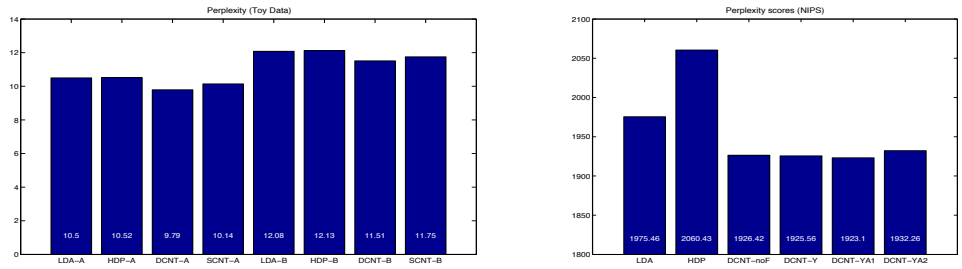

Figure 5: Perplexity scores (lower is better) computed via Chib-style estimators for several topic models. *Left:* Test performance for the toy datasets with uncorrelated bars (*-A*) and correlated bars (*-B*). *Right:* Test performance on the NIPS corpus with various metadata: no features (*-noF*), year features (*-Y*), year and prolific author features (over 10 publications, *-YA1*), and year and additional author features (over 5 publications, *-YA2*).

while the NIPS corpus was split into training and tests subsets containing 80% and 20% of the full corpus, respectively. Over the years 1988-1999, there were a total of 328 test documents.

We calculated predictive likelihood estimates using a Chib-style estimator [12]; for details see the supplemental material. In a previous comparison [19], the Chib-style estimator was found to be far more accurate than alternatives like the harmonic mean estimator. Note that there is some subtlety in correctly implementing the Chib-style estimator for our DCNT model, due to the possibility of rejection of our Metropolis-Hastings proposals.

Predictive negative log-likelihood estimates were normalized by word counts to determine perplexity scores [3]. We tested several models, including the SCNT and DCNT, LDA with $\alpha = 1$ and $\beta = 0.01$, and the HDP with full resampling of its concentration parameters. For the toy bars data, we set the number of topics to $K = 10$ for all models except the HDP, which learned $K = 15$. For the NIPS corpus, we set $K = 50$ for all models except the HDP, which learned $K = 86$.

For the toy datasets, the LDA and HDP models perform similarly. The SCNT and DCNT are both superior, apparently due to their ability to capture non-Dirichlet distributions on topic occurrence patterns. For the NIPS data, all of the DCNT models are substantially more accurate than LDA and the HDP. Including metadata encoding the year of publication, and possibly also the most prolific authors, provides slight additional improvements in DCNT accuracy. Interestingly, when a larger set of author features is included, accuracy becomes slightly worse. This appears to be an overfitting issue: there are 125 authors with over 5 publications, and only a handful of training examples for each one.

While it is pleasing that the DCNT and SCNT models seem to provide improved predictive likelihoods, a recent study on the human interpretability of topic models showed that such scores do not necessarily correlate with more meaningful semantic structures [4]. In many ways, the interactive visualizations illustrated in Sec. 4.2 provide more assurance that the DCNT can capture useful properties of real corpora.

5 Discussion

The doubly correlated nonparametric topic model flexibly allows the incorporation of arbitrary features associated with documents, captures correlations that might exist within a dataset's latent topics, and can learn an unbounded set of topics. The model uses a set of efficient MCMC techniques for learning and inference, and is supported by a set of web-based tools that allow users to visualize the inferred semantic structure.

Acknowledgments

This research supported in part by IARPA under AFRL contract number FA8650-10-C-7059. Dae Il Kim supported in part by an NSF Graduate Fellowship. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.