Regularized supervised topic models for high-dimensional multi-class regression

Måns Magnusson, **Leif Jonsson and Mattias Villani**

Linköping University, Sweden

mans.magnusson@liu.se

SIXTH IMS-ISBA JOINT MEETING
BAYES COMP at MCMSKI V
LENZERHEIDE
January 5-7, 2016

1. Abstract

Supervised topic models are becoming increasingly popular as an approach for modeling textual data. We introduce a supervised topic model to handle both many classes as well as many covariates. To handle many classes we use the recently proposed Diagonal Orthant (DO) probit model (Johndrow et al., 2013) for multi-class classification together with an efficient horseshoe prior for variable selection/shrinkage (Carvalho et al., 2010). We apply the model successfully to a bug prediction application where 161 classes are accurately predicted using 100 topics.

2. Related work

Supervised topic models try to infer topics that are both semantical **and** have a predictive performance. The sampling of the topic indicators in these models all contain a "supervised" part in the sampling procedure.

Model	Gibbs sampling $p(z_i = k \mid \cdot) \propto$
MedLDA (Jiang et al., 2012)	$\theta_{d,k} \cdot \phi_{v,k} \cdot \exp\left(\frac{1}{N_d}\sum_y (\lambda_d^y)^* \left(\kappa^*_{y_d k} - \kappa^*_{yk}\right)\right)$
HSLDA (Perotte et al., 2011)	$\theta_{d,k} \cdot \phi_{v,k} \cdot \exp\left(\sum_{l \in \mathcal{L}_d} -\frac{1}{2}\left[\bar{\mathbf{z}}_d^T \eta_l - a_{l,d}\right]^2\right)$
Logistic SLDA (Zhu et al., 2013)	$\theta_{d,k} \cdot \phi_{v,k} \cdot \exp\left(\gamma\kappa_d\eta_k - \frac{1}{2}\lambda_d\left[\gamma^2\eta_k^2 + 2\gamma(1-\gamma)\eta_k\Lambda^k_{dn}\right]\right)$

3. Diagonal orthant topic model

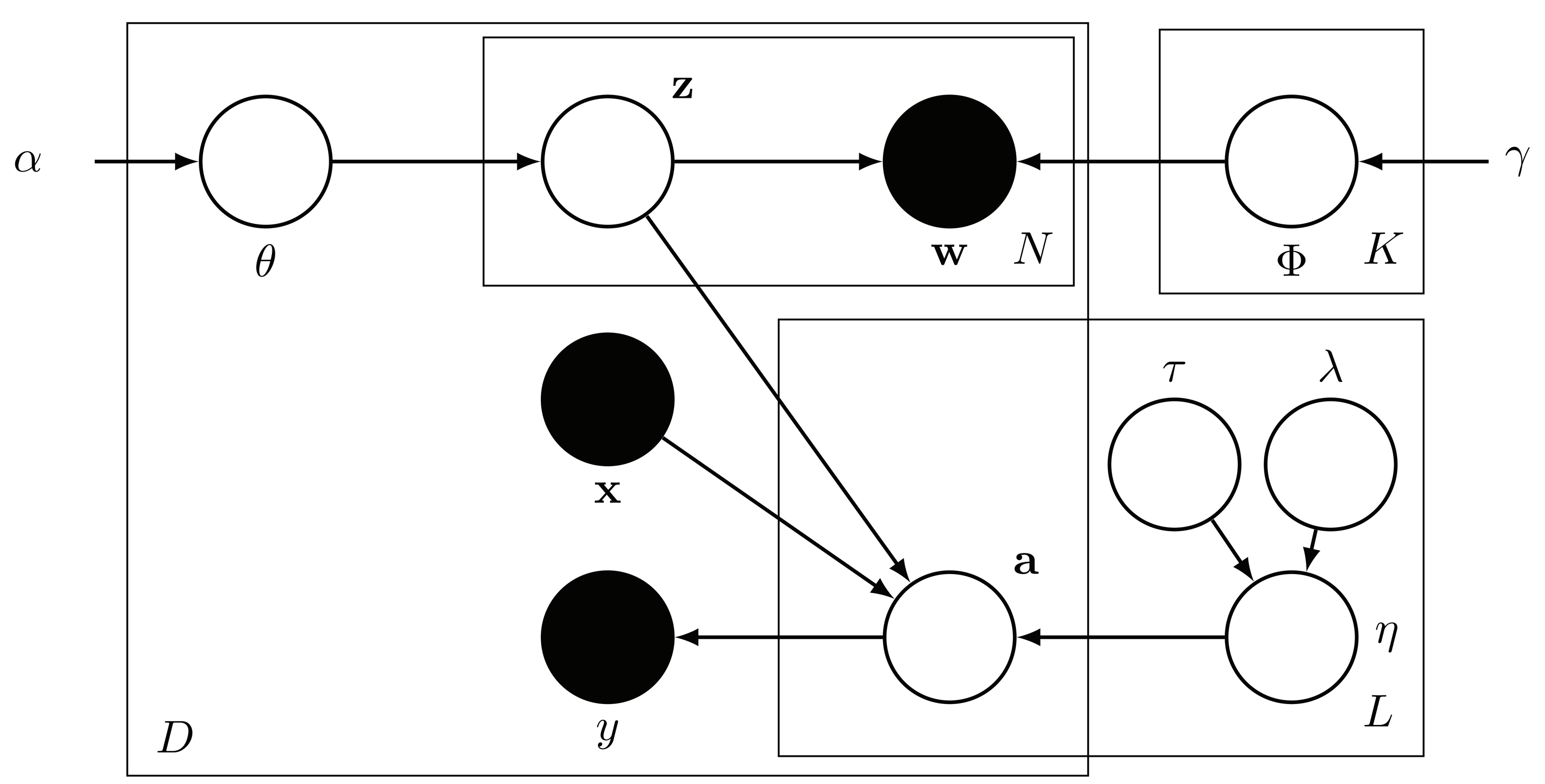

Figure 1: *Generative model*

4. Inference

We use MCMC to infer the model parameters in parallel using gibbs sampling and slice samling. Unlike the common fully collapsed gibbs for topic models, we use a partially collapsed sampler for $\mathbf{z}$ to sample the topic indicators in parallel. Similarly, we use the sparse sampler of Magnusson et al. (2015) to sample the test set efficiently.

1. Sample the latent variables $a^{(i)}_{d,l} \sim \mathcal{N}_+((\mathbf{x}\,\bar{\mathbf{z}})_d^T\eta_l, 1)$ for $l = y_d$ and $a_{d,l} \sim \mathcal{N}_-((\mathbf{x}\,\bar{\mathbf{z}})_d^T\eta_l, 1)$ for $l \neq y_d$, where $\mathcal{N}_+$ and $\mathcal{N}_-$ is the positive and negative truncated normal distribution, truncated at 0.
2. Sample all the regression coefficients as in an ordinary Bayesian linear regression per class label l where $\eta_l \sim \mathcal{MVN}\left(\mu_l, ((\mathbf{X}\,\bar{\mathbf{z}})^T(\mathbf{X}\,\bar{\mathbf{z}}) + \tau_l^2\Lambda_l)^{-1}\right)$ and Λ_l is a diagonal matrix with the local shrinkage parameters λ_i per parameter in η_l and $\mu_l = ((\mathbf{X}\,\bar{\mathbf{z}})^T(\mathbf{X}\,\bar{\mathbf{z}}) + \tau_l^2\Lambda_l)^{-1}(\mathbf{X}\,\bar{\mathbf{z}})^T\mathbf{a}_l$
3. Sample the global shrinkage parameters τ_l at iteration j using the following two step slice sampling:

$$u \sim \mathcal{U}\left(0, \left[1 + \frac{1}{\tau_{l,(j-1)}}\right]^{-1}\right)$$

$$\frac{1}{\tau^2_{l,j}} \sim \mathcal{G}\left((p+1)/2, \frac{1}{2}\sum_{p=1}^{P}\left(\frac{\eta_{l,p}}{\lambda_{l,p}}\right)^2\right) I\left[\frac{1}{\tau^2_{l,(j-1)}} < (1-u)/u\right]$$

   where I indicates the truncation region for the truncated gamma.

4. Sample each local shrinkage parameter $\lambda_{i,l}$ as

$$u \sim \mathcal{U}\left(0, \left[1 + \frac{1}{\lambda^2_{p,l,(j-1)}}\right]^{-1}\right)$$

$$\frac{1}{\lambda^2_{p,l,j}} \sim \mathcal{E}\left(\frac{1}{2}\left(\frac{\eta_{l,p}}{\tau_l}\right)^2\right) I\left[\frac{1}{\lambda^2_{p,l,(j-1)}} < (1-u)/u\right]$$

5. Sample the topic indicators $\mathbf{z}$

$$p(z_{i,d} = k \mid w_i, \mathbf{z}^{\neg i}, \eta, \mathbf{a}) \propto \phi_{v,k} \cdot \left(\mathbf{M}^{\neg i}_{d,k} + \alpha\right) \cdot \exp\left(-\frac{1}{2}\sum_{l}^{L}\left[-2\frac{\eta_{l,k}}{N_d}\left(a_{d,l} - (\bar{\mathbf{z}}_d^{\neg i}\,\mathbf{x}_d)\eta_l^\intercal\right) + \left(\frac{\eta_{l,k}}{N_d}\right)^2\right]\right)$$

5. Experiments

We use 5-fold cross validation, remove stopwords and rare words (< 10). We then run the MCMC chain for 10 000 iterations. Running one experiment with 100 topics for the 20 Newsgroups dataset took approximate 6.5 hours.

We evaluate the model on two datasets:

Dataset	Classes (L)	Vocabulary (V)	Documents (D)	Tokens (N)
IMDB	20	7530	8648	307569
20 Newsgroups	20	23941	15077	2008897

6. Results

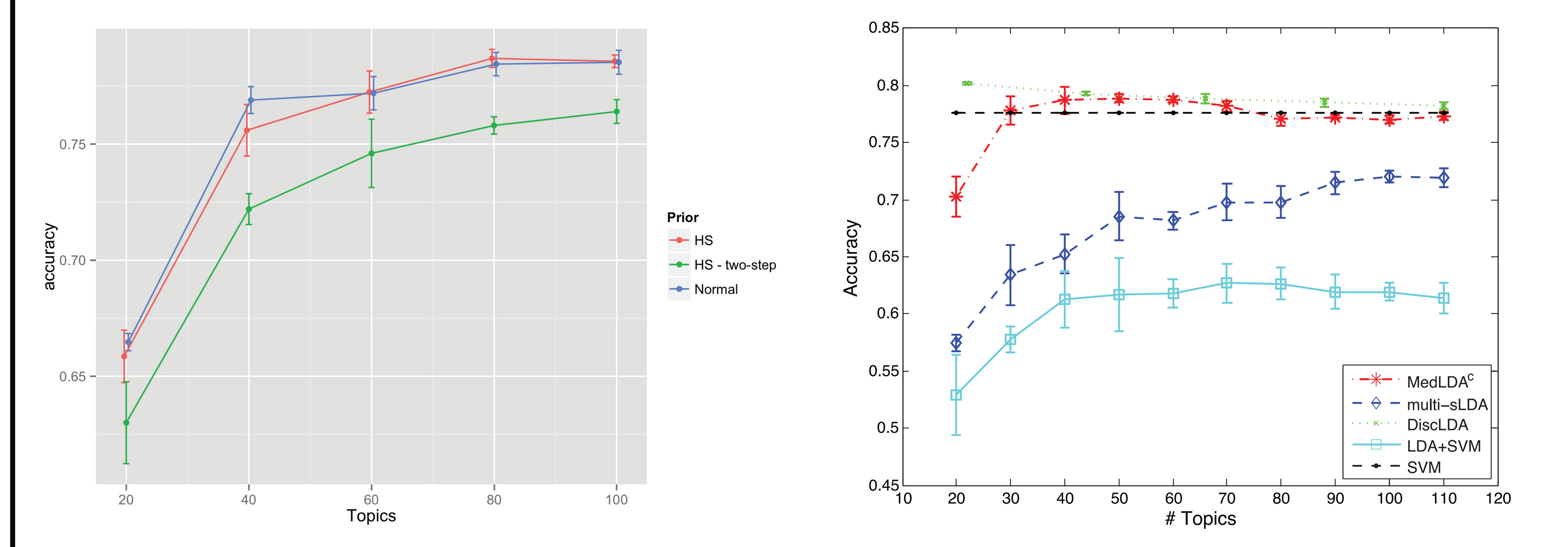

Figure 2: *Accuracy of the DO topic model (left) and other models in Zhu et al. (2012) (right)*

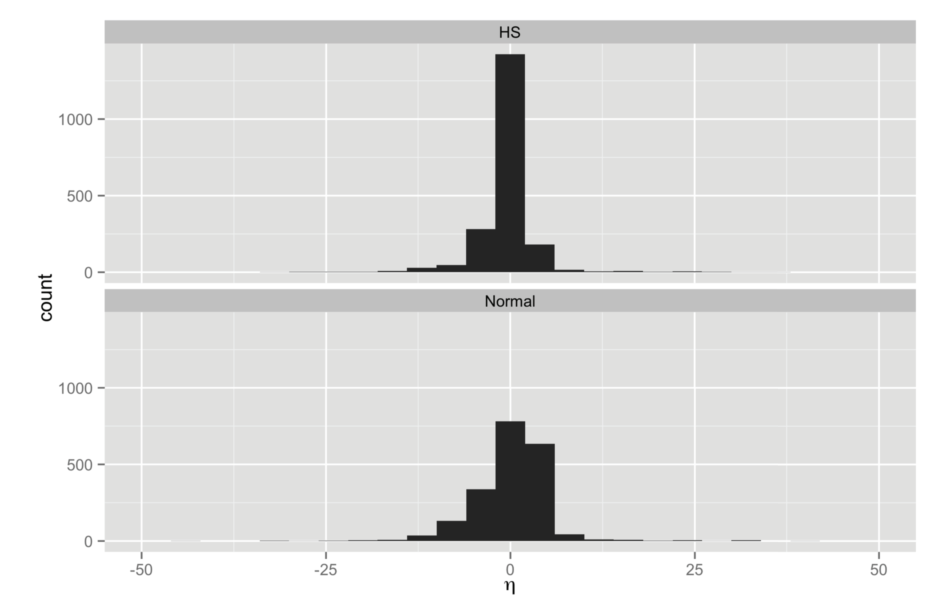

Figure 3: *Shrinkage of the coefficients for the 20 Newsgroups data with 100 topics*

References

Carvalho, C. M., Polson, N. G., Scott, J. G., 2010. The horseshoe estimator for sparse signals. Biometrika, asq017.

Jiang, Q., Zhu, J., Sun, M., Xing, E. P., 2012. Monte carlo methods for maximum margin supervised topic models. In: Advances in Neural Information Processing Systems. pp. 1592–1600.

Johndrow, J., Dunson, D., Lum, K., 2013. Diagonal orthant multinomial probit models. In: Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics. pp. 29–38.

Magnusson, M., Jonsson, L., Villani, M., Broman, D., 2015. Parallelizing lda using partially collapsed gibbs sampling. arXiv preprint arXiv:1506.03784.

Perotte, A. J., Wood, F., Elhadad, N., Bartlett, N., 2011. Hierarchically supervised latent dirichlet allocation. In: Advances in Neural Information Processing Systems. pp. 2609–2617.

Zhu, J., Ahmed, A., Xing, E. P., 2012. Medlda: maximum margin supervised topic models. the Journal of machine Learning research 13 (1), 2237–2278.

Zhu, J., Zheng, X., Zhang, B., 2013. Improved bayesian logistic supervised topic models with data augmentation. arXiv preprint arXiv:1310.2408.