scores $z_i = \boldsymbol{\varphi}'\mathbf{f}_i$ increase quickly with vote-share at low (mostly Democrat) values and more slowly for high (mostly Republican) values. This motivates our quadratic forward regression for bushvote onto SR score, the predictive mean of which is plotted in Figure 5 (with R^2 of 0.5). However, looking at the SR scores colored by party (red for Republicans, blue Democrats, green independents) shows that this curvature could instead be explained through different forward regression slopes by level of gop, implying that the relationship between language and ideology is party-dependent.

Given the above, a more useful model might consider text reduction that allows interaction between party and ideology. For example, we can build orthogonal bivariate sentiment factors as gop and bushvote minus the gop-level means, say votediff (again, rounded to the nearest whole percentage). Figure 6 shows fitted values for such a model, including random effects and with hyperprior shape increased to $s = 1/10$ to reflect a preference for smaller conditional coefficients. In detail, with z_{gop} and z_{votediff} the two dimensions of SR scores from MNIR $\mathbf{x} \sim \text{MN}(\mathbf{q}(v_{\text{gop}}, v_{\text{votediff}}), m)$, normalized for ease of interpretation, the fitted forward model is

$$\mathbb{E}[\text{bushvote}] = 51.9 + 6.2z_{\text{gop}} + 5.2z_{\text{votediff}} - 1.9z_{\text{gop}}z_{\text{votediff}}. \tag{13}$$

Thus a standard deviation increase in either SR direction implies a 5-6% increase in expected vote-share, and each effect is dampened when the normalized SR scores have the same sign.

The right panel of Figure 6 shows fitted expected counts $q_j m$ against true nonzero counts in our bivariate MNIR model fit; with random effects to account for model misspecification, there appears to be no pattern of overdispersion. The only clear outlier in forward regression is Chaka Fattah (D-PA) with a standardized residual of -5.2; he uttered a token in our sample only twice: once each for rate.return and billion.dollar. Finally, Figure 7 plots response factor loadings for a select group of tokens. Among other lessons, we see that racial identity rhetoric (african.american.latino, black.caucu) points towards the left wing of the Democratic party, while discussion of hate crimes is indicative of a moderate Republican. A few large loadings are driven by single observations: for example, violent.sexual.predator contributes more than 0.1% of speech for only Byron Dorgan, a Democratic Senator in Bush-supporting North Dakota. However, this is not the rule and most term loadings affect many speakers.

5.3 Application: on-line restaurant reviews

For the data of Section 1.1.2, our sentiment consists of five correlated restaurant ratings (each on a five point scale) that accompany every review. The left panel of Figure 8 shows MNIR for review content regressed onto the single overall response factor, as studied in Section 5.1. The true overall rating has high correlation (0.7) with our SR scores, despite considerable overlap between scores across rating levels. The right plot of Figure 8 shows probabilities for each increasing overall rating category, as estimated in the proportional-odds logistic forward regression, $\mathrm{p}(\text{overall} \leq c) = \exp[\alpha_c - \beta z_{\text{overall}}]/(1 + \exp[\alpha_c - \beta z_{\text{overall}}])$. Again, z_{overall} is normalized here to have mean zero and standard deviation of one in our sample. This model has $\beta = 2.3$, implying that the odds of being

ACCEPTED MANUSCRIPT

at or above any given rating level are multiplied by $e^{2.3} \approx 10$ for every standard deviation increase in the SR score.

Looking to explore aspect-specific factors, Figure 9 shows top-30 absolute value loadings in MNIR for review token-counts onto *all five* dimensions of sentiment. Influential terms on either side of the rating spectrum can be easily connected with elements of a good or bad meal: plan.return, best.meal, and big.portion are good, while sent.back, servic.terribl, and food.bland are bad. The largest loadings appear to be onto overall and food aspects, with service slightly less important and loadings for value and atmosphere quickly decreasing in size. This would indicate that the reviews focus on these elements in that order.

6 Discussion

The promising results of Section 5 reinforce a basic idea: a workable inverse specification can introduce information that leads to more efficient estimation. Given the multinomial model as a natural inverse distribution for token-counts, analysis of sentiment in text presents an ideal setting for inverse regression. While the approach of not *jointly* modeling a corresponding forward regression falls short of full Bayesian analysis, such inference would significantly complicate estimation and detract from our goal of providing a fast default method for supervised document reduction. We are happy to take advantage of parametric hierarchical Bayesian inference for the difficult MNIR estimation problem, and suggest that application appropriate techniques for low-dimensional forward regression should be readily available.

Although the illustrative applications in this article are quite simple, the methods scale to far larger datasets. Collapsing observations across sentiment factors for MNIR yields massive computational gains: training data need only include token counts tabled by sentiment level, and as an example, in Taddy (2012a) this allows MNIR runs of only a few seconds for 1.6 million twitter posts scored as positive or negative. Moreover, we see no reason why gamma-lasso logistic regression, which was developed specifically for large response settings, should not be viewed as an efficient option in generic penalized regression. Finally, current collaborations that use MNIR for text analysis include study of partisanship in the US congressional record from 1873 to present, and an attempt to quantify the economic content of news in 20 years of Wall Street Journal editions. In each case, we are considering a more rigorous treatment of the identification of single sentiment dimensions and controlling for related endogenous variables; this work shows MNIR's promise as the basis for a variety of text related inference goals.

Appendix

A.1 Slant and Partial Least Squares

The GS slant index for document i is $z_i^{\text{slant}} = \sum_{j=1}^{p} b_j(f_{ij} - a_j)/\sum_{j=1}^{p} b_j^2$, with parameters obtained through ordinary least-squares (OLS) as $[a_j, b_j] = \arg\min_{a,b} \sum_{i=1}^{n}[f_{ij} - (a + by_i)]^2$ for $j = 0 \dots p$. 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 is also the first direction in partial least-squares; 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. Using the usual normalization applied in PLS, an improved slant measure is given by $z_i^{\text{slant}} = \sum_{j=1}^{p} f_{ij}\text{cor}(f_j, y_i)$. For vote-share regressed onto congressional speech in the data of Section 1.1.1, this change increases within-sample R^2 from 0.37 to 0.57.

Given $\hat{\mathbf{F}} = [\hat{\mathbf{f}}_1 \cdots \hat{\mathbf{f}}_p]$ as a normalized covariate matrix with mean-zero and variance-one columns, a PLS algorithm which highlights its inverse regression structure is as follows.

1. Set the initial response factor $\mathbf{v}_0 = \mathbf{y} = [y_1 \dots y_n]'$, and for $k = 1, \dots, K$:
    - Loadings are $\boldsymbol{\varphi}_k = \text{cor}(\hat{\mathbf{F}}, \mathbf{v}_{k-1}) = [\text{cor}(\hat{\mathbf{f}}_1, \mathbf{v}_{k-1}) \dots \text{cor}(\hat{\mathbf{f}}_p, \mathbf{v}_{k-1})]'$.
    - The k^{th} PLS direction is $\mathbf{z}_k = \boldsymbol{\varphi}_k'\hat{\mathbf{F}}$.
    - The new response factors are $\mathbf{v}_k = \mathbf{v}_{k-1} - [\mathbf{z}_k'\mathbf{v}_{k-1}/(\mathbf{z}_k'\mathbf{z}_k)]\mathbf{z}_k$.

ACCEPTED MANUSCRIPT

2. Set $\hat{\mathbf{y}}$ as OLS fitted values for regression of $\mathbf{y}$ onto $\mathbf{Z}$, where $\mathbf{Z} = [\mathbf{z}_1 \cdots \mathbf{z}_K]$.

Orthogonalization of $\mathbf{v}_k$ with respect to $\mathbf{z}_k$ is algorithmically equivalent to predictor orthogonalization in the usual 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, such that PLS can be viewed as stagewise inverse regression.

A.2 Trust-region bound for logistic multinomial likelihood

The bounding used here is essentially the same as in Genkin et al. (2007) but for introduction of dependence upon v_{ik} that is missing from their version. We describe the bound for updates to φ_{jk}, but it applies directly to α_j or u_{ij} simply by replacing covariate values with one.

Given a trust region of $\varphi_{jk} \pm \delta$, the upper bound on $h_l(\varphi_{jk}) = \sum_{i=1}^{n} v_{ik}^2 m_i q_{ij}(1 - q_{ij})$ is $H_{jk} = \sum_{i=1}^{n} v_{ik}^2 m_i / F_{ij}$, where each F_{ij} is a lower bound on $1/(q_{ij} - q_{ij}^2) = 2 + e^{\eta_{ij}+\delta v_{ik}}/E_{ij} + E_{ij}/e^{\eta_{ij}+\delta v_{ik}}$, with $E_{ij} = \sum_{l=1}^{p} e^{\eta_{il}} - e^{\eta_{ij}}$. This target is convex in δ with minimum at $e^{\delta v_{ik}} = E_{ij}/e^{\eta_{ij}}$, such that

$$F_{ij} = \frac{e_{ij}}{E_{ij}} + \frac{E_{ij}}{e_{ij}} + 2 \text{ where } e_{ij} = \begin{cases} e^{\eta_{ij} - |v_{ik}|\delta} & \text{if } E_{ij} < e^{\eta_{ij} - |v_{ik}|\delta} \\ e^{\eta_{ij} + |v_{ik}|\delta} & \text{if } E_{ij} > e^{\eta_{ij} + |v_{ik}|\delta} \\ E_{ij} & \text{otherwise.} \end{cases}$$

We use unique δ_{jk} and update $\delta_{jk}^{\star} = \max\{\delta_{jk}/2, 2|\varphi_{jk}^{\star} - \varphi_{jk}|\}$ after each iteration.

A.3 Out-of-Sample Prediction Study Details

Each model was fit to 100 random data subsets and used to predict on the left-out sample. Tables report average root mean square error (RMSE) or percent misclassified (MC%), the percentage worse than best on this metric, and run-time in seconds (including count collapsing in MNIR).

We use R package implementations: lda for SLDA (Chang, 2011); glmnet for CV lasso regression (Friedman et al., 2010); monomvn for Bayesian lasso (Gramacy, 2012); kernlab for SVM (Karatzoglou et al., 2004); textir for MNIR, LDA, PLS, and gamma-lasso regression; and arm (Gelman et al., 2012) for the forward regression models that accompany MNIR and LDA. Penalty prior in MNIR is Ga(s, 1/2), (s)LDA Dirichlet precisions are $1/K$ for topic weights and $1/p$ for token probabilities, and sLDA assumes a forward error variance of 25% of marginal response variance. Unless otherwise specified, we apply package defaults. (S)LDA and MNIR use token counts; all others regress onto token frequencies.

Vote Share: Congressional speech with two-party vote share (%) as continuous response, training on 200 and predicting on 329. Constant mean RMSE is 13.4. MNIR models were fit *with* random effects; models without random effects are an average of 1.5% worse on RMSE but 20% faster. Bayes lasso uses a Ga(2,1/10) prior on λ and was run for 200 MCMC iterations after a burn-in of 100 (refer to monomvn for details).

	MNIR & Quadratic			MNIR & Linear			LDA & Linear					Supervised LDA					Lasso		PLS
	$s = 10^{-2}$	10^{-1}	1	$s = 10^{-2}$	10^{-1}	1	$K = 2$	5	10	25	50	$K = 2$	5	10	25	50	CV	Bayes	K=1
RMSE	10.7	10.7	10.8	10.9	10.9	10.9	11.7	11.3	11.1	10.9	10.9	12.9	12.1	11.7	12.3	15.1	13.7	15.7	15.9
% Worse	0	0	0	1	1	2	9	6	4	2	2	21	13	9	15	41	28	46	49
Run Time	2.2	2.3	2.1	2.2	2.3	2.1	1.2	2.4	6.2	29	112	43	75	128	288	508	0.9	410	0.1

Party Classification: Congressional speech data with 'Republican' as binary response, training on 200 and predicting on 329. Null model misclassification rate is 46%. MNIR models were fit *without* random effects which lead to the same misclassification but 40% longer average run-times. Lasso and gamma-lasso are applied in binary logistic regressions, with shape one and rate r for the latter, and SVM uses Gaussian kernels with misclassification cost C (refer to kernlab for details). LDA led to complete separation and SLDA failed to converge for $K > 10$.

	MNIR & Logistic			LDA & Logistic			Supervised LDA			Lasso	Gamma-Lasso				SVM		
	$s = 10^{-2}$	10^{-1}	1	$K = 2$	5	10	$K = 2$	5	10	CV	$r = 5$	25	50	100	C=1	100	1000
MC%	11	11	12	20	15	15	33	20	18	24	19	17	16	15	37	32	32
% Worse	0	0	2	76	36	30	188	75	54	115	68	49	42	35	224	182	180
Run Time	0.3	0.4	0.3	1.1	2.5	6.3	44	77	126	1.0	0.6	0.5	0.5	0.5	3.1	3.5	3.4

ACCEPTED MANUSCRIPT

Restaurant Rating: We8there reviews with ordinal rating response, training on 2000 and predicting on 4166. Constant mean RMSE is 1.35. Reported MNIR models were fit *without* random effects which lead to equivalent predictive performance but 15% longer average run-times.

	MNIR & POLR			MNIR & Linear			LDA & POLR					Supervised LDA					Lasso	PLS
	$s = 10^{-2}$	10^{-1}	1	$s = 10^{-2}$	10^{-1}	1	$K = 2$	5	10	25	50	$K = 2$	5	10	25	50	CV	$K = 1$
RMSE	1.08	1.08	1.07	1.09	1.09	1.10	1.19	1.17	1.20	1.23	1.23	1.15	1.13	1.14	1.15	1.16	1.24	1.25
% Worse	1	1	0	2	2	2	12	10	12	15	15	8	5	6	7	8	16	17
Run Time	0.6	0.6	0.5	0.3	0.4	0.3	2.5	13.4	28	61	167	53	90	154	341	651	54	2.2

References

Agresti, A. (2002). *Categorical Data Analysis* (2nd ed.). Wiley.

Bishop, Y., S. Fienberg, and P. Holland (1975). *Discrete Multivariate Analysis*. MIT Press.

Blei, D. M. and J. D. McAuliffe (2007). Supervised topic models. In *Neural Information Processing Systems*, Volume 21.

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

Bollen, J., H. Mao, and X.-J. Zeng (2011). Twitter mood predicts the stock market. *Journal of Computational Science 2*, 1–8.

Bura, E. and D. Cook (2001). Estimating the structural dimension of regressions via parametric inverse regression. *Journal of the Royal Statistical Society, Series B 63*, 393–410.

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.

Chang, J. (2011). *lda: Collapsed Gibbs sampling methods for topic models.* R package version 1.3.1.

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.

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

Friedman, J. H. (2008). Fast sparse regression and classification. Technical Report, Dept. of Statistics, Stanford University.

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

Gail, M. H., S. Wieand, and S. Piantadosi (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika 71*, 341–444.

Gelman, A., Y.-S. Su, M. Yajima, J. Hill, M. G. Pittau, J. Kerman, and T. Zheng (2012). *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models*. R package version 1.5-03.

ACCEPTED MANUSCRIPT

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.

Gramacy, R. B. (2012). *monomvn: Estimation for multivariate normal and Student-t data with monotone missingness*. R package version 1.8-10.

Gramacy, R. B. and N. G. Polson (2012). Simulation-based regularized logistic regression. *Bayesian Analysis 7*, 1–24.

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis 18*, 1–35.

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.

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

Karatzoglou, A., A. Smola, K. Hornik, and A. Zeileis (2004). kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software 11*(9), 1–20.

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.

Laver, M., K. Benoit, and J. Garry (2003). Extracting policy positions from political texts using words as data. *American Political Science Review 97*, 311–332.

Lehmann, E. L. and H. Sheffé (1950). Completeness, similar regions, and unbiased estimation – part 1. *Sankhyā: The Indian Journal of Statistics 10*, 305–340.

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

Li, L., R. D. Cook, and C.-L. Tsai (2007). Partial inverse regression. *Biometrika 94*, 615–625.

Loughran, T. and B. McDonald (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance 66*, 35–65.

Luenberger, D. G. and Y. Ye (2008). *Linear and Nonlinear Programming* (3rd ed.). Springer.

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

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

Mazunder, R., J. H. Friedman, and T. Hastie (2011). Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association 106*, 1125–1138.

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

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

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

Porter, M. F. (1980). An algorithm for suffix stripping. *Program 14*, 130–137.

Quinn, K., B. Monroe, M. Colaresi, M. Crespin, and D. Radev (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science 54*, 209–228.

Rossi, P. E., G. M. Allenby, and R. McCulloch (2005). *Bayesian Statistics and Marketing*. Wiley.

Schervish, M. J. (1995). *Theory of Statistics*. Springer.

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

Taddy, M. (2012a). Design and analysis of a text mining experiment. arXiv:1206.3776v1.

Taddy, M. (2012b). On estimation and selection for topic models. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS 2012)*.

Talley, E. L. and D. O'Kane (2011). The measure of a MAC: A machine-learning protocol for analyzing force majeure clauses in M&A agreements. *Journal of Institutional and Theoretical Economics 168*, 181–201.

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

Thomas, M., B. Pang, and L. Lee (2006). Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Emperical Methods in Natural Language Processing*.

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.

ACCEPTED MANUSCRIPT

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.

Yu, B., S. Kaufmann, and D. Diermeirer (2008). Classifying party affiliation from political speech. *Journal of Information Technology and Politics 5*, 33–49.

Zeger, S. L., K.-Y. Liang, and S. G. Self (1985). The analysis of binary longitudinal data with time-independent covariates. *Biometrika 72*, 31–38.

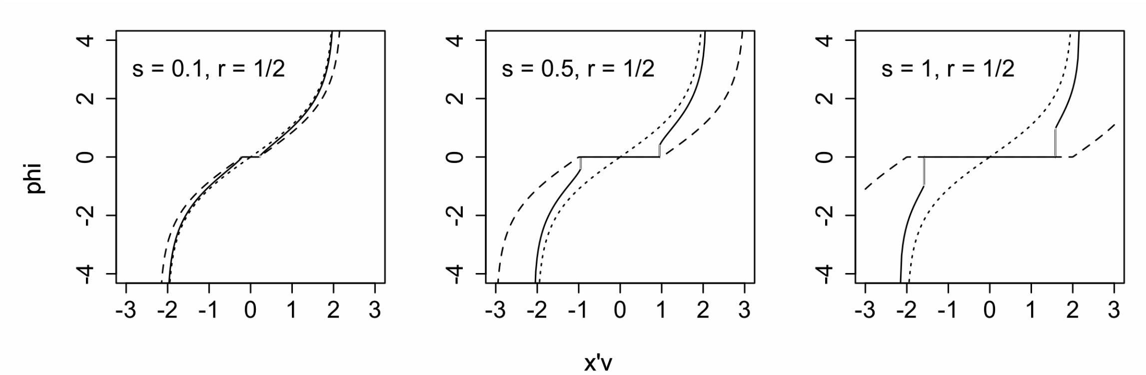

Figure 1: Maximizing solutions for univariate logistic regression log posteriors $L(\varphi) = \mathbf{x}'\mathbf{v}\varphi - \sum_i \log\left[1 + e^{\varphi v_i}\right] - \mathrm{pen}(\varphi)$, given $\mathbf{v} = [-1, -1, 1, 1]'$. The dotted line is the MLE, with $\mathrm{pen}(\varphi) = 0$, the dashed line is lasso $\mathrm{pen}(\varphi) = s|\varphi|/r$, and the solid line is gamma-lasso $\mathrm{pen}(\varphi) = s\log(1 + |\varphi|/r)$.

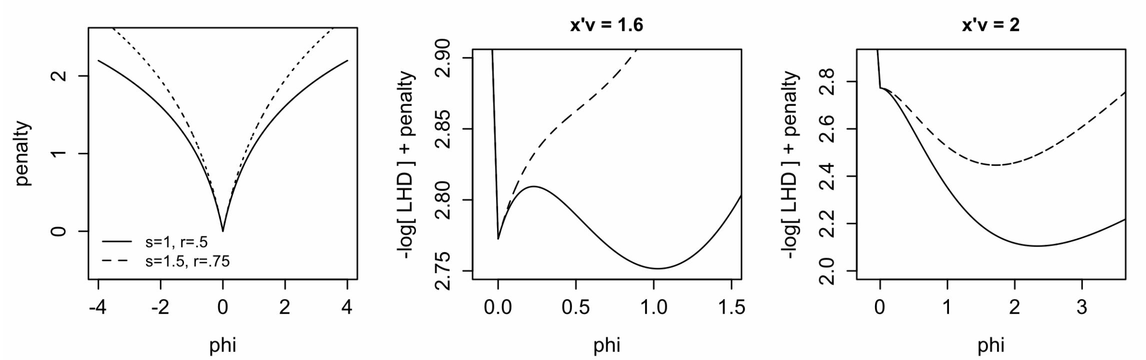

Figure 2: The left panel shows gamma-lasso penalty $s\log(1+|\varphi|/r)$ for $[s, r]$ of $[1, 1/2]$ (solid) and $[3/2, 3/4]$ (dashed). The right two plots show the corresponding minimization objectives, negative log likelihood plus GL penalty, near a solution discontinuity in the simple logistic regression of Figure 1.

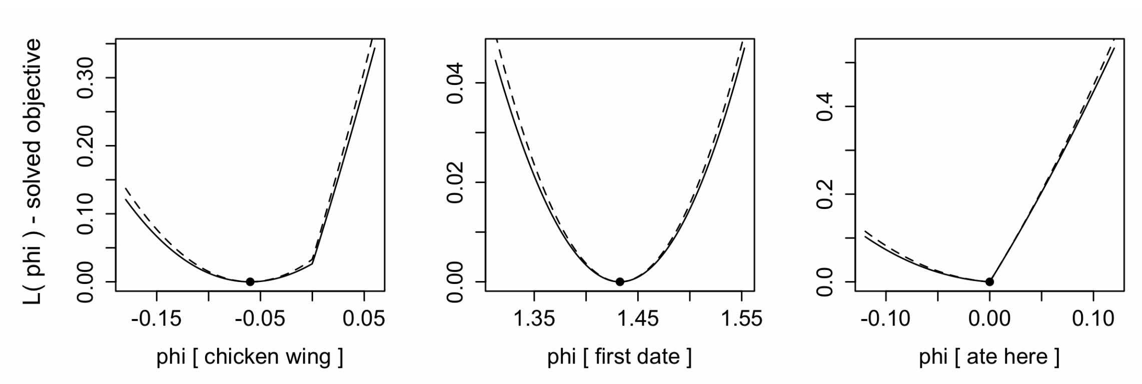

Figure 3: Coordinate objective functions at convergence in regression of we8there reviews onto overall rating. Solid lines are the true negative log likelihood and dashed lines are bound functions with $\delta = 0.1$. Both are shown for new $\varphi_j^\star$ as a difference over the minimum at estimated φ_j (marked with a dot).

ACCEPTED MANUSCRIPT

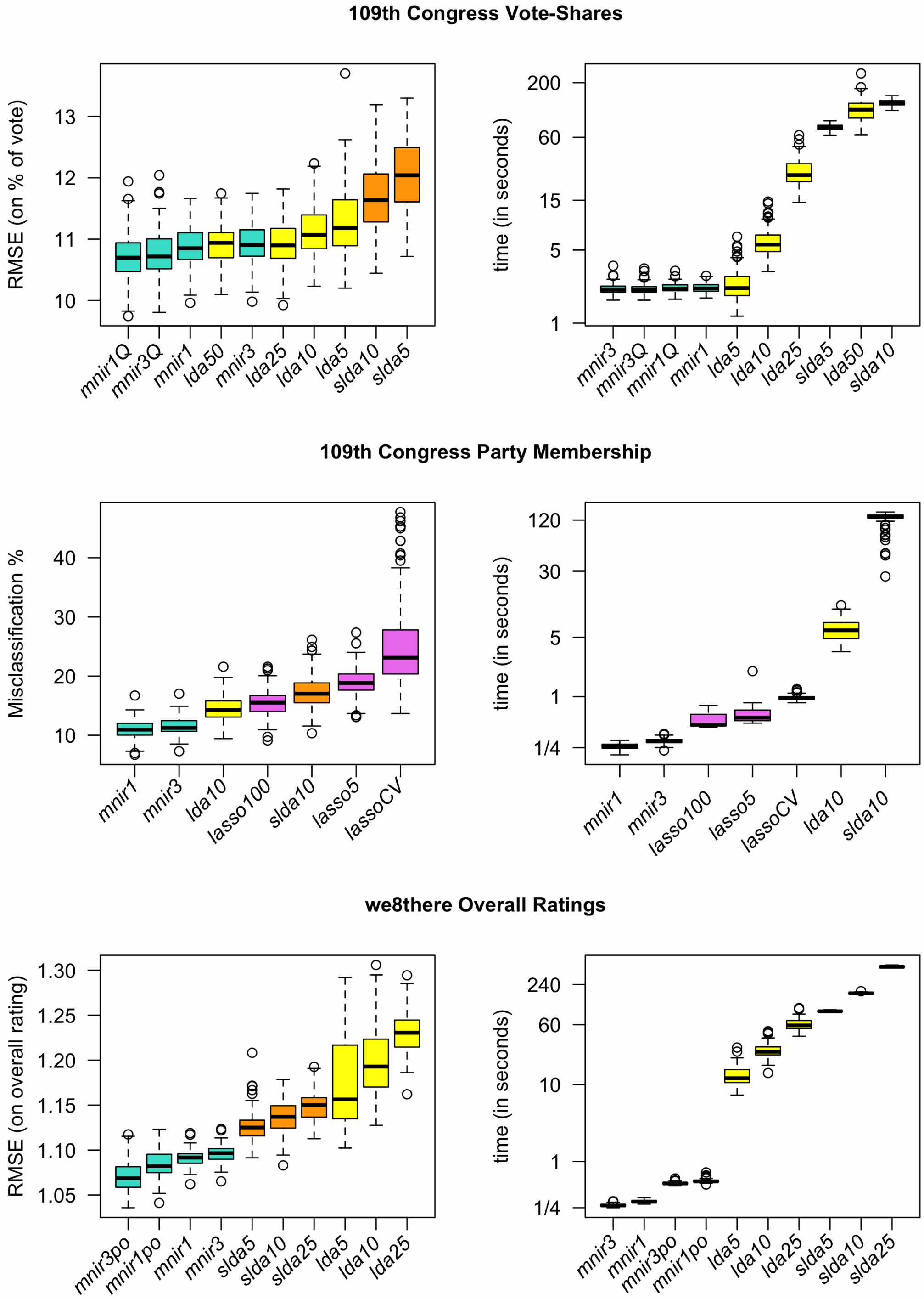

Figure 4: Out-of-sample performance and run-times for select models. For MNIR, 'Q' indicates quadratic and 'po' proportional-odds logistic forward regressions, while λ_j prior '1' is Ga(0.01, 0.5) and '3' is Ga(1, 0.5). We annotate with the number of topics for (s)LDA, and for binary Lasso regressions with either CV or the rate in an exponential penalty prior. Full details are in the appendix.

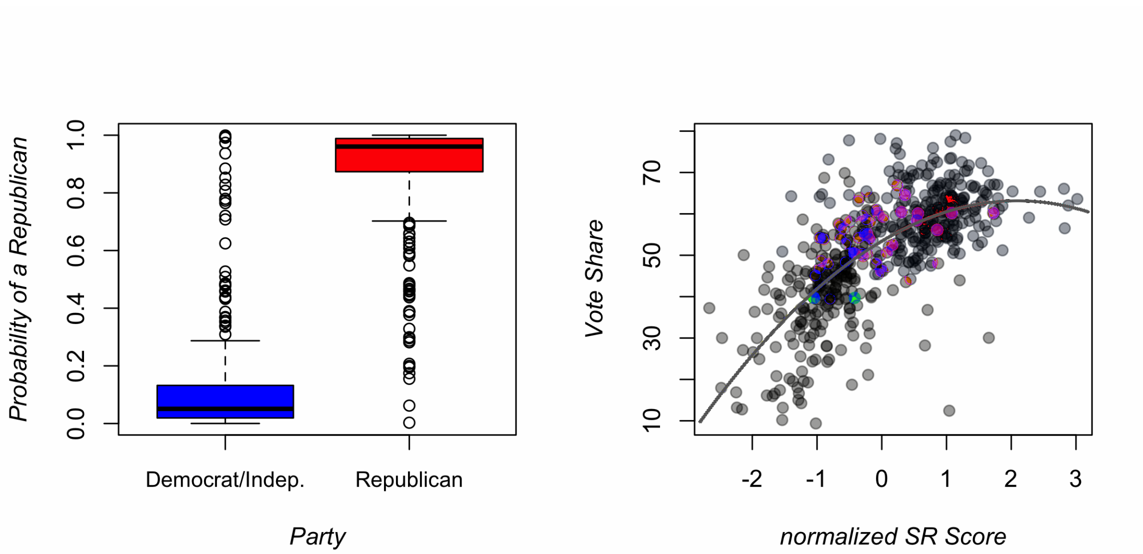

Figure 5: Separate MNIR fits for congressional speech onto each of party and vote-share. The right shows probabilities that each speaker is Republican and the left shows SR scores against bushvote.

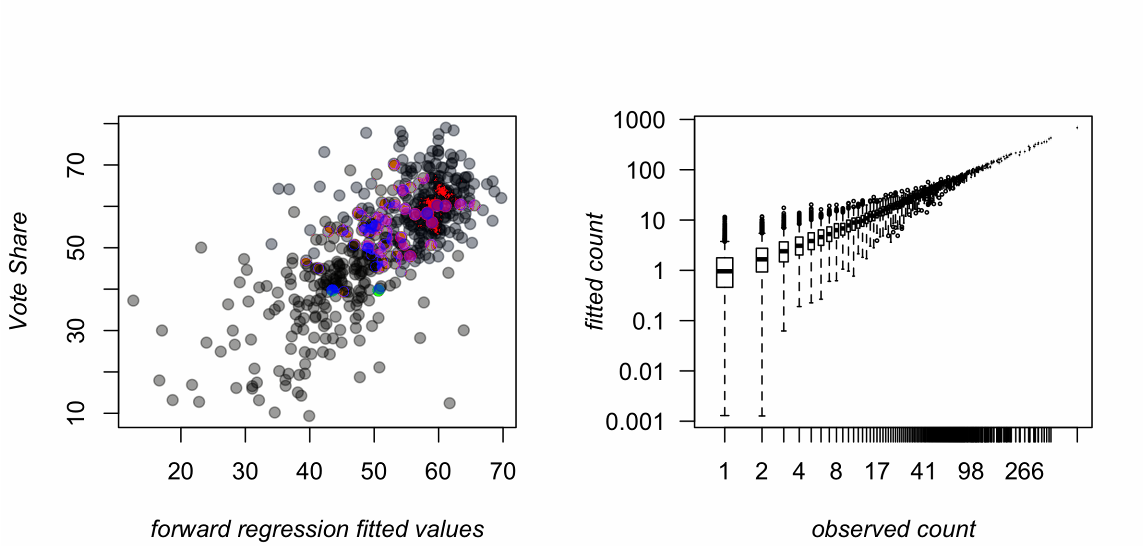

Figure 6: Bivariate ideology and partisanship MNIR. The left plot shows fitted values for a forward regression that interacts SR scores, and the right shows fitted vs observed token counts in MNIR.

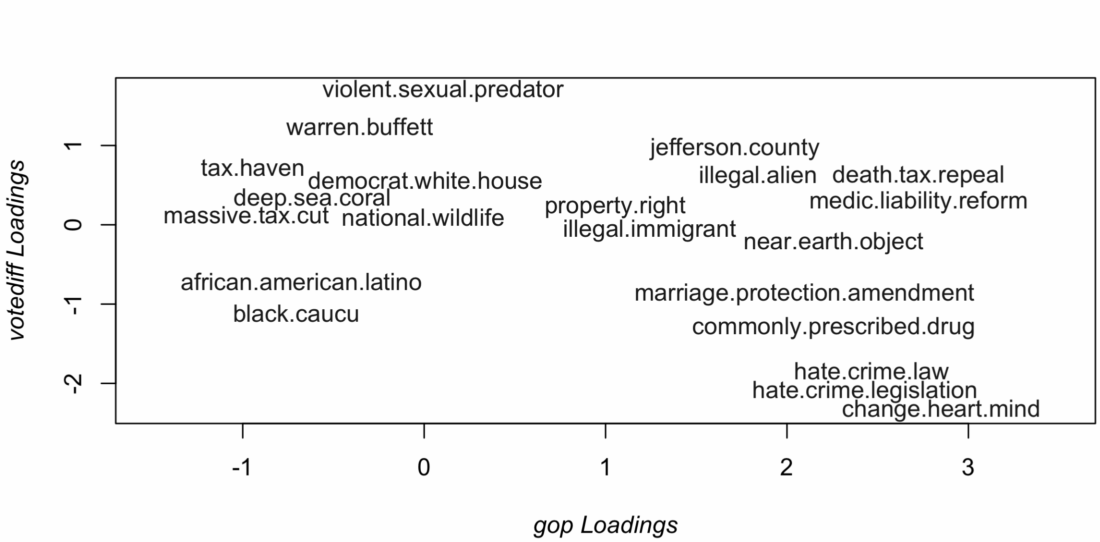

Figure 7: Select congressional speech term loadings in bivariate MNIR with party and vote-share.

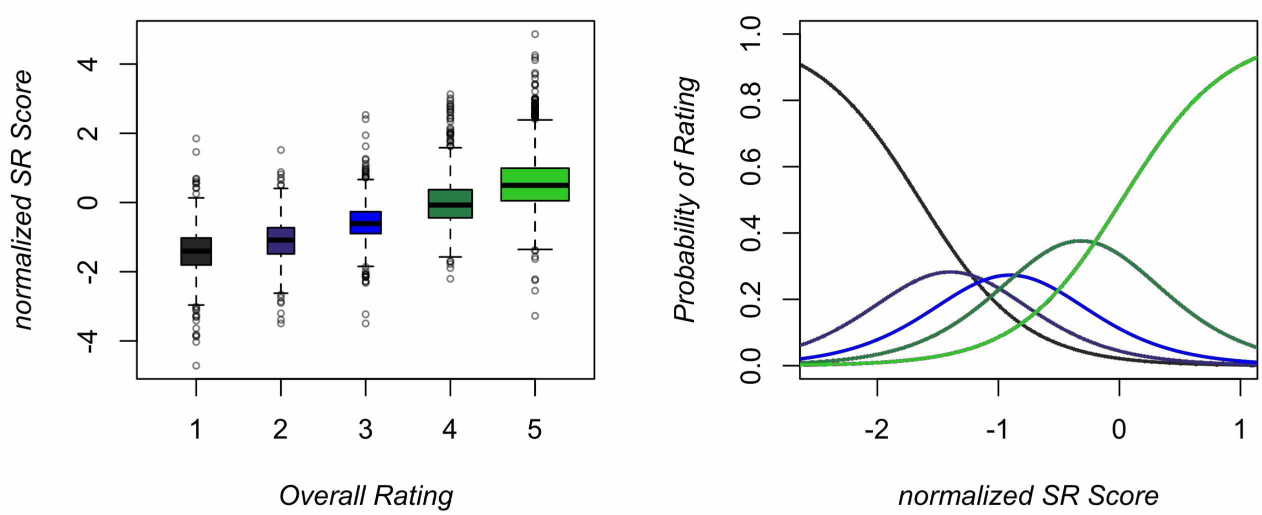

Figure 8: Sufficient reduction and forward model fit for inverse regression of we8there reviews onto the corresponding overall rating. The left plot shows SR score by true review rating, and the right shows proportional-odds logistic regression probabilities for each rating-level as a function of these SR scores.

Overall	Food	Service	Value	Atmosphere
plan.return	again.again	cozi.atmospher	big.portion	walk.down
feel.welcom	mouth.water	servic.terribl	around.world	great.bar
best.meal	francisco.bay	servic.impecc	chicken.pork	atmospher.wonder
select.includ	high.recomend	attent.staff	great.too	dark.wood
finest.restaur	cannot.wait	time.favorit	perfect.place	food.superb
steak.chicken	best.servic	servic.outstand	place.visit	atmospher.great
ask.waitress	kept.secret	servic.horribl	mahi.mahi	alway.go
love.restaur	food.poison	dessert.great	veri.reason	bleu.chees
good.work	outstand.servic	terribl.servic	babi.back	realli.cool
can.enough	far.best	never.came	low.price	recommend.everyon
after.left	food.awesom	experi.wonder	peanut.sauc	great.atmospher
come.close	best.kept	waitress.come	wonder.time	wonder.restaur
open.lunch	everyth.menu	time.took	live.entertain	love.atmospher
warm.friend	excel.price	servic.except	garlic.sauc	bar.just
spoke.manag	keep.come	final.came	great.can	expos.brick
definit.recommend	hot.fresh	new.favorit	absolut.best	back.drink
expect.wait	best.mexican	servic.awesom	place.best	fri.noth
great.time	best.sushi	sever.minut	year.alway	great.view
chicken.beef	pizza.best	best.dine	over.price	chicken.good
room.dessert	food.fabul	veri.rude	dish.well	bar.great
price.great	melt.mouth	peopl.veri	few.place	person.favorit
seafood.restaur	each.dish	poor.servic	authent.mexican	great.decor
friend.atmospher	absolut.wonder	ask.check	wether.com	french.dip
sent.back	foie.gras	real.treat	especi.good	pub.food
ll.definit	food.bland	never.got	like.sit	coconut.shrimp
anyon.look	menu.chang	non.exist	open.until	go.up
most.popular	noth.fanci	flag.down	open.daili	servic.fantast
order.wrong	back.time	least.minut	best.valu	gas.station
delici.food	food.excel	tabl.ask	just.great	pork.loin
fresh.seafood	worth.trip	won.disappoint	fri.littl	place.friend

Figure 9: High-loading phrases in each direction for regression of we8there reviews onto aspect ratings. Green tokens are positive, black are negative, and size is proportional to the absolute value of the loading.

ACCEPTED MANUSCRIPT