Mixed Membership Word Embeddings for Computational Social Science Supplementary Material

James R. Foulds
Department of Information Systems, University of Maryland, Baltimore County.

1 Related Work

In this supplementary document, we discuss related work in the literature and its relation to our proposed methods, provide a case study on NIPS articles, and derive the collapsed Gibbs sampling update for the MMSGTM, which we leverage when training the MMSG.

1.1 Topic Modeling and Word Embeddings

The *Gaussian LDA* model of Das et al. (2015) improves the performance of topic modeling by leveraging the semantic information encoded in word embeddings. Gaussian LDA modifies the generative process of LDA such that each topic is assumed to generate the vectors via its own Gaussian distribution. Similarly to our MMSG model, in Gaussian LDA each topic is encoded with a vector, in this case the mean of the Gaussian. It takes pre-trained word embeddings as input, rather than learning the embeddings from data within the same model, and does not aim to perform word embedding.

The topical word embedding (TWE) models of Liu et al. (2015) reverse this, as they take LDA topic assignments of words as input, and aim to use them to improve the resultant word embeddings. The authors propose three variants, each of which modifies the skip-gram training objective to use LDA topic assignments together with words. In the best performing variant, called *TWE-1*, a standard skip-gram word embedding model is trained independently with another skip-gram variant, which tries to predict context words given the input word's topic assignment. The skip-gram embedding and the topic embeddings are concatenated to form the final embedding.

At test time, a distribution over topics for the word given the context, $p(z_i|\text{context}(i))$ is estimated according to the topic counts over the other context words. Using this as a prior, a posterior over topics given both the input word and the context is calculated, and similarities between pairs of words (with their contexts) are averaged over this posterior, in a procedure inspired by those used by Reisinger and Mooney (2010); Huang et al. (2012). The primary similarity to our MMSG approach is the use of a training algorithm involving the prediction of context words, given a topic. Our method does this as part of an overall model-based inference procedure, and we learn mixed membership proportions $\theta^{(w)}$ rather than using empirical counts as the prior over topics for a word token. In accordance with the skip-gram's prediction model, we are thus able to model the context words in the data likelihood term when computing the posterior probability of the topic assignment. TWE-1 requires that topic assignments are available at test time. It provides a mechanism to predict contextual similarity, but not to predict held-out context words, so we are unable to compare to it in our experiments.

Other neurally-inspired topic models include replicated softmax (Hinton and Salakhutdinov, 2009), and its successor, DocNADE (Larochelle and Lauly, 2012). Replicated softmax extends the restricted Boltzmann machine to handle multinomial counts for document modeling. DocNADE builds on the ideas of replicated softmax, but uses the NADE architecture, where observations (i.e. words) are modeled sequentially given the previous observations.

1.2 Multi-Prototype Embedding Models

Multi-prototype embeddings models are another relevant line of work. These models address lexical ambiguity by assigning multiple vectors to each word type, each corresponding to a different meaning of that word. Reisinger and Mooney (2010) propose to cluster the occurrences of each word type, based on features extracted from its context. Embeddings are then learned for each cluster. Huang et al. (2012) apply a similar approach, but they use initial single-prototype word embeddings to provide the features used for clustering. These clustering methods have some resemblance to our topic model pre-clustering step, although their

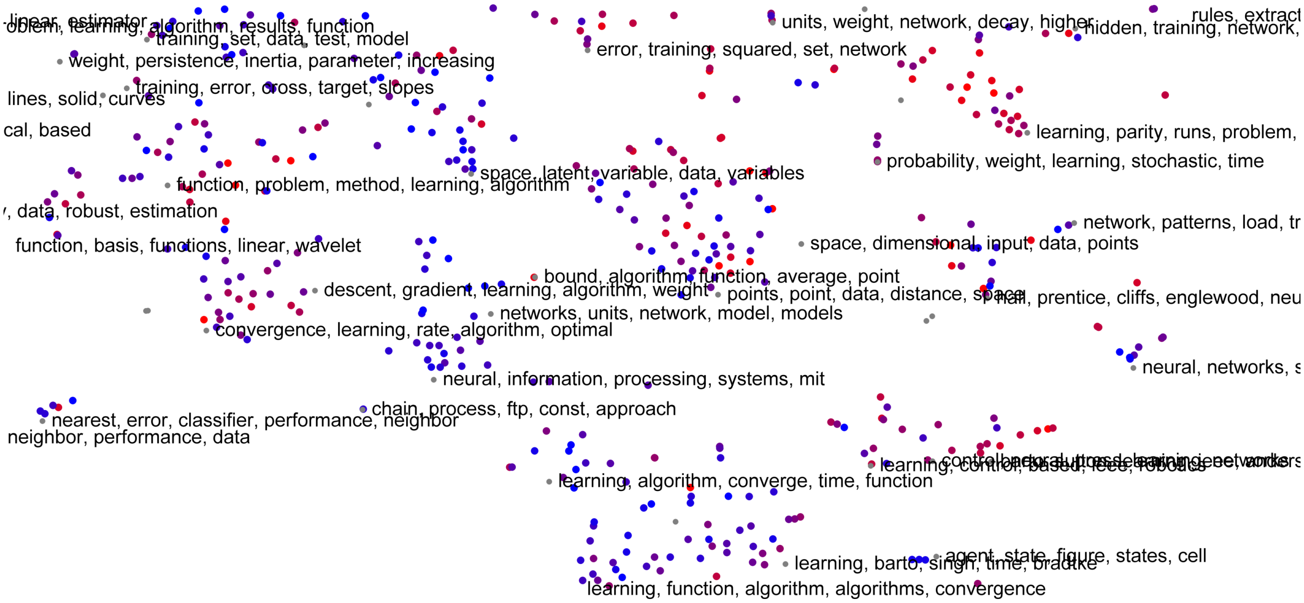

Figure 1: NIPS documents/topics, t-SNE, zoomed in. Blue/red = more recent/older, gray = topics.

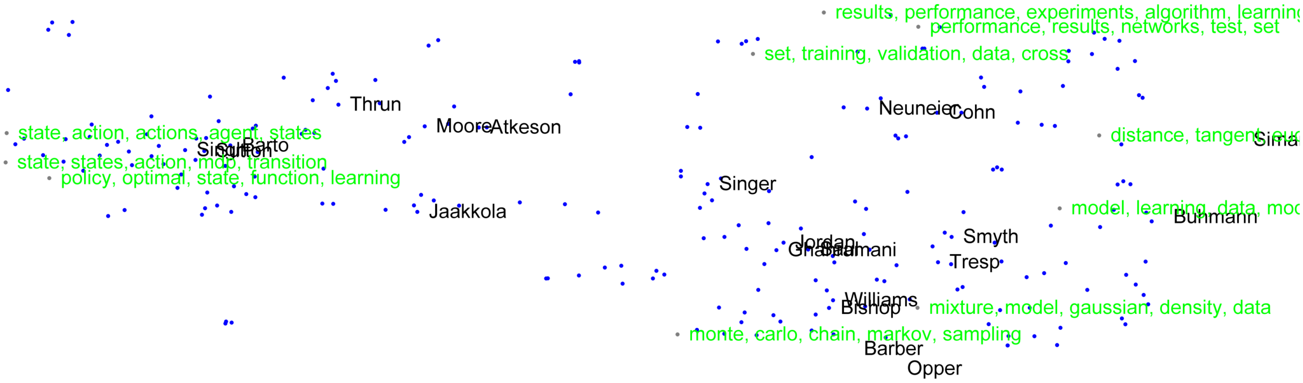

Figure 2: NIPS authors and topics, t-SNE, zoomed in. Blue = authors, gray = topics.

clustering is applied within instances of a given word type, rather than globally across all word types, as in our methods. This results in models with more vectors than words, while we aim to find fewer vectors than words, to reduce the model's complexity for small datasets. Rather than employing an off-the-shelf clustering algorithm and then applying an unrelated embedding model to its output, our approach aims to perform model-based clustering within an overall joint model of topic/cluster assignments and word vectors.

Perhaps the most similar model to ours in the literature is the probabilistic multi-prototype embedding model of Tian et al. (2014), who treat the prototype assignment of a word as a latent variable, assumed drawn from a mixture over prototypes for each word. The embeddings are then trained using EM. Our MMSG model can be understood as the mixed membership version of this model, in which the prototypes (vectors) are shared across all word types, and each word type has its own mixed membership proportions across the shared prototypes. While a similar EM algorithm can be applied to the MMSG, the E-step is much more expensive, as we typically desire many more shared vectors (often in the thousands) than we would prototypes per a single word type (Tian et al. use ten in their experiments). We use the Metropolis-Hastings-Walker algorithm with the topic model reparameterization of our model in order to address this by efficiently pre-solving the E-step.

1.3 Mixed Membership Modeling

Mixed membership modeling is a flexible alternative to traditional clustering, in which each data point is assigned to a single cluster. Instead, mixed membership models posit that individual entities are associated with multiple underlying clusters, to differing degrees, as encoded by a mixed membership vector that sums to one across the clusters (Erosheva et al., 2004; Airoldi et al., 2014). These mixed membership proportions are generally used to model lower-level grouped data, such as the words inside a document. Each lower-level data point inside a group is assumed to be assigned to one of the shared, global clusters according to the group-level membership proportions. Thus, a mixed membership model consists of a mixture model for each group, which share common mixture component parameters, but with differing mixture proportions.

This formalism has lead to probabilistic models for a variety of applications, including medical diagnosis (Manton et al., 1994), population genetics (Pritchard et al., 2000), survey analysis (Erosheva, 2003), computer vision (Barnard et al., 2003; Fei-Fei and Perona, 2005), text documents (Hofmann, 1999; Blei et al., 2003), and social network analysis (Airoldi et al., 2008). Nonparametric Bayesian extensions, in which the number of underlying clusters is learned from data via Bayesian inference, have also been proposed (Teh et al., 2006). In this work, dictionary words are assigned a mixed membership distribution over a set of shared latent vector space embeddings. Each instantiation of a dictionary word (an "input" word) is assigned to one of the shared embeddings based on its dictionary word's membership vector. The words in its context ("output" words) are assumed to be drawn based on the chosen embedding.

2 Case Study on NIPS

In Figure 1, we show a zoomed in t-SNE visualization of NIPS document embeddings. We can see regions of the space corresponding to learning algorithms (bottom), data space and latent space (center), training neural networks (top), and nearest neighbors (bottom-left). We also visualized the authors' embeddings via t-SNE (Figure 2). We find regions of latent space for reinforcement learning authors (left: "state, action,...," Singh, Barto,Sutton), probabilistic methods (right: "mixture, model," "monte, carlo," Bishop, Williams, Barber, Opper, Jordan, Ghahramani, Tresp, Smyth), and evaluation (top-right: "results, performance, experiments,...").

3 Derivation of the Collapsed Gibbs Update

Let $C_i = |\text{context}(i)|$ be the number of output words in the ith context, let $w_1^{(i)}, \ldots, w_{C_i}^{(i)}$ be those output words, and let $\mathbf{w}_{\neg i}$ be the input words other that w_i (similarly, topic assignments $\mathbf{z}_{\neg i}$ and output words $\mathbf{w}^{(\neg i)}$). Then the collapsed Gibbs update samples from the conditional distribution

$$\begin{aligned}
&p(z_i = k|\mathbf{z}_{\neg i}, w_i, w_1^{(i)}, \ldots, w_{C_i}^{(i)}, \mathbf{w}_{\neg i}, \mathbf{w}^{(\neg i)}, \alpha, \beta) \\
&\propto p(z_i = k, w_1^{(i)}, \ldots, w_{C_i}^{(i)}|\mathbf{z}_{\neg i}, w_i, \mathbf{w}_{\neg i}, \mathbf{w}^{(\neg i)}, \alpha, \beta) \\
&\quad = \int_{\phi^{(k)}} \int_{\theta^{(w_i)}} p(z_i = k, w_1^{(i)}, \ldots, w_{C_i}^{(i)}, \phi^{(k)}, \theta^{(w_i)}|\mathbf{z}_{\neg i}, \\
&\qquad\qquad w_i, \mathbf{w}_{\neg i}, \mathbf{w}^{(\neg i)}, \alpha, \beta) \\
&\quad = \int_{\phi^{(k)}} \int_{\theta^{(w_i)}} p(z_i = k, w_1^{(i)}, \ldots, w_{C_i}^{(i)}|\phi^{(k)}, \theta^{(w_i)}, w_i) \\
&\qquad\qquad \times p(\phi^{(k)}, \theta^{(w_i)}|\mathbf{z}_{\neg i}, w_i, \mathbf{w}_{\neg i}, \mathbf{w}^{(\neg i)}, \alpha, \beta) \\
&\quad = \int_{\phi^{(k)}} \int_{\theta^{(w_i)}} \theta_k^{(w_i)} \prod_{c=1}^{C_i} \phi_{w_c^{(i)}}^{(k)} \times p(\theta^{(w_i)}|\mathbf{z}_{\neg i : w_j = w_i}, \alpha) \\
&\qquad\qquad \times p(\phi^{(k)}|\mathbf{z}_{\neg i}, \mathbf{w}^{(\neg i)}, \beta) \\
&\quad = \int_{\theta^{(w_i)}} \theta_k^{(w_i)} p(\theta^{(w_i)}|\mathbf{z}_{\neg i : w_j = w_i}, \alpha) \\
&\qquad \times \int_{\phi^{(k)}} \prod_{c=1}^{C_i} \phi_{w_c^{(i)}}^{(k)} p(\phi^{(k)}|\mathbf{z}_{\neg i}, \mathbf{w}^{(\neg i)}, \beta) \ .
\end{aligned}$$

We recognize the first integral as the mean of a Dirichlet distribution which we obtain via conjugacy:

$$\begin{aligned}
p(\theta^{(w_i)}|\mathbf{z}_{\neg i : w_j = w_i}, \alpha) &= \text{Dirichlet}(\mathbf{n}_{:}^{(w_i)\neg i} + \alpha) \\
\int_{\theta^{(w_i)}} \theta_k^{(w_i)} p(\theta^{(w_i)}|\mathbf{z}_{\neg i : w_j = w_i}, \alpha) &= \frac{n_k^{(w_i)\neg i} + \alpha_k}{\sum_{k'} n_{k'}^{(w_i)\neg i} + \alpha_{k'}} \\
&\propto n_k^{(w_i)\neg i} + \alpha_k \ .
\end{aligned}$$

The above can also be understood as the probability of the next ball drawn from a multivariate Polya urn model, also known as the Dirichlet-compound multinomial distribution, arising from the posterior predictive distribution of a discrete likelihood with a Dirichlet prior. We will need the full form of such a distribution to analyze the second integral. Once again leveraging conjugacy, we have:

$$\begin{aligned}
&\int_{\phi^{(k)}} \prod_{c=1}^{C_i} \phi_{w_c^{(i)}}^{(k)} p(\phi^{(k)}|\mathbf{z}_{\neg i}, \mathbf{w}^{(\neg i)}, \beta) \\
&= \int_{\phi^{(k)}} \prod_{c=1}^{C_i} \phi_{w_c^{(i)}}^{(k)} \frac{\Gamma(\sum_{v=1}^{D}(n_v^{(k)\neg_i} + \beta_v))}{\prod_{v=1}^{D} \Gamma(n_v^{(k)\neg_i} + \beta_v)} \prod_{v=1}^{D} {\phi_v^{(k)}}^{n_v^{(k)\neg_i} + \beta_v - 1} \\
&= \int_{\phi^{(k)}} \frac{\Gamma(\sum_{v=1}^{D}(n_v^{(k)\neg_i} + \beta_v))}{\prod_{v=1}^{D} \Gamma(n_v^{(k)\neg_i} + \beta_v)} \prod_{v=1}^{D} {\phi_v^{(k)}}^{n_v^{(k)\neg_i} + \beta_v + n_v^{(i)} - 1}
\end{aligned}$$

$$= \frac{\Gamma(\sum_{v=1}^{D}(n_v^{(k)\neg_i} + \beta_v))}{\prod_{v=1}^{D}\Gamma(n_v^{(k)\neg_i} + \beta_v)} \frac{\prod_{v=1}^{D}\Gamma(n_v^{(k)\neg_i} + \beta_v + n_v^{(i)})}{\Gamma(\sum_{v=1}^{D}(n_v^{(k)\neg_i} + \beta_v + n_v^{(i)}))}$$

$$\times \int_{\phi^{(k)}} \frac{\Gamma(\sum_{v=1}^{D}(n_v^{(k)\neg_i} + \beta_v + n_v^{(i)}))}{\prod_{v=1}^{D}\Gamma(n_v^{(k)\neg_i} + \beta_v + n_v^{(i)})} \prod_{v=1}^{D} \phi_v^{(k)\, n_v^{(k)\neg_i} + \beta_v + n_v^{(i)} - 1}$$

$$= \frac{\Gamma(\sum_{v=1}^{D}(n_v^{(k)\neg_i} + \beta_v))}{\prod_{v=1}^{D}\Gamma(n_v^{(k)\neg_i} + \beta_v)} \frac{\prod_{v=1}^{D}\Gamma(n_v^{(k)\neg_i} + \beta_v + n_v^{(i)})}{\Gamma(\sum_{v=1}^{D}(n_v^{(k)\neg_i} + \beta_v + n_v^{(i)}))} ,$$

where $n_v^{(i)}$ is the number of times that output word v occurs in the ith context, since the final integral is over the full support of a Dirichlet distribution, which integrates to one. Eliminating terms that aren't affected by the z_i assignment, the above is

$$\propto \frac{\prod_{v=1}^{D}\Gamma(n_v^{(k)\neg_i} + \beta_v + n_v^{(i)})}{\Gamma(\sum_{v=1}^{D}(n_v^{(k)\neg_i} + \beta_v + n_v^{(i)}))}$$

$$= \frac{\prod_{v=1}^{D}\Big(\Gamma(n_v^{(k)\neg_i} + \beta_v)\prod_{j=0}^{n_v^{(i)}-1}(n_v^{(k)\neg_i} + \beta_v + j)\Big)}{\Gamma(\sum_{v=1}^{D}(n_v^{(k)\neg_i} + \beta_v))\prod_{j=0}^{C_i-1}(\sum_{v=1}^{D}(n_v^{(k)\neg_i} + \beta_v) + j)}$$

$$\propto \frac{\prod_{v=1}^{D}\prod_{j=0}^{n_v^{(i)}-1}(n_v^{(k)\neg_i} + \beta_v + j)}{\prod_{j=0}^{C_i-1}(\sum_{v=1}^{D}(n_v^{(k)\neg_i} + \beta_v) + j)}$$

$$= \prod_{c=1}^{C_i} \frac{n_{w_c}^{(k)\neg_i} + \beta_{w_c} + n_{w_c^{(i,c)}}}{n^{(k)\neg_i} + \sum_v \beta_v + c - 1}$$

where we have used the fact that $\Gamma(x+n) = (x+n-1)(x+n-2)...(x+1)x\Gamma(x)$ for any $x > 0$, and integer $n \geq 1$. We can interpret this as the probability of drawing the context words under the multivariate Polya urn model, in which the number of "colored balls" (word counts plus prior counts) is increased by one each time a certain color (word) is selected. In other words, in each step, corresponding to the selection of each context word, we draw a ball from the urn, then put it back, *along with another ball of the same color.* The $n_{w_c^{(i,c)}}$ and $c-1$ terms reflect that the counts have been changed by adding these extra balls into the urn in each step. The second to last equation shows that this process is exchangeable: it does not matter which order the balls were drawn in when determining the probability of the sequence. Multiplying this with the term from the first integral, calculated earlier, gives us the final form of the update equation,

$$p(z_i = k|\cdot) \propto (n_k^{(w_i)\neg i} + \alpha_k) \prod_{c=1}^{C_i} \frac{n_{w_c}^{(k)\neg_i} + \beta_{w_c} + n_{w_j^{(i,c)}}}{n^{(k)\neg_i} + \sum_v \beta_v + c - 1} .$$

References

Airoldi, E., Blei, D., Feinberg, S., and Xing, E. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014.

Airoldi, E. M., Blei, D. M., Erosheva, E. A., and Fienberg, S. E. (2014). Introduction to mixed membership models and methods. In *Handbook of Mixed Membership Models and Their Applications*. Chapman and Hall/CRC.

Barnard, K., Duygulu, P., Forsyth, D., Freitas, N. d., Blei, D. M., and Jordan, M. I. (2003). Matching words and pictures. *Journal of Machine Learning Research*, 3(Feb):1107–1135.

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

Das, R., Zaheer, M., and Dyer, C. (2015). Gaussian LDA for topic models with word embeddings. In *Proceedings of the 53nd Annual Meeting of the Association for Computational Linguistics*, pages 795–804.

Erosheva, E., Fienberg, S., and Lafferty, J. (2004). Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5220–5227.

Erosheva, E. A. (2003). Bayesian estimation of the grade of membership model. *Bayesian Statistics*, 7:501–510.

Fei-Fei, L. and Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. In *Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 524–531. IEEE.

Hinton, G. E. and Salakhutdinov, R. R. (2009). Replicated softmax: an undirected topic model. In *Advances in Neural Information Processing Systems*, pages 1607–1614.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57. ACM.

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics.

Larochelle, H. and Lauly, S. (2012). A neural autoregressive topic model. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 2708–2716.

Liu, Y., Liu, Z., Chua, T.-S., and Sun, M. (2015). Topical word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2418–2424.

Manton, K. G., Tolley, H. D., and Woodbury, M. A. (1994). *Statistical applications using fuzzy sets*. Wiley-Interscience.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Reisinger, J. and Mooney, R. J. (2010). Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117. Association for Computational Linguistics.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476).

Tian, F., Dai, H., Bian, J., Gao, B., Zhang, R., Chen, E., and Liu, T.-Y. (2014). A probabilistic model for learning multi-prototype word embeddings. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 151–160.