The Inverse Regression Topic Model

Maxim Rabinovich[†] MR608@CAM.AC.UK
Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ

David M. Blei BLEI@CS.PRINCETON.EDU
Department of Computer Science, Princeton University, 35 Olden Street, Princeton, NJ 08544

Abstract

Taddy (2013) proposed multinomial inverse regression (MNIR) as a new model of annotated text based on the influence of metadata and response variables on the distribution of words in a document. While effective, MNIR has no way to exploit structure in the corpus to improve its predictions or facilitate exploratory data analysis. On the other hand, traditional probabilistic topic models (like latent Dirichlet allocation) capture natural heterogeneity in a collection but do not account for external variables. In this paper, we introduce the inverse regression topic model (IRTM), a mixed-membership extension of MNIR that combines the strengths of both methodologies. We present two inference algorithms for the IRTM: an efficient batch estimation algorithm and an online variant, which is suitable for large corpora. We apply these methods to a corpus of 73K Congressional press releases and another of 150K Yelp reviews, demonstrating that the IRTM outperforms both MNIR and supervised topic models on the prediction task. Further, we give examples showing that the IRTM enables systematic discovery of in-topic lexical variation, which is not possible with previous supervised topic models.

1. Introduction

Probabilistic topic models are widely used to analyze large collections of documents. Given a corpus, topic models reveal its underlying themes, or *topics*, and how the documents exhibit them. Inferences from topic models can be used to navigate, organize, and analyze the collection.

Simple topic models are powerful, but they model documents in isolation. Typically, documents occur in a context, captured by associated variables, or *metadata*, and both predictive and exploratory applications require that this side information be taken into account. For example, political speeches are given in the context of a political affiliation; product reviews are written in the context of a star rating; and blog posts are written in the context of blogger demographics.

Consider political discourse. Republicans and Democrats both discuss health care, immigration, jurisprudence, and any number of other issues. But their divergent perspectives on these issues lead them to discuss them in different ways. For instance, a Democrat writing about immigration might be more likely to raise questions about economics and social policy, while a Republican might be more likely to speak about border security or amnesty programs. Context shapes the way topics are discussed. We cannot expect to predict party affiliation or analyze its effect on patterns of discourse without accounting for that influence.

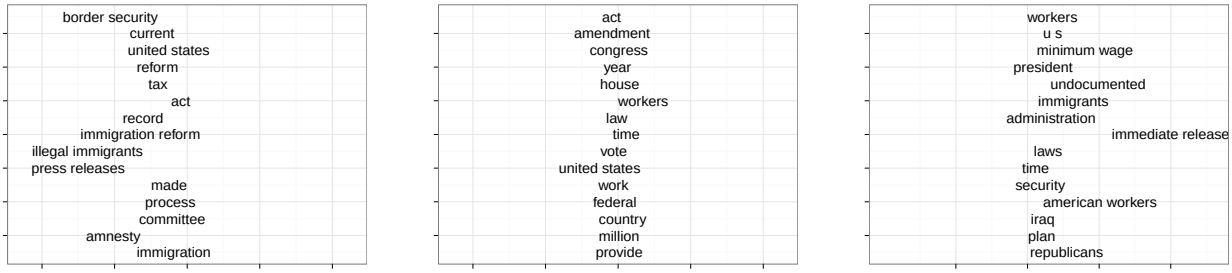
In this paper, we develop the inverse regression topic model (IRTM), a model that discovers and quantifies variation in topic expression. For example, Figure 1 illustrates what the model finds in a corpus of 10,000 political press releases. In the center, we show the most prevalent words in the 'neutral' form of a topic about immigration. On the left and right sides, we show words that are strongly associated with the topic as discussed by one or the other party. The difference between Republican terms ("illegal immigrants," "border security") and Democratic ones ("undocumented," "american workers") reflects the parties' differing stance on immigration.[1]

Our method builds on the multinomial inverse regression model (MNIR) (Taddy, 2013), a unigram model of text that uses the per-document context to distort a base distribution

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

[†]Work completed while the author was a student at Princeton University.

[1]We pursue this analysis further and explain how we constructed the plots in Section 3.3.



(a) Republican ($y < 0$) (b) Neutral ($y = 0$) (c) Democratic ($y > 0$)

Figure 1. A topic from the subsampled press release corpus, distorted in different ways. This topic corresponds to the legislative process itself, and, notably, immigration reform; the scoring function is that from (5). The neutral words are the top words in the base topic $\beta_k(0)$. Horizontal position indicates the value value of the distortion value Φ_w for the word (left being more negative).

over words. In the IRTM, each document exhibits several topics and each of them is distorted by the context. We found that the IRTM uncovers a rich structure for the exploration of topic variation, while simultaneously giving better predictions than the basic inverse regression model.

From a topic modeling perspective, the IRTM represents a different type of supervised topic model (Blei & McAuliffe, 2007). Previously, these have primarily focused on the relationship between metadata and the *choice* of topics in a text (Blei & McAuliffe, 2007; Lacoste-Julien et al., 2008; Mimno & McCallum, 2008).[2] Supervised topic models might detect that Republicans discuss climate change less than Democrats, but not how a Republican discussion of climate change is different from a Democratic discussion. We found that the IRTM improves over this approach, though combining the two perspectives on how metadata influences text is a promising area for future work.

Inverse regression models are difficult to fit, and Taddy's original algorithm exploited the simplicity of his unigram model. Unfortunately, the speedups he enjoys are not available when topics come into play. Thus, to fit our model, we designed two new variational algorithms for fitting inverse regression models: first, an efficient batch algorithm based on Taddy's original minorization strategy; and second, a faster online variant, based on stochastic subgradient descent and with sparse updates. These innovations could find application in other multinomial inverse regression settings.

The rest of this paper is organized as follows. In Section 2.1, we specify the assumptions of the IRTM. In Sections 2.2-2.3, we describe two methods for variational MAP estimation of its parameters and how to predict metadata from unlabeled documents. In Sections 3.1-3.2, we study the IRTM's predictive performance on several large corpora, demonstrating that it outperforms both the original MNIR, supervised topic modeling, and Dirichlet multinomial regression. Finally, in Section 3.3, we demonstrate how to use the IRTM to give a new exploratory window into text and its context.

[2]One notable exception is SAGE (Eisenstein et al., 2011), which does however assume discrete metadata. In our terminology, SAGE uses a separate distortion vector for each metadata class, making it unsuitable for continuous metadata (and less parsimonious than the IRTM in the case of discrete but ordered annotations).

2. Model Specification and Inference

In this section, we specify the inverse regression topic model (IRTM) and derive its corresponding inference algorithms. We then discuss how to form predictions from unlabeled documents.

2.1. The Inverse Regression Topic Model

Like latent Dirichlet allocation (Blei et al., 2003), the IRTM is a mixed-membership model comprising K topics $\beta_{1:K}$, each a distribution over terms. Documents are drawn by first choosing a distribution θ_d over topics, then allocating each word slot to one of the topics and drawing the word from that topic.

But whereas a topic in LDA is just a single distribution, in the IRTM it is a family of distributions. Each document is associated with *metadata* $y_d \in \mathbf{R}^M$, and its words are drawn from a document-specific set of topics derived by a distortion effect from the base corpus-wide topics. In the remainder of this paper, we take $M = 1$, but the generalization to $M > 1$ is straightforward. The variation in topic probabilities depends on the *distortion* $\Phi \in \mathbf{R}^W$, a vector mapping each word to a distortion value or weight, and the metadata y_d.

Concretely, let β_k be one of K *base topics*. For document

d, the probability of drawing word w from this topic in document d with metadata y_d is

$$\beta_{kw}(y_d) = \frac{\beta_{kw} \exp(\Phi_w y_d)}{C_k(y_d)}, \quad (1)$$

where $C_k(y_d) = \sum_v \beta_{kv} \exp(\Phi_v y)$ is the normalizer. For any given y_d value, the deviation from the base probabilities will be greatest for words with large values of $|\Phi_w|$.

We put a Laplace prior on the distortion, Dirichlet priors on the base topics, and Gaussian priors on the metadata. This leads to the following generative process:

1. For each base topic k, draw $\beta_k \sim \text{Dir}(\eta)$.
2. For each word w, draw the distortion $\Phi_w \sim \text{Lap}(\lambda)$.
3. For each document d
   (a) Draw topic proportions $\theta_d \sim \text{Dir}(\alpha)$.
   (b) Draw metadata $Y_d \sim \mathcal{N}(\mu, \sigma^2)$.
   (c) For each word
      i. Draw topic indicator $Z_{dn} \sim \theta_d$.
      ii. Draw word $W_{dn} \sim \beta_{z_{dn}}(y_d)$.

The graphical model is illustrated in Figure 2.

Although fully generative rather than discriminative, the IRTM is a generalization of Taddy's unigram inverse regression model of text (Taddy, 2013). We see below that our model is more powerful, though this power comes at the cost of more complex inference—particularly when large data sets are involved.

In closing, we note that our choice of a single distortion vector, shared across topics, was motivated by the negative results of preliminary experiments. These showed that topic-specific distortion vectors led to lower predictive accuracy and reduced interpretability. We believe this is essentially due to overfitting and that it might be possible (and useful) to permit topic-specific distortion vectors, provided further model structure linking them is also added.

2.2. MAP Inference

In training, we are given a collection of documents and their metadata. We obtain MAP estimates of the distortion and topics using a variational EM algorithm. We alternate between approximating the local variables—the topic proportions and topic assignments—and the corpus-wide parameters—the base topics and distortion.

We optimize a variational lower bound on the log-likelihood:

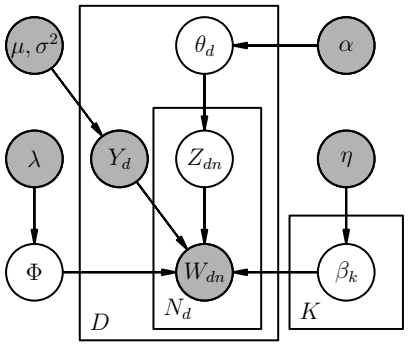


Figure 2. Graphical model for the inverse regression topic model. The metadata Y_d is observed during training, but not at test time.

$$\mathcal{L} = (\eta - 1) \sum_k \sum_w \log \beta_{kw} - \lambda ||\Phi||_1 + \sum_d \left((\alpha - 1) \sum_k \log \theta_{dk} + \ell_d(\boldsymbol{\beta}, \Phi) \right) + H(q), \quad (2)$$

where we have denoted the expected log-likelihood of a document given θ_d, $\boldsymbol{\beta}$, and Φ by

$$\ell_d(\boldsymbol{\beta}, \Phi) = \sum_w \sum_k n_d^w \gamma_{dwk} (\log \beta_{kw} + \Phi_w \cdot y_d - \log C_{dk}). \quad (3)$$

Here, $\gamma_{dwk} \approx p(Z_{dn} = k \mid W_{dn} = w)$. A maximum of this function gives an approximate MAP estimate of the model parameters.

To fit the per-document variables we use coordinate ascent on the topic proportions θ_d and variational topic assignments γ_{dwk}. We iterate the following updates until convergence:

$$\gamma_{dwk} \propto \theta_{dk} \beta_{kw}(y_d); \quad \theta_{dk} \propto \sum_w n_d^w \gamma_{dwk} + \alpha - 1.$$

This differs from variational inference for LDA only in that the distorted form of the topics appears in the update equation for the topic assignments. Note that we only compute γ values for words that appear in the document.

We optimize the model parameters in two ways. In batch inference, we adapt the minorization strategy of Taddy (2013), iteratively updating parameters by optimizing a Taylor expansion based lower bound on $\mathcal{L}$ in a small neighborhood of the current estimate. This algorithm is effective for moderately sized data sets, but cannot handle large corpora. To scale up, we design an online inference scheme based on stochastic subgradient descent, a different form of which has recently proved successful in topic modeling

(Hoffman et al., 2013). Our algorithm repeatedly subsamples the corpus, fits the document variables on the sample, and then updates the global parameters by following a noisy subgradient of the (negative) ELBO. To make this procedure more efficient, we introduce a second level of stochasticity that sparsifies the noisy subgradient before the update.

In our experiments, the batch algorithm typically took about 10 passes over the data, each of which requires $O(DWK)$ time; in contrast, the online algorithm required only about 200 iterations, each of which takes $\leq (1 + p)SWK$ time, where $0 < p \leq 1$ is a sparsity parameter that can be controlled by the user and S is the minibatch size. On some corpora, the resulting speedup was as great as a factor of 10. Inference is slower than for simpler topic models; this reflects the added complexity of the inverse regression setting.

In both algorithms, we initialize the topics $\boldsymbol{\beta}$ using an LDA inference algorithm,[3] and, in the stochastic case, we only optimize the distortion Φ holding topics fixed to the LDA estimate. The mathematical details of both algorithms are in the supplement.

2.3. Prediction

A natural application of the IRTM is to the prediction of the metadata y_d (e.g., a product rating) from document text. This can be done in a few ways. The most immediate—though also the least effective—is to follow the approach from Taddy (2013), based on the projected document representation $u_{\text{SRN}} = \frac{1}{N} \cdot (n \cdot \Phi)$, where n is the word count vector. The corresponding prediction is $\hat{y}_{\text{SRN}} = a u_{\text{SRN}} + b$ for scalars a and b that must be fit.

We can also predict y directly using MAP estimation. Given model parameters and a document, we choose $\hat{y}_{\text{MAP}}$ to maximize the variational objective

$$\mathcal{L}_{\text{pred}}(\theta,\ \gamma,\ y) = (\alpha - 1) \sum_k \log \theta_k - \frac{(y-\mu)^2}{2\sigma^2} + \sum_k \sum_w n^w \gamma_{wk} \Big(\log \beta_{kw} + \Phi_w y - \log C_k(y) \Big). \tag{4}$$

We fit y by coordinate ascent, iteratively optimizing γ_{wk}, θ_d, and y. In the supplement, we use the exponential family structure of $\beta_k(y)$ to show that the MAP estimate is unique and to show that MAP prediction chooses $\hat{y}_{\text{MAP}}$ to approximately match the model expected distortion $\left(\sum_k \theta_k \mathbf{E}_{W \sim \beta_k(\hat{y}_{\text{MAP}})}[\Phi_W]\right)$ and its empirical realization $\left(\frac{1}{N}\sum_w n^w \Phi_w\right)$

Notice, however, that the overall scale of MAP predictions is dictated by the degree of lexical variation; it does not necessarily match the scale of the metadata. Likewise, it can happen that MAP predictions systematically have nonzero mean—if, for instance, $\frac{1}{W}\sum_w \Phi_w \neq 0$. This difference is the analogue, in our case, of the scale difference between MNIR's reduced-dimensional document representation u_{SRN} and the metadata. Taddy (2013) resolves the difficulty by regressing metadata values onto the reduction. Following this lead, we improve on MAP estimation by predicting metadata with a linear model $y_{\text{MAP-LM}} = a\hat{y}_{\text{MAP}} + b$ for scalars a and b. This mechanism is an informal variant of MAP prediction in an extended model where an intermediate latent variable u_d dictates the distortion via $\Phi_w u_d$ and $y_d = a u_d + b$. As in Taddy (2013), we found that this adjustment lead to good empirical performance.

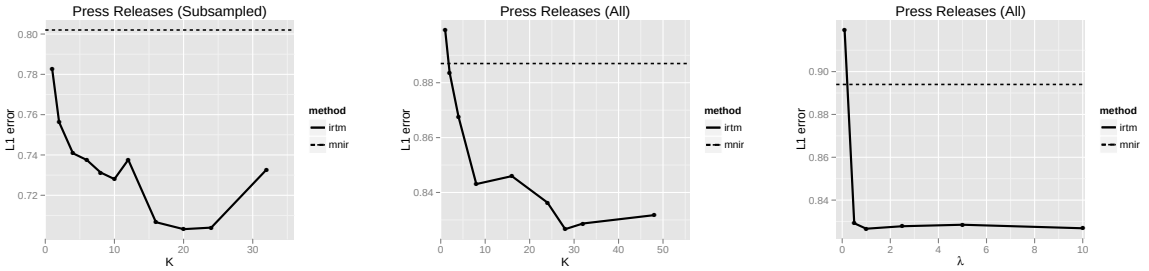
3. Empirical Study

We analyze our algorithms' performance on several collections of documents and their associated metadata. We first examine the predictive performance of the IRTM. We find that the IRTM outperforms MNIR (Taddy, 2013) and several supervised topic modeling approaches (for which open-source code is available): supervised latent Dirichlet allocation (sLDA, (Blei & McAuliffe, 2007)), LDA and regression (Blei & McAuliffe, 2007), and Dirichlet-multinomial regression (Mimno & McCallum, 2008). We find the best performance when we combine the IRTM with word-based features in an extended text regression model (Joshi et al., 2010). Finally, we demonstrate how to use the IRTM to explore patterns of word use in topics, and how they relate to metadata.

3.1. Data Sets, Preprocessing, and Training

We studied two kinds of documents—product reviews and Congressional press releases. The product reviews came from Amazon and Yelp, with metadata given by the star rating in $\{1, 2, 3, 4, 5\}$. The Amazon corpus contains 13,528 reviews from eleven categories (e.g. Books, Health & Personal Care). We used a vocabulary of 8,309 terms. The corpus was obtained from the raw Multi-Domain Sentiment Dataset (Blitzer et al., 2007). The Yelp corpus contains 152,280 reviews of numerous kinds of businesses, though restaurant reviews account for most of the data. For the full corpus, we used a vocabulary of 61,515 terms. In some experiments, we also work with a subset of 15,305 randomly selected reviews from the full corpus. The vocabulary for the subsampled corpus consists of 9,146 terms. All Yelp reviews came from the Yelp Academic Dataset (Yelp, Inc.,

[3] For the corpora we used, 100 iterations of collapsed Gibbs sampling sufficed. For even bigger corpora, faster algorithms can be used.



(a) Learning curve (vs. K) with $\alpha = 0.1,\ \eta = 0.01$, and $\lambda = 1.0$

(b) Learning curve (vs. K) with $\alpha = 0.1,\ \eta = 0.01$, and $\lambda = 1.0$

(c) Learning curve (vs. λ) with $\alpha = 0.1,\ \eta = 0.01$, and $K = 28$

Figure 3. The IRTM's error on the evaluation sets in the Press Release corpora, as a function of hyperparameter values. Predictive error using MNIR with its default $\mathrm{Ga}(1.0,\ 0.5)$ prior on λ is superimposed. In all cases, the IRTM outperforms MNIR and takes advantage of its ability to use multiple topics. Performance is highly robust to the choice of penalty λ, justifying our use of the default setting $\lambda = 1.0$ in further experiments.

Table 1. Comparison of the IRTM to MNIR and supervised topic models on the prediction task. The biggest gains from the IRTM are seen on the complex press release corpora, whose optimal numbers of topics were large ($K = 20,\ 24$, and 28, respectively); by comparison, the product review corpora benefited much less—and this is reflected in the lower number of topics ($K = 4, 4$, and 1). These K values, and those for the other topic models in the table, were chosen to minimize prediction error on an evaluation set. Default values of $\lambda = 1.0$, $\alpha = 0.1$, and $\eta = 0.01$ were used in all of these experiments; for MNIR, the default rate 0.5 and shape 1.0 for the gamma prior on the penalty was used. The missing values in the supervised LDA row correspond to corpora too large for the implementation of supervised LDA to handle.

Method	Test error (L_1)					
	Amazon	Press Releases (Subsampled)	Press Releases (Top Members)	Press Releases (All)	Yelp (Subsampled)	Yelp (All)
IRTM (This paper)	**0.996**	**0.703**	**0.420**	**0.826**	**0.741**	**0.704**
MNIR (Taddy, 2013)	1.03	0.802	0.597	0.894	0.765	0.721
LDA and regression (Blei & McAuliffe, 2007)	1.27	0.952	0.735	0.961	0.866	0.767
Supervised LDA (Blei & McAuliffe, 2007)	1.13	0.917	0.711	-	0.805	-
Dirichlet-multinomial regression (Mimno & McCallum, 2008)	1.25	0.978	0.915	0.970	0.853	0.850

2012).

The press releases were collected from United States Senators between 2005 and 2008. The metadata is the party affiliation of the source: Republican documents are coded as -1; Democratic documents are coded as 1; and Independent/No Affiliation documents are coded as 0. We looked at two subsets of this data. The "subsampled" subset contains 9,868 documents, chosen at random, and a vocabulary of 4,789 terms. The "top members" subset contains 13,023 documents from the most prolific legislators and a vocabulary of 4,818 terms. The full set contains 72,224 documents from more than 100 Senators, and uses a vocabulary of 19,882 terms. This data comes from Justin Grimmer's work on representational style (Grimmer, 2010).

The vocabulary was initially selected using χ^2-tests for collocations with a significance level of $p = 0.00001$. We then kept only the terms that occurred more than 50 times. For batch inference, convergence in training was defined as a change in the objective less than 0.1%. For online inference, the learning rate was chosen by optimizing evaluation error on a subsampled corpus and the number of iterations was set to optimize prediction error on an evaluation set drawn from the full corpus.

3.2. Predictive Performance

We first assess the predictive power of the IRTM, showing that the IRTM gives better predictions than MNIR and various supervised topic models. In these studies, we use the adjusted MAP estimate of y (Section 2.3), as this was the most effective strategy. Section 4 of the supplement com-

Table 2. Results of experiments comparing L_1-penalized text regression and the IRTM alone to a hybrid approach. Text regression is competitive with, and often better than, prediction based on generative models. Nonetheless, we find that adding a feature reduces test error on all our corpora—by as much as 12% (relative) compared to IRTM-only prediction or lasso-only prediction. An L_1 penalty was used for the Words, Words + $\hat{y}_{\text{MAP}}$, and Words + θ feature sets; the other regressions were unpenalized.

Features	Test error (L_1)					
	Amazon	Press Releases (Subsampled)	Press Releases (Top Members)	Press Releases (All)	Yelp (Subsampled)	Yelp (All)
Words	1.03	0.703	0.384	0.711	0.753	0.696
θ	1.27	0.952	0.718	0.961	0.866	0.767
$\hat{y}_{\text{MAP}}$	0.996	0.703	0.420	0.826	0.741	0.704
Words + $\hat{y}_{\text{MAP}}$	**0.976**	**0.647**	**0.337**	**0.703**	**0.718**	**0.642**
Words + θ	1.03	0.705	0.383	0.711	0.739	0.661
θ + $\hat{y}_{\text{MAP}}$	0.986	0.709	0.417	0.825	0.727	0.679

pares the test error of this approach to those of direct MAP estimation and prediction in the manner of (Taddy, 2013). We demonstrate that the IRTM's predictive performance is robust to changes in the distortion regularizer λ, which obviates the empirical Bayes approach taken in Taddy (2013). Finally, we explain how we can use the IRTM prediction in a text regression on word counts (Joshi et al., 2010), demonstrating that adding the reduced-dimension representation obtained from IRTM as a (single) additional feature by as much as 12%.

Figure 3 illustrates the results of model selection, exploring the number of topics K and the regularization λ. In all cases, we fit to a training set (80% of the data), optimized the number of topics on an evaluation set (10%), and assessed on a test set (10%). These plots show that the IRTM's success is due to its use of topics; the best results on the complex press release corpora, for instance, came with $K \geq 20$. Further, the Press Releases (All) learning curve versus λ shows that a default value of $\lambda = 1.0$ fares about as well as the best choice of λ. Results with the other corpora were similar.

Table 1 shows prediction errors on the gold-standard test set and compares them to those of several models: MNIR (Taddy, 2013), supervised LDA (Blei & McAuliffe, 2007), LDA and regression (Blei & McAuliffe, 2007), and Dirichlet-Multinomial regression (Mimno & McCallum, 2008). The table shows that IRTM performed best on every corpus, improving on MNIR by as much as 29.6%. The gains over MNIR are most pronounced on the complex press release corpora, which require large numbers of topics to model optimally ($K \geq 20$ in our experiments); our model's edge on the more homogeneous Yelp corpora is considerably less. This supports our intuition that topics are important for modeling context-specific distortion in heterogeneous corpora. Likewise, the IRTM's improvement over the supervised topic models confirms that variation in the manner of topic expression is more informative for sentiment and opinion analysis than the expression levels of topics (as manifested in the mixing distribution θ_d).

In pure prediction applications, text regression—where word counts are covariates in a regularized regression—is a popular alternative to topic-based methods (Joshi et al., 2010). Table 2 compares L_1-penalized (i.e. lasso) text regression to regression onto reduced-dimension representations obtained from LDA (θ) and the IRTM ($\hat{y}_{\text{MAP}}$), as well as to regressions that combine these features. Using the IRTM significantly improves prediction on every corpus: adding $\hat{y}_{\text{MAP}}$ to the feature vector of word counts produces drops in error as large as 12% (relative). In addition, in these fits, we found that the lasso puts most of its weight on the $\hat{y}_{\text{MAP}}$ feature; it uses the words (i.e., the usual features in text regression) to slightly correct errors associated with prediction through $\hat{y}_{\text{MAP}}$.

3.3. Exploring Topic Variation

We now show how to discover variation in the expression of fixed background topics by combining the topics and distortion to identify the most metadata-dependent terms. In this section, we concentrate on the subsampled and full press release corpora.

We base our analysis on the following scoring function:

$$\text{score}(k,\ w) = \beta_{kw}\Phi_w. \tag{5}$$

This score is the contribution of word w to the expected weight $\mathbf{E}_{V\sim\beta_k}[\Phi_V]$ under topic distribution β_k. It makes sense as a measure of a word's sentiment content because each word influences $\hat{y}_{\text{MAP}}$ for a document through $\mathbf{E}_{V\sim\beta_k(y)}[\Phi_V]$ (see Section 2.3). When we wish to find words in topic k most characteristic of documents with positive metadata, we can look at the words with most positive scores. Likewise, when we wish to find words in topic k most characteristic of negative metadata documents, we can look at the most negative-scoring words. This approach consistently gives revealing results supported by actual variation in word usage. Figures 1 and 4 illustrate

this for two topics. In both cases, Republican and Democratic words are obtained as the lowest- and highest-scoring words, respectively, according to (5). Neutral words are just the highest probability words in the base topic. The continuity between the representations shows this combination is reasonable. The supplement discusses exploration via direct examination of the families $\beta_k(y)$ and explains the shortcomings of that approach.

Table 3. High- (Democratic) and low-scoring (Republican) words in the immigration topic, shown with the proportion of occurrences in documents from the respective parties. Proportions were based on LDA pseudocounts and defined as $\frac{\text{count}^+}{\text{count}^y}$ (Democratic) and $1 - \frac{\text{count}^+}{\text{count}^y}$, respectively, per equations (6) and (7).

Democratic Terms	% Democratic In-Topic	Republican Terms	% Republican In-Topic
workers	72%	border security	77%
minimum wage	87%	reform	54%
undocumented	80%	immigration reform	63%
immigrants	75%	illegal immigrants	77%
security	62%	amnesty	70%
american workers	87%	immigration	51%
rights	71%	borders	56%
wages	92%	visa	63%
voting rights act	80%		
families	78%		
health care	69%		

We now examine differences in the discourse of Democrats ($y = 1$) and Republicans ($y = -1$) on issues of immigration reform, which falls under a broader topic about the legislative process itself.[4] As shown in Figure 1, high scoring words tended to concern employers, workers, immigrants' rights, and economic matters, while low scoring words were more likely to refer to border security. Putting ourselves in the position of a practitioner, we might then search among the high-scoring words for some that appear especially relevant to immigration and group these into thematic categories. This leads to the 11 Democratic and 8 Republican words shown in Table 3 (the proportions are explained below), which can be decomposed into a few different semantically coherent categories as follows. On the Democratic side, emergent themes include employment ("workers," "minimum wage," "american workers," "wages"), immigrants ("workers," "undocumented," "immigrants," "families"), welfare and rights ("minimum wage," "rights," "wages," "voting rights act," "families," "health care"), and, to a lesser degree, border control ("security"). On the Republican side, these are instead border control ("border security," "amnesty," "borders," and "visa"), immigrants ("illegal immigrants"), and immigration reform ("reform," "immigration reform"). All of these groupings represent possible lines of investigation for a political scientist.

Based on these breakdowns, we might expect that border security dominates Republican discourse, while Democrats are more concerned with the economic and social condition and impact of immigrants. The difference between "undocumented"[5] and "illegal immigrants" as labels also suggests differing attitudes toward the people involved.

Such speculations require validation to become conclusions. Even before asking whether the differing vocabularies reflect differing attitudes, we need to show that the vocabularies do differ. Ideally, we would do this by seeing how often each word occurs in each topic in Democratic documents and in Republican documents. Since we do not observe topic labels, however, we must settle for expected counts

$$\text{count}^y(k,\ w) = \sum_d n_d^w \gamma_{dwk} \tag{6}$$

Restricting to Democratic documents ($y_d = 1$) gives us

$$\text{count}^+(k,\ w) = \text{count}^y(k,\ w)\big|_{d:\ y_d=1} = \sum_{d:\ y_d=1} n_d^w \gamma_{dwk}. \tag{7}$$

Based on this, we define the proportion $\frac{\text{count}^+}{\text{count}^y}$ of Democratic uses of a word in a topic and the analogous Republican proportion $1 - \frac{\text{count}^+}{\text{count}^y}$. These proportions, converted to percentages, are the metrics we use in Table 3. Essentially by definition, these metrics capture the overall and Democratic word frequencies in each LDA topic.[6] As Table 3 shows, they confirm that the words that scored highly according to (5) are those that are much more prevalent in Democratic discourse on the topic— likewise, those with low scores are more prevalent in Republican discourse.

Anyone seeking a deeper understanding of the views on immigration policy articulated in this corpus would still need to go deeper than this; ultimately, it would be necessary to look at individual documents somehow believed to be representative of Republican and Democratic views on immigration, respectively. The IRTM facilitates the retrieval of such documents through a scoring function:

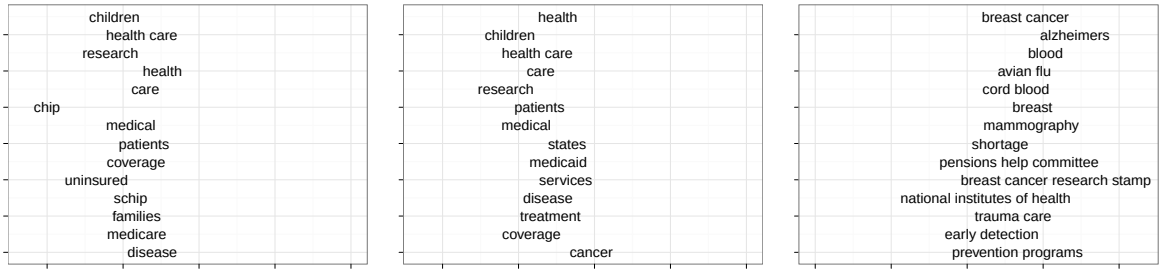
$$\text{score}(k,\ d) = \frac{\sum_w n_d^w \gamma_{dwk} \Phi_w}{\sum_w n_d^w \gamma_{dwk}}. \tag{8}$$

The highest- and lowest-scoring documents will be those with greatest expected weight in the topic–and, thus, representative of the ends of the spectrum of document sentiment in that topic. Table 4 shows representative phrases on immigration reform from the two sides of the aisle, considering only documents with $\theta_{dk} \geq 0.8$ in this topic.

[4] Immigration likely falls under this topic in the corpus because of the extensive discussion of the Dream Act in the final years of the Bush presidency.

[5] Followed, in 145 of 178 instances, by one of "immigrants" (69), "workers" (29), "aliens" (12), "immigration" (10), "persons" (5), "students" (4), "people" (4), "population" (3), "residents" (2), "children" (2), "individual" (2), "immigrant" (2), and "relatives" (1).

[6] Note that these metrics' values depend on the properties of LDA, but not on our specific model of how LDA topics are distorted.



(a) Republican ($y < 0$) (b) Neutral ($y = 0$) (c) Democratic ($y > 0$)

Figure 4. Republican (low-scoring) and Democratic (high-scoring) words in a topic family from the full press release corpus corresponding to medicine and health care using the score (5). The neutral words are the top words in the base topic $\beta_k(0)$. Horizontal position indicates Φ value (left being more negative).

Table 4. Representative phrases from press releases on immigration policy. Sentences were chosen from among the top 25 most positive (Democratic) and most negative (Republican) documents according to score (8), subject to the requirement that $\theta_{dk} \geq 0.8$ in the immigration topic.

Democrats	Republicans
...includes extensive labor protections for the temporary **workers** ...	It is in our national security interest to secure our **borders** ....
Millions of **undocumented immigrants** persist in the shadows ...	I have always opposed **amnesty** ....
...muster the political will to pass comprehensive reform that protects our **security**, bolsters our economy, and preserves America's tradition as a nation of **immigrants**	...praised the passage of an amendment he sponsored to the emergency supplemental spending that would beef up funding for **border security** ....
...shouldn't be making criminals out of hardworking **families**.	...strengthens **border security** by increasing border patrol ....
holding employers accountable if they hire illegal **immigrants** and dealing with the 12 million **undocumented immigrants** a way that is practical and fair to **american workers** and taxpayers	...the vulnerability of our inadequately protected **borders** and the need to allow an earned path to citizenship ....
We should not punish **undocumented** children who were brought to this country	...temporary **visa** workers ....

4. Conclusion and Future Work

We have described the inverse regression topic model (IRTM), a new model combining the strengths of topic modeling with those of the recently-proposed multinomial inverse regression (MNIR) framework. The IRTM extends MNIR by accounting for the heterogeneity of text corpora, and gains expressive power by capturing the way context distorts topic *expression*. This is in contrast to previous supervised topic models, which exclusively model the influence of context on the relative *prevalence* of different topics in a document (Blei & McAuliffe, 2007; Lacoste-Julien et al., 2008; Mimno & McCallum, 2008).

To estimate parameters for these models, we introduced an efficient variational MAP inference algorithm. We then developed a fast online inference algorithm with sparse updates based on stochastic subgradient descent.

We applied both algorithms to several text corpora, including 150K Yelp reviews and 73K Congressional press releases. Comparison to other models on the prediction task consistently demonstrated the advantages of our approach. The large gains over MNIR illustrated the benefit of adding topics, while improvements over supervised topic models provide evidence that variation in topic expression is more informative for opinion analysis than variation in topic prevalence.

Finally, we analyzed the divergence in discourse on immigration reform. This showed how our model can be used to uncover topic-specific variation in word usage and indicated how this might be useful in exploratory data analysis.

These results reveal the benefits of modeling variation in topic expression. Making inference in the IRTM even more scalable, extending the IRTM to account for metadata's influence on topic prevalence, and adding a forward regression component would all be natural next steps in the exploration of this methodology. Additionally, analyzing the exploratory power of the IRTM in detail would be an interesting applied extension of our work.

Acknowledgements. The authors would like to thank David Mimno, Rajesh Ranganath, and the anonymous reviewers for helpful feedback. MR is supported in part by a Cambridge Overseas Trust Scholarship. DMB is supported by NSF IIS-0745520, IIS-1247664, IIS-1009542, ONR N00014-11-1-0651, DARPA FA8750-14-2-0009, and the Alfred P. Sloan foundation.

References

Blei, David M. and McAuliffe, Jon D. Supervised topic models. In *Advances in Neural Information Processing Systems (NIPS) 20*, 2007.

Blei, David M., Ng, Andrew Y., and Jordan, Michael I. Latent Dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, 3:993–1022, March 2003. ISSN 1532-4435.

Blitzer, John, Dredze, Mark, and Pereira, Fernando. Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 187–205, 2007.

Eisenstein, Jacob, Ahmed, Amr, and Xing, Eric P. Sparse additive generative models of text. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pp. 1041–1048, 2011.

Grimmer, Justin. Replication data for: The central role of communication in representation. `http://hdl.handle.net/1902.1/14596`, 2010.

Hoffman, Matthew D., Blei, David M., Wang, Chong, and Paisley, John. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Joshi, Mahesh, Das, Dipanjan, Gimpel, Kevin, and Smith, Noah A. Movie reviews and revenues: An experiment in text regression. In *2010 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 293–296, 2010.

Lacoste-Julien, Simon, Sha, Fei, and Jordan, Michael I. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Advances in Neural Information Processing Systems (NIPS) 21*, pp. 897–904, 2008.

Mimno, David M. and McCallum, Andrew. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2008.

Taddy, Matthew A. Multinomial inverse regression for text analysis. *Journal of the American Statistical Association (JASA)*, 108(3):755–770, 2013.

Yelp, Inc. Yelp Academic Dataset. `http://www.yelp.com/academic_dataset`, 2012.

The Inverse Regression Topic Model (Supplement)

This supplement includes brief elaborations on the main paper that may be of interest to some readers. In Section 1, we explain the minorization procedure underlying MAP inference. In Section 2, we lay out the details of our stochastic subgradient approximation procedure for online MAP inference. In Section 3, we lay out a useful interpretation of MAP prediction. In Section 4, we summarize the results of experiments with the two other prediction methods for the IRTM (MAP and sufficient reduction based) mentioned in the main paper. In Section 5, we discuss exploration of topic variation via the topic families $\beta_k(y)$ themselves and explain why we found it inadequate.

1. Minorization Scheme

Our goal in minorization is to maximize a lower bound on the objective $\mathcal{L}$ of equation (2). This maximization is done separately for the topics $\boldsymbol{\beta}$ and Φ, and distinct lower bounds are produced for each. For concreteness, we focus on Φ; the procedure for $\boldsymbol{\beta}$ is entirely analogous. As in the main paper, we assume real-valued metadata $y_d \in \mathbf{R}$. The general case is a straightforward extension.

In coordinate-wise minorization, the lower bounds, valid in a neighborhood of the current estimate $\Phi^{(0)}$, come from second-order Taylor expansion; they take the form

$$\tilde{Q}_w(\Phi_w) = \ell(\boldsymbol{\beta},\ \Phi^{(0)}) + \frac{\partial \ell}{\partial \Phi_w}(\boldsymbol{\beta},\ \Phi^{(0)})(\Phi_w - \Phi_w^{(0)}) + \frac{1}{2} H_w (\Phi_w - \Phi_w^{(0)})^2 - \lambda |\Phi_w|,$$

where

$$H_w = -\sum_d \sum_k \sum_v \left[n_d^v \gamma_{dvk} \quad \times \sup_{|\Phi_w - \Phi_w^{(0)}| \le \delta_w} \beta_{kw}(y_d)(1 - \beta_{kw}(y_d)) \right] \tag{1}$$

is a lower bound on the second derivative of ℓ with respect to Φ_w valid for $|\Phi_w - \Phi_w^{(0)}| \le \delta_w$. Since $\tilde{Q}_w(\Phi_w^{(0)}) = \mathcal{L}(\boldsymbol{\beta},\ \Phi^{(0)})$, it is easy to see that in fact $\tilde{Q}_w \le \mathcal{L}$ for $|\Phi_w - \Phi_w^{(0)}| \le \delta_w$ as a function of Φ_w with all other parameters held fixed.

At the end of this section, we explain how to compute H_w explicitly using techniques from Genkin et al. (2007).

This means that, if $|\Phi'_w - \Phi_w^{(0)}| \le \delta_w$ has $\tilde{Q}(\Phi'_w) \ge \tilde{Q}(\Phi_w^{(0)})$, and if Φ is obtained from $\Phi^{(0)}$ by setting $\Phi_w = \Phi'_w$, then

$$\mathcal{L}(\boldsymbol{\beta},\ \Phi) \ge \tilde{Q}_w(\Phi'_w) \ge \tilde{Q}_w(\Phi_w^{(0)}) = \mathcal{L}(\boldsymbol{\beta},\ \Phi^{(0)}),$$

so any update to Φ_w that stays within the δ_w-neighborhood of $\Phi_w^{(0)}$ and increases $\tilde{Q}_w$ also increases $\mathcal{L}$. Taking advantage of this, we use coordinate ascent updates of the form

$$\begin{aligned} &\Phi_w \leftarrow \operatorname{argmax}_{\phi \in A_w} \tilde{Q}_w(\phi), \quad \text{where} \\ &A_w = [-\delta_w,\ \delta_w], \quad \text{if } \Phi_w^{(0)} = 0 \\ &A_w = \{\phi \in \mathbf{R} \colon |\phi - \Phi_w^{(0)}| \le \delta_w,\ \operatorname{sgn}(\phi)\operatorname{sgn}(\Phi_w^{(0)}) \ge 0\}, \\ &\qquad \text{otherwise.} \end{aligned}$$

In other words, each update either maximizes $\tilde{Q}_w$ over the whole δ_w neighborhood of $\Phi_w^{(0)}$ (if $\Phi_w^{(0)} = 0$ or $|\Phi_w^{(0)}| \ge \delta_w$), or maximizes the lower bound over a truncated version of the neighborhood cut off so as to remain on the same side of 0 as $\Phi_w^{(0)}$. An analogous update applies to $\log \beta_{kw}$, albeit without truncation. We point out that, in fact, truncation does not appear strictly necessary for this algorithm to succeed, though it does seem natural in light of the choice of the sparsity-inducing Laplace prior: the mechanism that produces sparsity is precisely the difficulty of escaping from the critical point (of non-differentiability) at 0.

A naive implementation of this algorithm would update the Φ_w and β_{kw} sequentially. This is impractical, however, as it requires recomputation of the log-normalizer $C_k(y_d)$ for every topic-document pair after each update, with the result that updating the weights costs $\Omega(DWK)$ time.

We therefore adopt a lazy updating strategy that computes all the new Φ values before updating them, then computes all the new $\boldsymbol{\beta}$ values before updating them. Essentially, this approach amounts to a non-coordinate-wise minorization algorithm. Indeed, if $\boldsymbol{H} - \boldsymbol{\nabla^2}_{\Phi} \ell(\boldsymbol{\beta},\ \Phi)$ is positive semidefinite on $\prod_{w,\ m} A_w$,

$$\begin{aligned} \ell(\boldsymbol{\beta}, \Phi) &\ge \ell(\boldsymbol{\beta}, \Phi^{(0)}) + \nabla \ell(\boldsymbol{\beta}, \Phi^{(0)})^T (\Phi - \Phi^{(0)}) \\ &\qquad + \frac{1}{2} (\Phi - \Phi^{(0)})^T \boldsymbol{H} (\Phi - \Phi^{(0)}) \\ &=: Q(\boldsymbol{\beta}, \Phi), \quad \Phi \in \prod_w A_w. \end{aligned} \tag{2}$$

This implies $\mathcal{L} \geq \tilde{Q} \colon = Q - \lambda||\Phi||_1$ on the product neighborhood.

Unfortunately, optimizing this function directly is infeasible, so we replace $\boldsymbol{H}$ by a diagonal matrix $\boldsymbol{D} = \text{diag}(H_w)$, where H_w is given by (3). This results in updates of the form prescribed above but whose independence of each other allows for lazy updating.[1] Mathematically, the approximation makes sense in this context because $H_{v,w} = \sum_d \sum_k O\left(y_d^2 \beta_{kv}(y_d)\beta_{kw}(y_d)\right)$ if $v \neq w$, while $H_{w,w} = -\sum_d \sum_k \Omega\left(y_d^2 \beta_{kw}(y_d)(1-\beta_{kw}(y_d))\right)$. This means that, in the typical case when $\beta_{kw}(y_d) \ll 1$ for all k, w, and d, the off-diagonal entries of the lower bound on the Hessian are much smaller than the diagonal terms. Empirically, we find that the optimization scheme resulting from this approximation runs quickly and performs parameter estimation effectively.

We now give an explicit value for H_w using estimates similar to those of Genkin et al. (2007) and Taddy (2013). Begin by noting

$$\frac{\partial \ell}{\partial \Phi_w} = \sum_d \sum_k \left(n_d^w \gamma_{dwk} - \sum_v n_d^v \gamma_{dvk} \cdot \beta_{kw}(y_d) \right) \cdot y_d$$

and

$$\frac{\partial^2 \ell}{\partial \Phi_w^2} = -\sum_d \left(\sum_v n_d^v \gamma_{dvk} \right) \cdot \beta_{kw}(y_d)(1-\beta_{kw}(y_d)) y_d^2,$$

and letting

$$H_w \colon = -\sum_d y_d^2 \sum_k \left[\left(\sum_v n_d^v \gamma_{dvk} \right) \times \sup_{|\Phi_w - \Phi_w^{(0)}| \leq \delta} \beta_{kw}(y_d)(1-\beta_{kw}(y_d)) \right]. \quad (3)$$

[1] In the general case of $y_d \in \mathbf{R}^M$, we would replace by $\boldsymbol{H}$ by a *block-diagonal* matrix $\boldsymbol{D}$ instead, where each block would have dimensions $M \times M$.

We can compute these suprema exactly:

$$\begin{aligned}
&2 + \frac{\sum_{v \neq w} \beta_{kv} \exp(\Phi_v^{(0)} \cdot y_d)}{\beta_{kw} \exp(\Phi_w^{(0)} \cdot y_d + \Delta\Phi_w \cdot y_d)} + \frac{\beta_{kw} \exp(\Phi_w^{(0)} \cdot y_d + \Delta\Phi_w \cdot y_d)}{\sum_{v \neq w} \beta_{kv} \exp(\Phi_v^{(0)} \cdot y_d)} \\
&= 2 + \frac{\left(\sum_{v \neq w} \beta_{kv} \exp(\Phi_v^{(0)} \cdot y_d)\right)^2}{\beta_{kw} \exp(\Phi_w^{(0)} \cdot y_d + \Delta\Phi_w \cdot y_d) \sum_{v \neq w} \beta_{kv} \exp(\Phi_v^{(0)} \cdot y_d)} + \frac{\left(\beta_{kw} \exp(\Phi_w^{(0)} \cdot y_d + \Delta\Phi_w \cdot y_d)\right)^2}{\beta_{kw} \exp(\Phi_w^{(0)} \cdot y_d + \Delta\Phi_w \cdot y_d) \sum_{v \neq w} \beta_{kv} \exp(\Phi_v^{(0)} \cdot y_d)} \\
&= \frac{\left(\sum_{v \neq w} \beta_{kv} \exp(\Phi_v^{(0)} \cdot y_d) + \beta_{kw} \exp((\Phi_w^{(0)} + \Delta\Phi_w) \cdot y_d\right)^2}{\beta_{kw} \exp((\Phi_w^{(0)} + \Delta\Phi_w) \cdot y_d) \sum_{v \neq w} \beta_{kv} \exp(\Phi_v^{(0)} \cdot y_d)} \\
&= \frac{1}{\beta_{kw}(y_d)(1-\beta_{kw}(y_d))},
\end{aligned}$$

where $\beta(y_d)$ is formed at $\Phi_w = \Phi_w^{(0)} + \Delta\Phi_w$. Since the first expression in this chain has the form

$$2 + \frac{1}{ax} + ax,$$

where $a = \frac{\beta_{kw} \exp(\Phi_w^{(0)} \cdot y_d)}{\sum_{v \neq w} \beta_{kv} \exp(\Phi_v^{(0)} \cdot y_d)}$ and $x = \exp(\Delta\Phi_w y_d)$, its minimum, hence the maximum (supremum) of $\beta_{kw}(y_d)(1-\beta_{kw}(y_d))$, is attained at $x = \frac{1}{a}$ or, equivalently, when $\beta_{kw} \exp(\Phi_w \cdot y_d) = \sum_{v \neq w} \beta_{kv} \exp(\Phi_v^{(0)} \cdot y_d)$. This may not always be attainable with $|\Delta\Phi_w| \leq \delta$, so we end up with the bound

$$\begin{aligned}
F_{dwk} \colon &= \inf_{|\Delta\Phi_w| \leq \delta} \frac{1}{\beta_{kw}(y_d)(1-\beta_{kw}(y_d))} \\
&= 2 + \frac{f_{dw}}{\sum_{v \neq w} \beta_{kv} \exp(\Phi_v^{(0)} \cdot y_d)} + \frac{\sum_{v \neq w} \beta_{kv} \exp(\Phi_v^{(0)} \cdot y_d)}{f_{dw}},
\end{aligned}$$

where

$$\begin{aligned}
f_{dwk} &= \exp(\Phi_w \cdot y_d + \delta|y_d|), \\
&\quad \text{if } \exp(\Phi_w \cdot y_d + \delta|y_d|) < \sum_{v \neq w} \beta_{kv} \exp(\Phi_v^{(0)} \cdot y_d); \\
f_{dwk} &= \exp(\Phi_w \cdot y_d - \delta|y_d|), \\
&\quad \text{if } \exp(\Phi_w \cdot y_d - \delta|y_d|) > \sum_{v \neq w} \beta_{kv} \exp(\Phi_v^{(0)} \cdot y_d); \\
f_{dwk} &= \sum_{v \neq w} \beta_{kv} \exp(\Phi_v^{(0)} \cdot y_d), \\
&\quad \text{otherwise.}
\end{aligned}$$

Finally, we compute H_w exactly as

$$H_w = -\sum_d y_d^2 \cdot \sum_k \frac{\sum_v n_d^v \gamma_{dvk}}{F_{dwk}}. \tag{4}$$

2. Stochastic Subgradient Descent Scheme

We now describe our stochastic subgradient descent (SSGD) scheme for online MAP inference. As noted in the paper, this method is for fitting the distortion matrix Φ, with the topics $\boldsymbol{\beta}$ held fixed.

In this setting, we wish to minimize the negative ELBO, given up to constants independent of Φ, by

$$\begin{aligned}\mathcal{M} = \sum_d \Bigg[&- \sum_k \sum_w n_d^w \gamma_{dwk} \log \beta_{kw} \\ &- \sum_w n_d^w \Phi_w \cdot y_d \\ &+ \sum_k \left(\sum_w n_d^w \gamma_{dwk} \right) \log C_k(y_d) \Bigg] \\ &- (\eta - 1) \sum_k \sum_w \log \beta_{kw} + \lambda \left|\left|\Phi\right|\right|_1 .\end{aligned}$$

Switching to $\mathcal{M}$ allows us to frame our algorithm in the standard terms of convex optimization—in particular, to work with the subdifferential $\partial_\Phi \mathcal{M}(\Phi)$ rather than the 'superdifferential' needed for maximization.

Our stochastic approximation is based on a two-tier sampling approach. First, we sample a minibatch $B \subset [D]$ of documents and form the approximate objective

$$\begin{aligned}\hat{\mathcal{M}}_B = -\frac{D}{S} \cdot \sum_{d \in B} \Bigg[&\sum_k \sum_w n_d^w \gamma_{dwk} \log \beta_{kw} \\ &+ \sum_w n_d^w \Phi_w \cdot y_d \\ &- \sum_k \left(\sum_w n_d^w \gamma_{dwk} \right) \log C_k(y_d) \Bigg] \\ &- (\eta - 1) \sum_k \sum_w \log \beta_{kw} + \lambda \left|\left|\Phi\right|\right| .\end{aligned} \tag{5}$$

We then choose a subgradient $g \in \partial_\Phi \hat{\mathcal{M}}$ and replace it in turn by a sparse approximation $\hat{g}$. To compute $\hat{g}$, we first sample a minibatch $B' \subset [W]$ of terms and define $V_{\text{seen}} = \{w \in [W] \colon \sum_{d \in B} n_d^w > 0\}$. The sparse approximate subgradient is then given by

$$\hat{g}_{wm} = \begin{cases} g_{wm} & \text{if } w \in V_{\text{seen}} \\ \frac{W}{S'} \cdot g_{wm} & \text{if } w \in B' \cap [W] \setminus V_{\text{seen}} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Since $p(w \in B' \mid w \in V_{\text{unseen}}) = \frac{S'}{W}$, we see that $\mathbf{E}_{B'}[\hat{g}] = g$. Further, any mapping $g \colon B \mapsto g(B) \in \partial_\Phi \hat{\mathcal{M}}_B$ necessarily satisfies $\mathbf{E}_B[g] \in \partial_\Phi \mathcal{M}_B$, so $\mathbf{E}_{B,B'}[\hat{g}] \in \partial_\Phi \mathcal{M}_B$, as required for SSGD.

As usual in stochastic optimization, we maintain an estimate $\Phi^{(t)}$ and update it iteratively, letting $t \to \infty$. An individual update has three stages:

1. Sample a minibatch of documents $B^{(t)} \subset [D]$ of size S and a minibatch of terms $B^{\prime,(t)} \subset [W]$ of size S'.
2. Choose a subgradient $g^{(t)} \in \partial_\Phi \hat{\mathcal{M}}(\Phi^{(t)})$ and compute the stochastic approximation $\hat{g}^{(t)}$.
3. Update $\Phi^{(t+1)} = \Phi^{(t)} - \epsilon^{(t)} \hat{g}^{(t)}$, where $\epsilon^{(t)}$ is the current step size.

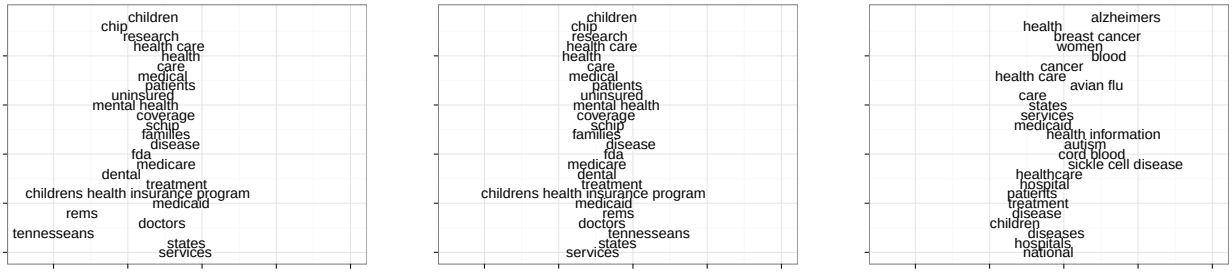
The first stage is carried out by repeatedly sampling without replacement; the second and third, on the other hand, require further elucidation. In the second stage, for each $w \in [W]$ and $1 \le m \le M$, we set

$$\begin{aligned}g^{(t)}_{w,\,\text{main}} = \frac{D}{S} \Bigg[&\sum_{d \in B^{(t)}} n_d^w y_d \\ &- \sum_d \sum_k \left(\sum_v n_d^v \gamma_{dvk} \right) \beta_{kw} (y_d) y_d \Bigg]\end{aligned} \tag{7}$$

and

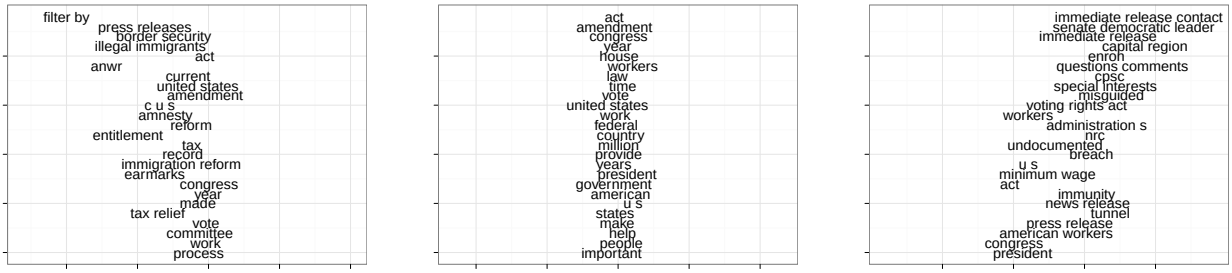
$$g_w^{(t)} = \begin{cases} g^{(t)}_{w,\,\text{main}} + \lambda \cdot \operatorname{sgn}(\Phi_w^{(t)}) & \text{if } \Phi_w^{(t)} \neq 0, \\ g^{(t)}_{w,\,\text{main}} - \lambda & \text{o.w. if } g^{(t)}_{w,\,\text{main}} > \lambda, \\ g^{(t)}_{w,\,\text{main}} + \lambda & \text{o.w. if } g^{(t)}_{w,\,\text{main}} < -\lambda, \\ 0 & \text{otherwise.} \end{cases}$$

In words, each component of the subgradient is either just the derivative in the appropriate direction ($\Phi_w \neq 0$), chosen to point in the same direction as the main term $g^{(t)}_{w,\,\text{main}}$ ($\Phi_w = 0$ and $|g^{(t)}_{w,\,\text{main}}| > \lambda$), or set to zero if 0 is a subgradient in dimension w ($\Phi_w = 0$ and $|g^{(t)}_{w,\,\text{main}}| \le \lambda$). After computing $g^{(t)}$, we use (6) to compute $\hat{g}^{(t)}$. Note that an actual implementation should compute g_w only for $w \in B'$. We also point out that, while this scheme does not itself have a provable rate of convergence, a simple modification using projections to a ball of radius R after each step and outputting averaged iterates $\hat{\Phi}^{(t)} = \frac{1}{\sum_{\tau=1}^t \epsilon^{(\tau)}} \cdot \sum_{\tau=1}^t \epsilon^{(\tau)} \Phi^{(\tau)}$ can easily be proven to converge (Polyak, 1987; Shor, 1998).



(a) Republican ($y = -4$) (b) Neutral ($y = 0$) (c) Democratic ($y = 5$)

Figure 1. Top words in $\beta_k(y)$ for $y \in \{-4,\ 0,\ 5\}$ in the topic family corresponding to medicine and health care. We obtained these results using the full press release corpus. Color and horizontal position indicates Φ value (red and left are more negative).



(a) Republican ($y = -4$) (b) Neutral ($y = 0$) (c) Democratic ($y = 4$)

Figure 2. Top words in $\beta_k(y)$ for $y \in \{-4,\ 0,\ 5\}$ in the topic family corresponding to immigration. We obtained these results using the subsampled press release corpus. Color and horizontal position indicates Φ value (red and left are more negative).

3. MAP Prediction

We show that $\log C_k(y)$ is convex in y.

Proposition 3.1. *In the usual notation, we have*

$$\frac{\partial \log C_k(y)}{\partial y} = \mathbf{E}_{W\sim\beta_k(y)}[\Phi_W]$$

and

$$\frac{\partial^2 \log C_k(y)}{\partial y^2} = \mathbf{E}_{W\sim\beta_k(y)}[\Phi_W^2] - \mathbf{E}_{W\sim\beta_k(y)}[\Phi_W]^2.$$

In particular, $\log C_k(y)$ *is convex in* y*.*

Proof. We show that $\beta_k(y)$ is an exponential family with natural parameter $y \in \mathbf{R}$. Indeed, we see that

$$p(w \mid \beta_k,\ \Phi,\ y) = \beta_{kw}(y) = \beta_{kw} \exp\left(y \cdot \Phi_w - \log C_k(y)\right).$$

Thus, if $t(w) = \Phi_w \in \mathbf{R}$, $h(w) = \beta_{kw}$, and $a(y) = \log C_k(y)$,

$$p(w \mid \beta_k,\ \Phi,\ y) = \exp\left(y \cdot t(w) - a(y)\right) h(w),$$

proving that $p(w \mid \beta_k,\ \Phi,\ y)$ for fixed β_k and Φ is an exponential family with parameter $y \in \mathbf{R}$.

Now, by the usual exponential family identity (see, e.g., Lehmann & Casella (1998)),

$$\frac{\partial a(y)}{\partial y} = \mathbf{E}_{W\sim\beta_k(y)}\left[t(W)\right]$$

and

$$\frac{\partial^2 a(y)}{\partial y^2} = \mathbf{E}_{W\sim\beta_k(y)}\left[t(W)^2\right] - \mathbf{E}_{W\sim\beta_k(y)}\left[t(W)\right]^2$$

Since $a(y) = \log C_k(y)$ and $t(w) = \Phi_w$, the equalities follow. Now, $\mathbf{E}_{W\sim\beta_k(y)}\left[\Phi_W^2\right] \geq \mathbf{E}_{W\sim\beta_k(y)}\left[t(W)\right]$ by Jensen's inequality, so convexity follows. □

Since $\mathcal{L}_{\text{pred}}$ is a negative linear combination of terms $\log C_k(y)$ plus the strictly concave penalty $-\frac{1}{2\sigma^2}(y-\mu)^2$, Proposition 3.1 shows that $\mathcal{L}_{\text{pred}}$ is strictly concave in y.

It likewise allows a simple probabilistic interpretation of MAP prediction in the IRTM. Indeed, if β^{emp} denotes a document's empirical word distribution, the proposition immediately implies

$$\frac{\partial \mathcal{L}}{\partial y} = -\frac{1}{\sigma^2}(y-\mu) + N \cdot \left(\mathbf{E}_{w\sim\beta^{\text{emp}}}[\Phi_w] - \sum_k \left(\frac{\sum_w n^w \gamma_{wk}}{N}\right) \mathbf{E}_{w\sim\beta_k(y)}[\Phi_w]\right).$$

After letting $\tilde{\theta}_k = \frac{\sum_w n^w \gamma_{dwk}}{N}$ and $\tilde{\beta}^{\text{mod}}(y) = \sum_k \tilde{\theta}_k \beta_k(y)$, we then find that the MAP estimate

Table 1. Though not as effective as our primary method, direct MAP estimation often still outperforms MNIR and the supervised topic models, whereas sufficient reduction based prediction is considerably less competitive. The Primary column lists the error when using the IRTM prediction method from the main paper.

Test error (L_1)	Method		
	MAP	Suff. Red.	Primary
Amazon	0.989	1.03	0.996
Press Releases (Subsampled)	0.777	0.756	0.703
Press Releases (Top Members)	0.437	0.524	0.420
Press Releases (All)	0.924	0.901	0.826
Yelp (Subset)	0.751	0.766	0.741
Yelp (All)	0.705	0.734	0.704

$\hat{y}_{\text{MAP}}(\theta, \gamma)$ given θ and γ satisfies

$$\mathbf{E}_{W \sim \tilde{\beta}^{\text{mod}}(\hat{y}_{\text{MAP}}(\theta,\gamma))}[\Phi_W] = \mathbf{E}_{W \sim \beta^{\text{emp}}}[\Phi_W] - \frac{1}{N\sigma^2}(\hat{y}_{\text{MAP}}(\theta, \gamma) - \mu). \tag{8}$$

Note that, at optimality, $\tilde{\beta}^{\text{mod}} \approx \beta^{\text{mod}} := \sum_k \theta_k \beta_k(y)$, since $\tilde{\theta} \approx \theta$; further, if N is large, the penalty term is dominated by the empirical distortion vector term. This means that, intuitively, the model picks $\hat{y}_{\text{MAP}}$ to bring its expected distortion vector $\mathbf{E}_{W \sim \beta^{\text{mod}}(\hat{y}_{\text{MAP}})}[\Phi_W]$ as close to the empirical distortion vector as possible, up to adjustments due to the prior and the variational approximation.

4. Alternate IRTM Prediction Methods

Section 2.3 of the main paper discussed two methods of prediction with the IRTM that fare worse than our chosen adjusted MAP strategy: first, prediction via a regression onto the sufficient reduction $u_{\text{SRN}} = \frac{1}{N} \cdot \sum_w n^w \Phi_w$, as for MNIR in Taddy (2013); second, direct MAP prediction. For completeness, we show the results of these methods on the test sets. Though not as effective as our primary method, direct MAP estimation often still outperforms MNIR and the supervised topic models, whereas sufficient reduction based prediction is considerably less competitive. Table 1 summarizes the results.

5. Exploration through Topic Families

Rather than using the scoring function of the main paper, we can attempt to explore corpora by examining the most probable words in $\beta_k(y)$ for varying y values. Figure 1 illustrates this approach. There, the topic corresponds to medicine and health care, and the varying high- probability words already suggest interesting biases in Republican and Democratic discourse on those subjects. We might guess, for example, that Democrats discuss breast cancer and Alzheimer's research much more than Republicans do and that, obversely, Republicans prioritize childrens' health care in their discourse, at least in the large press release corpus. In this case, both of these guesses turn out to be correct.

Unfortunately, examination of the top topic words often does not yield such illuminating patterns; Figure 2 shows an example of how things can go wrong. The problem is twofold. First, when y is small $(-1,\ 1)$, the most likely words in the distorted topic strongly resemble those in the base topic. Second, as y becomes larger $(-4,\ 4)$, the words at the top tend to become those with high (positive or negative) weight, and these may have no relation to the specific topic. Words both strongly associated with the topic *and* highly variable in prevalence depending on party affiliation appear interleaved with others that are simply likely in the topic or prone to sentiment-dependent variability but not strongly associated with the topic. Moreover, the most variable words need not be the most common, so that deep examination of the topic is necessary to unearth them. It is worth noting that these problems appear most pronounced on the smaller corpora, suggesting that this approach to topic exploration might be much more effective on big data sets than on small ones.

References

Genkin, Alexander, Lewis, David D., and Madigan, David. Large-scale Bayesian logistic regression for text categorization. *Technometrics*, 49:291–304(14), 2007.

Lehmann, Erich L. and Casella, George. *Theory of Point Estimation (Springer Texts in Statistics)*. Springer, 2nd edition, 1998.

Polyak, Boris. *Introduction to Optimization*. Optimization Software, Inc., 1987.

Shor, Naum Z. *Nondifferentiable Optimization and Polynomial Problems*. Springer, 1998.

Taddy, Matthew A. Multinomial inverse regression for text analysis. *Journal of the American Statistical Association (JASA)*, 2013. To appear.