Online supplement for: "A model of text for experimentation in the social sciences"

A Estimation of topic prevalence and content

Optimization of coefficients governing the topical content and prevalence models are dependent on the choice of priors.

Topic prevalence coefficients are given a Gaussian prior with topic-specific variance. These variances are given conjugate priors such that,

$$\gamma_{p,k} \sim \text{Normal}(0, \sigma_k^2) \tag{1}$$

$$\sigma_k^2 \sim \text{Inverse-Gamma}(1, 1) \tag{2}$$

where p indexes the covariates in the design matrix X. We leave the intercept unpenalized and compute the posterior mean of $\gamma_{p,k}$ using standard variational linear regression (Drugowitsch 2013, e.g.). The $\boldsymbol{\gamma}_k$ update for a given penalty parameter takes the form of a penalized linear regression

$$\hat{\gamma}_k = \left(X^T X + \text{diag}(1/\sigma_k^2)\right)^{-1} X^T \lambda_k \tag{3}$$

and the update for the variance is

$$\hat{\sigma_k^2} = \left(.5 + \sum_p \gamma_{p,k}^2\right) / (.5 + p)$$

We iterate between the coefficients and the shared variance until convergence.

In our software we also provide the option to estimate γ using the `glmnet` package (Friedman et al. 2010) which allows for penalized estimation using the lasso or elastic-net. This can be particularly useful for high-dimensional factor variables where many of the topic-effects are expected to be exactly zero. In this case the full regularization path is estimated and the value of the penalty parameter is selected using a modified information criterion (Taddy 2013b).

Estimation of the topical content covariate coefficients κ is equivalent to a multinomial logistic regression on the token level latent variables ϕ. In order to induce sparsity in κ we assign Laplace priors and compute the posterior mode using the lasso. In order to make the procedure more computationally efficient we adopt the distributed multinomial regression approach of Taddy (2013a). The idea is to use a plugin estimator for the document fixed effects which decouples the parameters of the single multinomial logistic regression into independent poisson regressions (one for each element of the vocabulary). This approach is not only faster but also allows for the operations to be parallelized over the vocabulary. The regularization parameter controlling sparsity is chosen using a modified information criterion as described in Taddy (2013b,a).

We also provide the option of estimating the topical content coefficients using a Normal-Jeffreys prior such that $\kappa_{k,v} \sim$Normal$(0, \tau_{k,v})$ and $\tau_{k,v} \sim 1/\tau_{k,v}$. The prior for τ is the improper Jeffreys prior. Here estimation involves alternating between maximization of κ and τ. Following Eisenstein et al. (2011) we use a block relaxation approach where each V-length vector in κ_k is updated using quasi-Newton methods, followed by an update of the variances. Taking as an example the vector for topic k we obtain the objective and gradient

$$\mathcal{L}_{\boldsymbol{\kappa}_k} = \langle \mathbf{c_k} \rangle \boldsymbol{\kappa}_k - \langle C_k \rangle \log \sum_v \exp(\kappa_{k,v} + m_v)) - \frac{1}{2} \boldsymbol{\kappa}_k^2 / \boldsymbol{\tau_k} \tag{4}$$

$$\nabla \mathcal{L}_{\boldsymbol{\kappa}_k} = \langle \mathbf{c_k} \rangle - \sum_j \langle C_{jk} \rangle \boldsymbol{\beta_{jk}} - \boldsymbol{\kappa_k} / \boldsymbol{\tau_k} \tag{5}$$

where $\langle c_k \rangle$ is V-length vector of expected counts for each term in the vocabulary for topic k. C_k is the summation over that vector producing a scalar equal to the expected number of words assigned to topic k. Updates for each covariate and interaction proceed analogously.

After each update of $\boldsymbol{\kappa}$ we update the corresponding penalty vector $\boldsymbol{\tau}$ with its variational expectation $\hat{\tau}_{v,k} = \kappa_{v,k}^2$.

References

Drugowitsch, J. (2013). Variational bayesian inference for linear and logistic regression. *arXiv preprint arXiv:1310.5438*.

Eisenstein, J., Ahmed, A., and Xing, E. (2011). Sparse additive generative models of text. In *Proceedings of ICML*, pages 1041–1048.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Taddy, M. (2013a). Distributed multinomial regression. *arXiv preprint arXiv:1311.6139*.

Taddy, M. (2013b). The gamma lasso. *arXiv preprint arXiv:1308.5623*.