2013

Regularization of Latent Variable Models to Obtain Sparsity

Ramnath Balasubramanyan
Carnegie Mellon University

William W. Cohen
Carnegie Mellon University, wcohen@cs.cmu.edu

Published In
Proceedings of the 2013 SIAM International Conference on Data Mining, 414-422.

Regularization of Latent Variable Models to Obtain Sparsity

Ramnath Balasubramanyan
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
rbalasub@cs.cmu.edu

William W. Cohen
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213
wcohen@cs.cmu.edu

Abstract

We present a *pseudo-observed* variable based regularization technique for latent variable mixed-membership models that provides a mechanism to impose preferences on the characteristics of aggregate functions of latent and observed variables. The regularization framework is used to regularize topic models, which are latent variable mixed membership models for language modeling. In many domains, documents and words often exhibit only a *slight* degree of mixed-membership behavior that is inadequately modeled by topic models which are overly liberal in permitting mixed-membership behavior. The regularization introduced in the paper is used to control the degree of polysemy of words permitted by topic models and to prefer sparsity in topic distributions of documents in a manner that is much more flexible than permitted by modification of priors. The utility of the regularization in exploiting sentiment-indicative features is evaluated internally using document perplexity and externally by using the models to predict star counts in movie and product reviews based on the content of the reviews. Results of our experiments show that using the regularization to finely control the behavior of topic models leads to better perplexity and lower mean squared error rates in the star-prediction task.

1 Introduction

Latent variable mixed membership models such as topic models [1] and mixed membership stochastic block models [2] allow words in a language model or nodes in a graph to participate in different roles each time they are observed in data. We propose a regularization technique to impose preferences on the nature of the distribution over latent roles in such models. In this approach, we extend the model to include a noisy copy of an aggregate function over latent variables, such as the entropy of the topic distributions. By pretending to see a desired value (*pseudo-observe*) for the copy the model is coaxed to push the variables that participate in the aggregate functions to values that make the pseudo-observed variables likely. More specifically, the generative model posits that the pseudo-observed variables are distributed as Gaussians, which are in part parameterized by the aggregate functions of latent variables. This approach therefore permits a simple way to specify preferences about interactions among variables that are iid in an unregularized model, without requiring the need for complicated priors. The approach also has the advantage of keeping posterior inference in the model simple; we need only extend the Gibbs sampler used for approximate inference in latent variable models, by adding a few additional terms, which do not increase the computational order of complexity of inference.

Topic models [1] have become a widely used tool to uncover structure in text corpora. The structure that is uncovered is useful in several ways. For instance, posterior topic distributions from topic models can be used for data exploration to get a big picture of the nature of discourse in the corpus. Topic models are also useful in reducing dimensionality while representing documents. Instead of representing documents as bag of words which lead to high-dimensional sparse representations, using topic distributions provides a low-dimensional dense representation of documents. The low-dimensional representation of documents is often used as input for text regression and classification tasks where the reduced number of input features is beneficial. Blei et al.[3] proposed a model that learns the lower dimensionality projection and regression/classification jointly through their model termed Supervised LDA (SLDA). In this paper, we use the SLDA framework to predict the number of stars assigned to movie and product reviews based on the text in the review.

Topic models permit words to exhibit *polysemy* by allowing the same word to be generated by different topic distributions albeit with different probabilities. While this flexibility is useful and necessary for model-

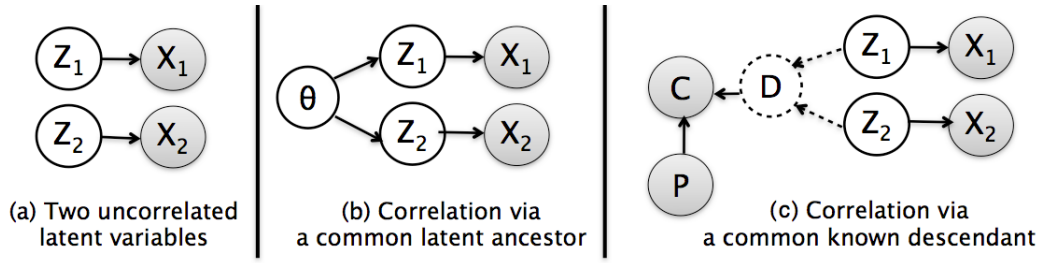


Figure 1: Different ways to impose a correlation between latent variables.

ing natural language, the model often permits excessive freedom for words to be generated by multiple topics. We use the regularization technique introduced in the paper to directly control the freedom of words to be generated by multiple topics by penalizing high entropy in the *topic distributions of words*, which reduces the degree of polysemy permitted by the model. Our experiments show that restricting the ability of words to span topics provides better performance both in terms of perplexity and in terms of predicting the star score of reviews in all the datasets studied. This is because the proposed regularization enables the model to better exploit sentiment-indicative words by restricting the latent topics they can occur under.

Topic models uncover hidden structures in corpora by positing that documents have a distribution over topics that are manifested in the documents and by representing topics as distributions over words in the vocabulary. In practice, it is often seen that posterior topic distributions in documents tend to be sparse. While the naturally emergent sparsity is sufficient for dimensionality reduction and language modeling practices in larger sized documents, it is often not the case with shorter documents. We therefore use the pseudo-observation based regularization technique to control the sparsity of the topic distributions directly. Our experiments show that reducing the entropy i.e. making the topic distributions sparser helps achieve better perplexity and lowers error rates in all the datasets in the star prediction task.

The rest of the paper is organized as follows. Section 2 describes the regularization framework. Sections 3 and 4 introduce the word entropy and document topic distribution entropy regularization techniques. Related work is discussed in Section 5 followed by the conclusion.

2 Entropic Regularization in Latent Variable Models

In constructing probabilistic models, it is often the case that we wish to have the freedom to assign latent roles to observed variables, but we nonetheless anticipate certain regularities among those latent assignments. For instance, in topic modeling, we typically assign latent roles to every occurrence of a polysemous word like *bank*, thus allowing it to take part in different topics (e.g., *finance* and *fishing*); however, we still anticipate that most words will have few meanings, and hence that the set of latent topics associated with most words will be small. In short, we hope that our models will show certain types of correlations between the latent topic assignments—correlations that lead to a moderate amount of polysemy.

Generally, a modeler has two options to ensure correlations between latent variables: including a common latent ancestor of the latent nodes, or including a common observed descendant. Figure 1 shows a very simple example of this modeling choice, in the case of a Bayes network with four binary variables: for the sake of concreteness, suppose that Z_1 and Z_2 represent some genetic property, and X_1 and X_2 encode an observable phenotype (say baldness). In part (a) of the figure the latent variables Z_1 and Z_2 are independent. In part (b) the variables are dependent, by virtue of the new node θ.

The model of (b) is quite useful if θ has a plausible causal interpretation—here, e.g., X_1 and X_2 might be observations of two brothers, θ genetic properties of their father, and $P(Z_i|\theta)$ a model of inheritance. In some settings, however, the causal interpretation is either poorly understood or difficult to model, but information is available about correlation. For instance, we may know that one in four men is bald, but that one pair of brothers in eight is also bald (rather than the expected one in 16). We can model this information conveniently by explicitly introducing variables that measure such correlations, and then using priors to express our preferences over them. As shown in Figure 1 part (c), we can specify D to be true iff both X_1 and X_2 are true and let C be a noisy copy of D whose expectation is encoded by the known variable P (which acts as a prior) and $P(C|P)$. (In this case, D is a deterministic function of its parents, so it is

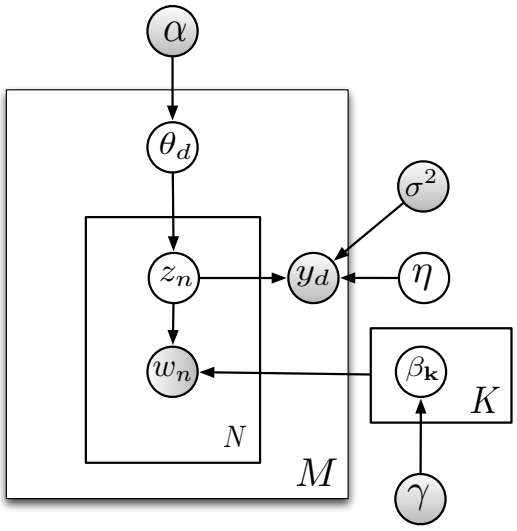


Figure 2: Supervised LDA

conceptually awkward to introduce a prior directly on D). Approach (c) also makes Z_1 and Z_2 dependent, and in this simple case, it is easy to show that the same set of probability distributions can be modeled by the networks in (b) and (c); however, the approach has different implications computationally, as we will argue below.

Modeling choice (c) can be generalized as follows: given an existing generative model, we can bias it by introducing new variables that, like D, measure some aggregate property of the latent variables, and then imposing a prior preference over this aggregate value. One computationally convenient way to introduce such a prior is to allow the aggregate node D to be latent, but introduce a new node C that is a "noisy copy" of C, with a specified noise model.

We will show that we can use this technique to encourage lower entropy in the latent topic assignments to a particular word, or to encourage lower entropy in the latent role assignments to nodes in a mixed-membership stochastic block model. Perhaps surprisingly, these bias variables can be introduced with minimal computational change: for instance, they can be added to topic models without impacting conjugacy. Importantly, they can be conveniently added to modify aggregates that are never explicitly sampled, and hence are difficult to predictably modify by simply changing the priors of a model.

3 Word Distribution Regularization

The LDA model and its extensions allow the same word to belong to different topics when they are instantiated multiple times in the corpus. This freedom is essential in modeling *polysemy*. While this freedom is useful, it is often necessary and advantageous to control the freedom. Following the idea illustrated in Figure 3, we present a entropy based regularization technique based on pseudo-observed variables, that directly controls the freedom of words to take on different latent topics, by penalizing high entropies in their topic distributions. In the review analysis task addressed in this paper, the regularization enables us to exploit sentiment-indicative words more efficiently. We apply this regularization to a Supervised LDA model (Figure 2). The SLDA model extends the Latent Dirichlet Allocation (LDA) model by incorporating response variables in documents that are dependent on the their topic proportions.

The inference procedure learns parameters for the topic model and the regression model in a joint fashion. The generative procedure to obtain a corpus D in this model is as follows:

1. Draw K topics $\beta_k \sim \text{Dir}(\gamma)$, for $k \in 1, \ldots, K$.
2. For every document $d \in D$
   - Draw topic proportions $\theta_d \sim \text{Dir}(\alpha)$.
   - For each word w_i in document $d, i \in 1, \ldots, N_d$:
     - Sample topic $z_i \sim \text{Mult}(\theta_d)$.
     - Sample word $w_i \sim \text{Mult}(\beta_{z_i})$.
   - Draw response variable $y_d \sim \mathcal{N}(\boldsymbol{\eta}.\overline{\mathbf{z}_d}, \sigma^2)$.

$\overline{\mathbf{z}_d}$ represents the observed topic proportion distribution of the document and is given by $\overline{z_d}^{(k)} = \frac{\sum_{i=1}^{N_d} \mathbf{I}(z_i=k)+\alpha}{N_d+K\alpha}$. The likelihood function for the model is

(3.1)
$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{y}, \mathbf{z}, \mathbf{w}|\alpha, \gamma, \boldsymbol{\eta}, \sigma^2) = \prod_{k=1}^{K} \text{Dir}(\beta_k|\gamma) \cdot \left(\prod_{d=1}^{D} \text{Dir}(\theta_d|\alpha) \left(\prod_{i=1}^{N_d} \theta_d^{(z_i)} \beta_{z_i}^{(w_i)} \right) \mathcal{N}(y_d|\boldsymbol{\eta}.\overline{\mathbf{z}_d}, \sigma^2) \right)$$

Since exact inference in the model is intractable, we use a collapsed Gibbs sampler for approximate inference, that collapses θ_d and $\boldsymbol{\beta}$. To sample a topic indicator z_i for word i in a document d, we use the expression

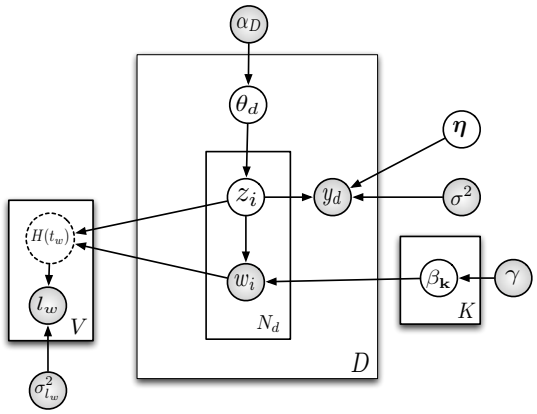


Figure 3: Supervised LDA with Word Distribution Regularization.

$$p(z_i = k|w_i, \mathbf{z}^{\neg \mathbf{i}}, \mathbf{w}^{\neg \mathbf{i}}, y_d, \alpha, \gamma, \boldsymbol{\eta}, \sigma^2) \propto$$
$$(n_{dk}^{\neg i} + \alpha)\frac{n_{kw_i}^{\neg i} + \gamma}{\sum_{w' \in V} n_{kw'}^{\neg i} + |V|\gamma}\exp\left(\frac{-(y_d - \boldsymbol{\eta}.\overline{\mathbf{z}_d})^2}{2\sigma^2}\right)$$

V refers to the set of all words i.e. the vocabulary. The n's are counts of observations in the training set. n_{kw} denotes the number of times a word w is observed under topic k, and n_{dk} the number of words with topic k in document d.

The regression parameters $\boldsymbol{\eta}$ are then obtained by training a linear regression model that uses $\overline{\mathbf{z}_d}$ as input features and y_d as the target. The inference procedure therefore alternates between sampling z_i for all words in the documents and training the linear regression model to obtain estimates for the $\boldsymbol{\eta}$, the regression parameters.

For the star-prediction task discussed later in the paper, reviews are represented as documents in the LDA framework and their star-ratings are represented as the real valued targets $\mathbf{y}$. Therefore, the model jointly learns a LDA like model and a regressor that maps topic proportion distributions of documents to their star-ratings.

3.1 Word Distribution Regularization The topic distribution of a word w belonging to V (i.e. the vocabulary) in a topic model can be defined as

$$t_w^{(k)} = \frac{n_{kw}}{\sum_{k'} n_{k'w}}, k \in 1, \ldots, K \tag{3.2}$$

The Shannon entropy of this distribution can be computed using the expression $H(t_w) = -\sum_k t_w^{(k)} \log_2(t_w^{(k)})$.

We now introduce word topic distribution entropy regularization by adding pseudo-observed variables, l_w (Figure 3), one for each word in V, which are noisy copies of $H(t_w)$. These noisy copies are drawn from Gaussians with mean $H(t_w)$ and variance $\sigma^2_{l_w}$, which is a hyperparameter to the model. The addition of these terms, which are set to 0, penalizes large entropies in the topic distributions of words, with $\sigma^2_{l_w}$ dictating the strictness of the penalty. The penalization of large entropies therefore drives the inference procedure to return models that exhibit lower entropies in their word distributions.

The joint distribution of the model with regularization is defined as:

(3.3)

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w}, \mathbf{y}, \mathbf{l_w}|\alpha, \gamma, \boldsymbol{\eta}, \sigma^2, \sigma^2_{l_w}) =$$
$$\prod_{k=1}^{K} \mathrm{Dir}(\beta_k|\gamma)\left(\prod_{d=1}^{D} \mathrm{Dir}(\theta_d|\alpha)\left(\prod_{i=1}^{N_d} \theta_d^{(z_i)}\beta_{z_i}^{(w_i)}\right)\right) \times$$
$$\prod_{w \in V} \exp\frac{-\left(l_w - H(t_w)\right)^2}{2\sigma^2_{l_w}}$$

The Gibbs sampling equation to sample a topic indicator for a word w_i for the regularized model is now given by

(3.4)

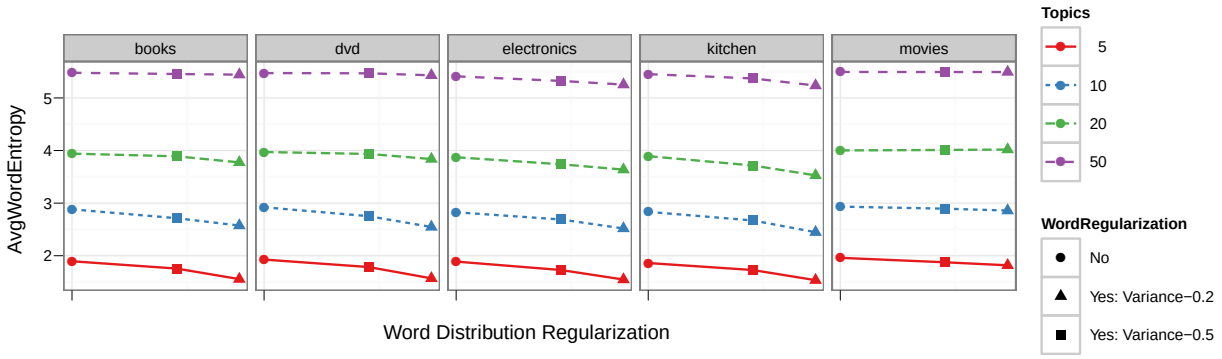
$$p(z_i = k|\mathbf{l_w}, w_i, \mathbf{z}^{\neg \mathbf{i}}, \mathbf{w}^{\neg \mathbf{i}}, y_d, \alpha, \gamma, \boldsymbol{\eta}, \sigma^2, \sigma^2_{l_w}) \propto$$
$$(n_{dk}^{\neg i} + \alpha)\frac{n_{kw_i}^{\neg i} + \gamma}{\sum_{w'} n_{kw'}^{\neg i} + |V|\gamma}\exp\left(\frac{-(y_d - \boldsymbol{\eta}.\overline{\mathbf{z}_d})^2}{2\sigma^2}\right) \times$$
$$\exp\left(\frac{-(H(t_{w_i}) - l_{w_i})^2}{2\sigma^2_{l_w}}\right)$$

During the Gibbs sampling process, the inference procedure tends to push the mean of the Gaussians i.e. $H(t_w)$ close to the *pseudo-observed* l_w values. We therefore set (*pseudo-observe*) l_w to be 0 to coax the inference procedure to return low entropy topic distributions for observed tokens. The variance parameter $\sigma^2_{l_w}$ can be used to adjust the tightness of the Gaussian to permit more or less entropy in the label distributions.

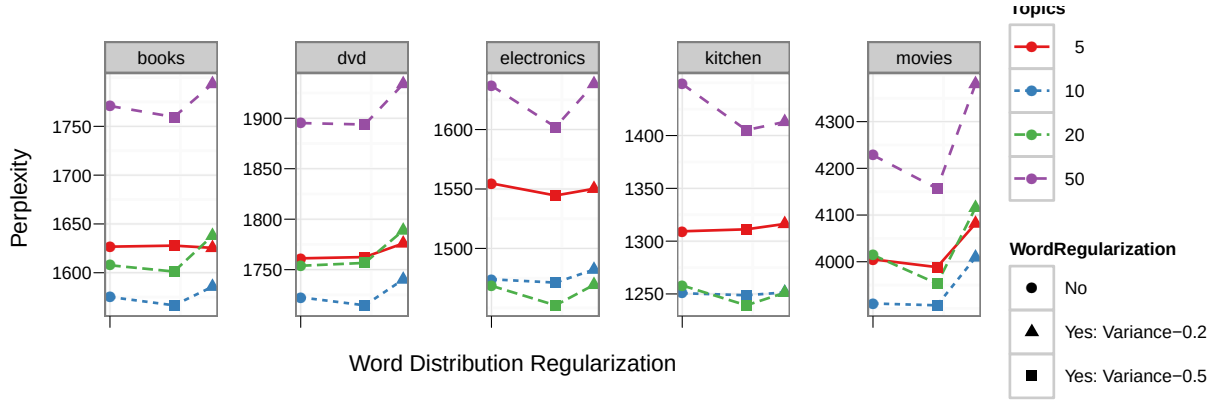
It should be noted that an alternate method to achieve sparsity is to modify the priors. Replacing the Dirichlet priors to obtain preferences in word distribution characteristics however requires complicated priors (which can no longer be Dirichlet) that are capable of producing topic distributions that are not iid, leading to complications in the inference procedure. The regularization technique described achieves the same objective while requiring minimal additions to the existing Gibbs sampling inference procedure.

Dataset	books	dvd	kitchen	electronics	movies
Size	5501	5118	5149	5901	2000
Vocabulary	13743	14548	10857	8377	12639

Table 1: Dataset statistics.



(a) Average Word Distribution Entropy



(b) Perplexity

Figure 4: Studying the effect of Word Distribution Regularization.

3.2 Datasets We test the performance of the regularized topic model by modeling star-rating annotated reviews of products from Amazon.com. The datasets provided by Blitzer et al. [4] consist of 4 sets of reviews corresponding to reviews about products in the books, dvd, electronics and kitchen categories. Each review is annotated with a star-rating that ranges from 0 to 5. We use only the text of reviews as inputs in this study. We also test the model on a movie reviews dataset [5] that contains 2000 reviews labeled as positive or negative, which are tagged with targets 1 and 0 respectively in our experiments. Statistics about the datasets are shown in Table 1.

3.3 Experiments We first investigate the effect of the proposed regularization on the entropy of words' topic distributions by studying the change in the word topic distribution entropy averaged across all words in the vocabulary i.e. $\sum_{w \in V} H(t_w)/|V|$. Figure 4(a) shows the change in the average word topic entropy for the datasets described above as the regularization is applied and increasingly tightened by decreasing the value of the variance parameter $\sigma^2_{l_w}$. The different lines in the plots indicate results of runs with different values of K (the number of topics). The first point in each plot indicates results with no regularization applied. The second and third points in each line show the entropy values with the variance value set to 0.5 and 0.2 respectively. Firstly, it can be seen that the models with higher K

Dataset	books	dvd	kitchen	electronics	movies
No regularization	2.169	2.042	1.986	1.885	0.218
With regularization	**2.136**	**2.003**	**1.761**	**1.874**	**0.208**

Table 2: Effect of Word Distribution Regularization on MSE in star-rating prediction.

Dataset	books	dvd	kitchen	electronics	movies
SVM	1.643	1.773	1.506	1.732	0.158
SVM + LDA (unregularized) features	1.614	1.757	1.373	**1.456**	0.145
SVM + LDA (with regularization) features	**1.592**	**1.748**	**1.356**	**1.456**	**0.142**

Table 3: MSE Scores: SVM text regression with topic modeling features.

show higher absolute entropy values since words have a greater number of topic indicators that they can appear under in. All the plots show a general downward trend in the average word entropy as regularization is applied and the regularization hyperparameter variance is decreased. This is explained due to the heavier penalization of higher entropies with lower variance that leads to lower average word topic entropies.

Next, we study the effect of regularization on document perplexity (shown in Figure 4(b)). It can be observed that the plots exhibit a "U-shape" i.e. the perplexity values dip below the value of the unregularized model when the variance value is set to 0.5 but rise again when the regularization is further tightened and set to 0.2. This indicates that there is a "sweet spot" for the regularization, which is $\sigma_{l_w}^2 = 0.5$ in the case of the datasets studied here: the polysemy freedom afforded by the unregularized model is overly expressive and setting the variance to a low value such as 0.2 reduces the freedom to a level where it is insufficient to represent the inherent polysemy in the corpus.

Finally, we test the performance of the model in predicting the star-ratings of reviews. We evaluate the star-rating predictions of models using mean-squared error (MSE) which is the square of the difference between the true and predicted ratings. Table 2 shows the mean squared error of the star-rating predictions for the different datasets. The table compares the error rates of the regularized models to the error rates of a baseline unregularized model. All results are obtained using 10-fold cross validation. The number of topics K is set to 20. For the regularized model, we use use a variance values of 0.5 for $\sigma_{l_w}^2$ which shows the best results in the perplexity plot. It can be seen from the table that using word topic regularization consistently significantly improves the MSE in star-prediction. On average this improvement in error rate is 3.98%. This improvement is statistically significant at the 95% level.

We also run experiments where we add the topic distributions as additional features to the original bag of words vector representation and train a SVM regression model. We therefore study the effect of the topic distributions obtained with and without regularization in such a setting. We first evaluate the SVM regression model with a linear kernel as a baseline model using 10-fold cross validation for the star-rating prediction task. We then add the topic proportion distributions of documents obtained from a LDA model trained with 50 topics as additional features to the SVM model. The slack parameter value was set at 0.05 based on cross-validation. The results of the experiments are shown in Table 3. It can be seen from the results that adding topic model features helps in improving SVM regression performance and adding regularization further improves performance in all the datasets.

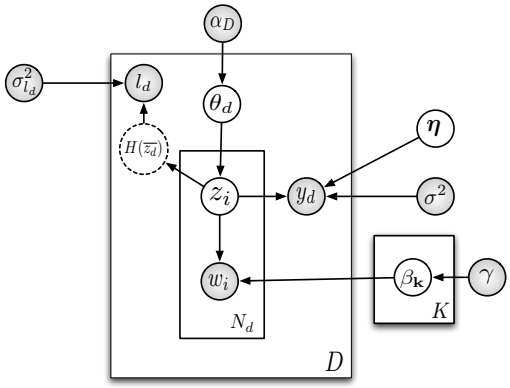


Figure 5: Supervised LDA with Document Topic Regularization.

4 Document Topic Proportion Regularization

In models based on LDA, the topic proportions of documents are typically drawn from a Dirichlet prior distribution. It is commonly seen in practice that the posteriors of these distributions obtained after inference are

often sparse. In this section, we present a regularization technique to explicitly increase and control the sparsity in the topic proportion distribution using an entropic regularization framework that is is similar to the one presented in the previous section.

For every document d in the corpus, we define $H(\overline{\mathbf{z}_d})$ as the Shannon entropy of the observed topic proportion distribution. Since $\overline{\mathbf{z}_d}$ is obtained by repeatedly sampling θ_d, it can be considered as an approximation of θ_d. To incorporate the regularization, we add a noisy copy of the entropy score l_d, which is sampled from a Gaussian that is parameterized by the mean $H(\overline{\mathbf{z}_d})$ and variance $\sigma^2_{l_d}$ that is a hyperparameter to the model. (See Figure 5) As $\sigma^2_{l_d}$ tends to 0, the model reduces to a mixture of multinomials model since the regularization will require the entropies to be close to 0 implying that the distribution over topics has all its mass on one topic. Similarly, as the variance tends to ∞, the model reduces to a fully unconstrained LDA-like model.

The joint likelihood of the model with the document topic proportion regularization added is given by

(4.5)

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w}, \mathbf{y}, \mathbf{l_d} | \alpha, \gamma, \boldsymbol{\eta}, \sigma^2, \sigma^2_{l_d}) \propto$$
$$\prod_{k=1}^{K} \mathrm{Dir}\left(\beta_k | \gamma\right) \times$$
$$\left(\prod_{d=1}^{D} \mathrm{Dir}(\theta_d | \alpha) \left(\prod_{i=1}^{N_d} \theta_d^{(z_i)} \beta_{z_i}^{(w_i)} \right) \exp \frac{-\left(l_d - H(\overline{\mathbf{z}_d})\right)^2}{2\sigma^2_{l_d}} \right)$$

The Gibbs sampling equation to sample a topic indicator for a word is now given by

(4.6)

$$p(z_i = k | \mathbf{l_w}, l_d, w_i, \mathbf{z}^{\neg \mathbf{i}}, \mathbf{w}^{\neg \mathbf{i}}, y_d, \alpha, \gamma, \boldsymbol{\eta}, \sigma^2, \sigma^2_{l_d}) \propto$$
$$(n_{dk}^{\neg i} + \alpha) \frac{n_{kw_i}^{\neg i} + \gamma}{\sum_{w'} n_{kw'}^{\neg i} + |V|\gamma} \exp\left(\frac{-(y_d - \boldsymbol{\eta}.\overline{\mathbf{z}_d})^2}{2\sigma^2} \right)$$
$$\exp\left(\frac{-(l_d - H(\overline{\mathbf{z}_d}))^2}{2\sigma^2_{l_d}} \right)$$

4.1 Results We study the effect of the regularization on document perplexity in Figure 6. In all the datasets, the perplexity values decrease as the regularization is enforced and tightened. This effect is more prominently seen in the models with higher topic values. For instance, the purple lines at the top in each plot which represents the results with $K = 50$, show a more noticeable drop in perplexity whereas the solid red lines which indicate results with $K = 5$ show little movement with different levels of regularization. As the number of topics is increased, the unregularized model tends to use more diffuse distributions over topics and the regularization helps in reducing this tendency resulting in better perplexities.

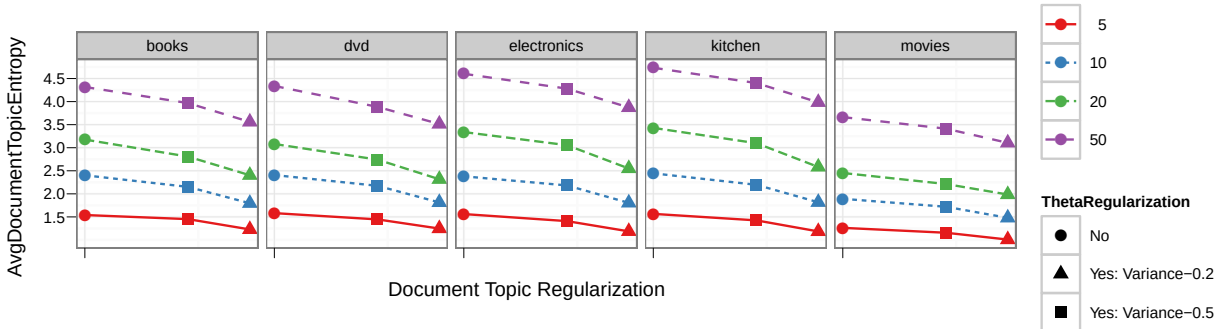
Next, we study the effect of regularization on star-prediction tasks, the results for which are shown in Table 4. For the results in the table, the variance hyperparameter for the regularization $\sigma^2_{l_d}$ was set to 0.2 (which is the optimal value based on the perplexity plots). It can be seen from the table that adding the regularization improves the MSE for all the datasets by 4.82% on average. This improvement is statistically significant at the 95% level.

As discussed in an earlier section, sparsifying the document topic distribution can also be done by adjusting the Dirichlet prior hyperparameter. The regularization technique presented in this technique achieves the sparsification in a much more direct manner.

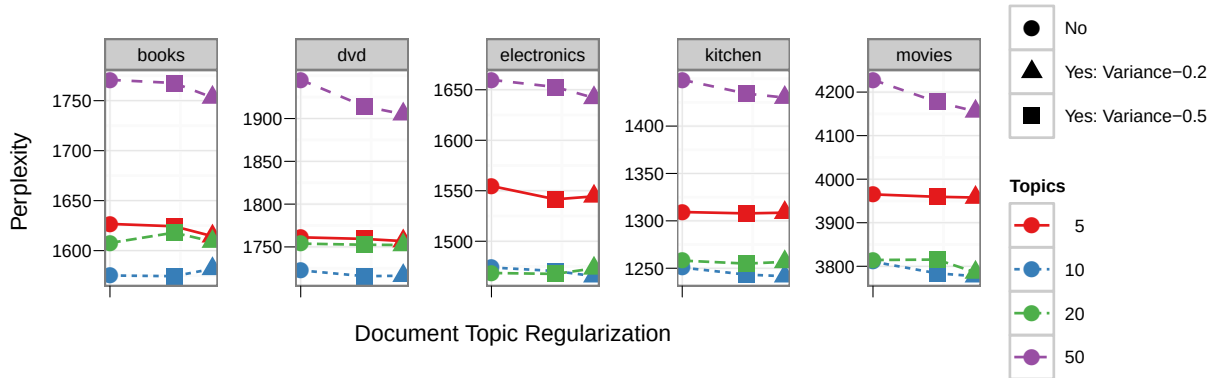
5 Related Work

Ganchev et al. [6] proposed Posterior Regularization (PR), a method to incorporate indirect supervision via constraints on posterior distributions of probabilistic models with latent variable. They demonstrate the use of the technique in models for several tasks such as POS induction, word alignment etc. While the approach proposed in this paper is similar in spirit to PR in that both approaches provide a method for preferences for the posteriors of latent variables to be specified, there are significant differences in the approaches. The PR framework is used in applications where exact inference is possible and the authors present a modified EM procedure to learn parameters for the model and incorporate constraints in an interleaved manner. In the regularization approach introduced in this paper, we focus on incorporating constraints on latent role distributions in models where exact inference is intractable by incorporating the constraints into the model instead of imposing them in a separate distinct step during inference; the constraints are introduced in a manner which only minimally affects the approximate inference procedure (by the addition of a few terms), through the use of pseudo-observed noisy copy of aggregate function values.

Mann and McCallum [7] also proposed a general framework to introduce preferences in model expectations by adding terms called *generalized expectation(GE) criteria* to the objective function. Examples of such criteria were explored in the domain of log linear models. The approach in this paper is similar to the GE framework in that the regularization operates on entropies of distributions of inferred latent variables. The manner in which deviations from expectations are pe-



(a) Average Document Topic Proportion Distribution Entropy



(b) Effect on Perplexity

Figure 6: Effect of Document Topic Proportion regularization.

Dataset	books	dvd	kitchen	electronics	movies
No regularization	2.169	2.042	1.986	1.885	0.218
With regularization	**2.121**	**2.031**	**1.784**	**1.743**	**0.210**

Table 4: Document Topic Regularization: Effect on MSE in star-rating prediction.

nalized, however differs from the criteria used by Mann and McCallum; the method introduced in this paper proposes that a desired value is drawn from a distribution parameterized by the inferred latent variables' values. The GE framework has not been applied to latent variable mixed-membership models as far as we know.

Blei and Chang [8] presented a model that jointly models documents and the network between them. The joint model encourages topics to have regularity in order to explain the documents and the network between them, but differs from the regularization scheme presented in this paper in that the regularization is based on characteristics of documents that are linked together rather than characteristics of the observed and latent variables as presented in this paper.

Opinion mining using topic models has been an active area of research in recent years. Titov and McDonald [9] presented a topic modeling approach that recovers topics that are aligned with aspects. Jo and Oh [10] also proposed a technique to recover topics that incorporate both sentiment and aspect. These approaches add tools to the topic modeling toolbox to specifically address opinion and aspect modeling in corpora annotated with sentiments. The regularization approach proposed in this paper is a general method to regularize latent variable models and could be potentially used with any of the models mentioned. We use the star prediction task, which is a specific task in opinion mining, only as a case-study to study the utility of the regularization.

Conclusion

We presented an entropic regularization approach to obtain finer control in permitting mixed-membership in latent variable models. We used the method to spar-

sify topic models; firstly we used it to softly constrain words' ability to participate in multiple topics thus providing a way to control the ability of the model to permit polysemy. We then used the entropic regularization approach to make the topic proportion distribution of documents sparse thus permitting LDA-like models to span the spectrum from a mixture of multinomials to a fully unconstrained LDA model. Our experiments show that the word entropy and document topic regularization result in better perplexity and mean-squared error scores in the star-rating prediction task because it enables the model to utilize sentiment-indicative words more efficiently. The entropic regularization technique presented enables finer control over latent variable modeling than that possible by prior modification without the disadvantage of making inference more complicated.

References

[1] David. M Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] Edoardo M. Airoldi, David Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, September 2008.

[3] David M. Blei Blei and Jon D. McAuliffe. Supervised Topic Models. In *NIPS*, 2007.

[4] John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, Bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In *ACL*, pages 187–205, 2007.

[5] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*, pages 115–124, 2005.

[6] Kuzman Ganchev, João Graça, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *J. Mach. Learn. Res.*, 11:2001–2049, August 2010.

[7] Gideon S Mann and Andrew Mccallum. Generalized Expectation Criteria for Semi-Supervised Learning with Weakly Labeled Data. *Journal of Machine Learning Research*, 11:955–984, 2010.

[8] Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, October 2010.

[9] Ivan Titov and Ryan McDonald. A Joint Model of Text and Aspect Ratings for Sentiment Summarization. pages 308–316, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[10] Yohan Jo and Alice H. Oh. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11*, page 815, New York, New York, USA, February 2011. ACM Press.