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

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

where

$$H_w = -\sum_d \sum_k \sum_v \Bigg[n_d^v \gamma_{dvk} \quad \times \qquad\qquad (1)$$

$$\sup_{|\Phi_w - \Phi_w^{(0)}| \le \delta_w} \beta_{kw}(y_d)(1 - \beta_{kw}(y_d)) \Bigg]$$

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.

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

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

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 $\prod_{w,\ m} A_w$,

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

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

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 $\boldsymbol{D} = \mathrm{diag}(H_w)$, 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

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

We can compute these suprema exactly:

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

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

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

where

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

[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 $M \times M$.

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 $B^{',(t)} \subset [W]$ 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.

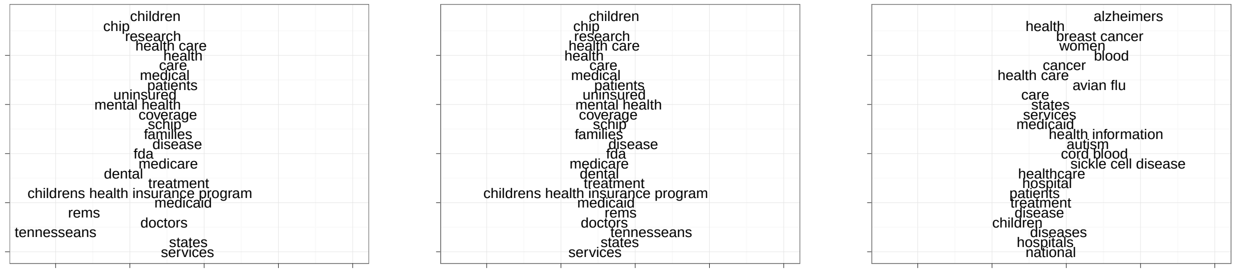
The first stage is carried out by repeatedly sampling without replacement; the second and third, on the other hand, require further elucidation. In the second stage, for each $w \in [W]$ and $1 \leq m \leq M$, we set

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

and

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

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 $g_{w,\text{ main}}^{(t)}$ ($\Phi_w = 0$ and $|g_{w,\text{ main}}^{(t)}| > \lambda$), or set to zero if 0 is a subgradient in dimension w ($\Phi_w = 0$ and $|g_{w,\text{ main}}^{(t)}| \leq \lambda$). 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).

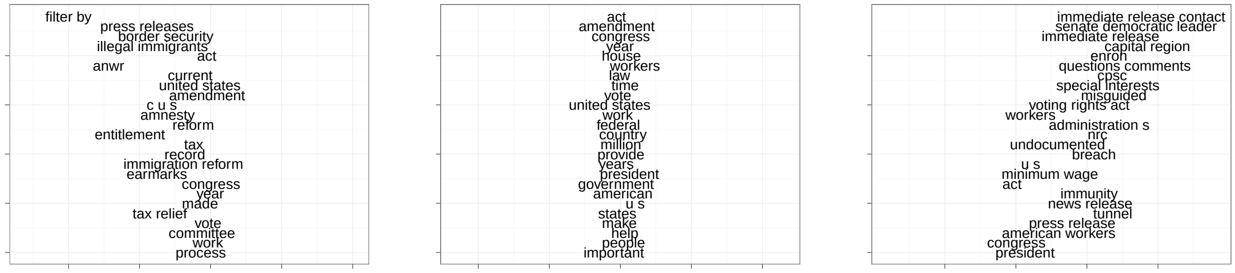


(a) Republican $(y = -4)$

(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

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

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

Note that, at optimality, $\tilde{\beta}^{\mathrm{mod}} \approx \beta^{\mathrm{mod}} :\ = \sum_k \theta_k \beta_k(y)$, 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 $\hat{y}_{\mathrm{MAP}}$ to bring its expected distortion vector $\mathbf{E}_{W \sim \beta^{\mathrm{mod}}(\hat{y}_{\mathrm{MAP}})}[\Phi_W]$ 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 $u_{\mathrm{SRN}} = \frac{1}{N} \cdot \sum_w n^w \Phi_w$, 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.