Text Clustering with the Gibbs Sampler

Liyi Zhang

Model. A mixture model is implemented via the Gibbs sampler on the AP dataset to cluster texts. The generative model can be summarized as:

1. Draw proportions $\theta \sim \text{DIRICHLET}_K(\alpha)$
2. for k in $1 : K$
    - Draw a component vector $\lambda_k \sim \text{GAMMA}(a, b)$
3. for i in $1 : N$
    - Draw a cluster assignment $z_i \sim \text{CATEGORICAL}(\theta)$
    - Draw a data-point $X_i \sim \text{MULTI-POISSON}(\lambda_{z_i})$

where we have N articles, or data-points, and K clusters. To focus on analyzing the behavior of the Gibbs sampler at a given K, we set $K = 5$. We denote the size of vocabulary by V, and thus each λ_k is a V-length vector. Data X is an N-by-V matrix, with entry $X_{i,j}$ defined as the count of word j in article i.

Implementation Details. *Data preparation.* We clean the data so that only words that appeared in more than one article are counted. We also train on 90% of the data, with 10% heldout data.

Initialization. Each component vector λ_k starts at the mean of $X_{:,j}$ with random Gaussian perturbation, where the perturbation has mean 0 and standard deviation 0.5 times the mean of $X_{:,j}$.

Prior. α, the DIRICHLET prior, is a vector of 1. a and b, the GAMMA priors, are 0.1. These priors are non-informative, since they are washed out by the data. In computing the complete-conditionals, we utilize conjugacy, and, for Gamma-Poisson, the posterior is $\text{GAMMA}(a + \Sigma X^{(k)}_{:,j}, b + n^{(k)})$; for Dirichlet-Categorical, the posterior is $\text{DIRICHLET}(\alpha + \Sigma z^{(k)})$.

Technicals. We implement from scratch in the R programming language.

Results. *Deviance.* We run two chains, each for 1000 iterations (each chain with different initialization, with the random perturbation). First, we check the deviance, or log of the joint-distribution of data and parameters. Figure 1 shows that the deviance converges for each chain. The log-joint increases to a certain mode and remains stable. However, we can also see that the two chains converge at different deviance values, and we will analyze this phenomenon in section *across-chain convergence,* or rather, divergence.

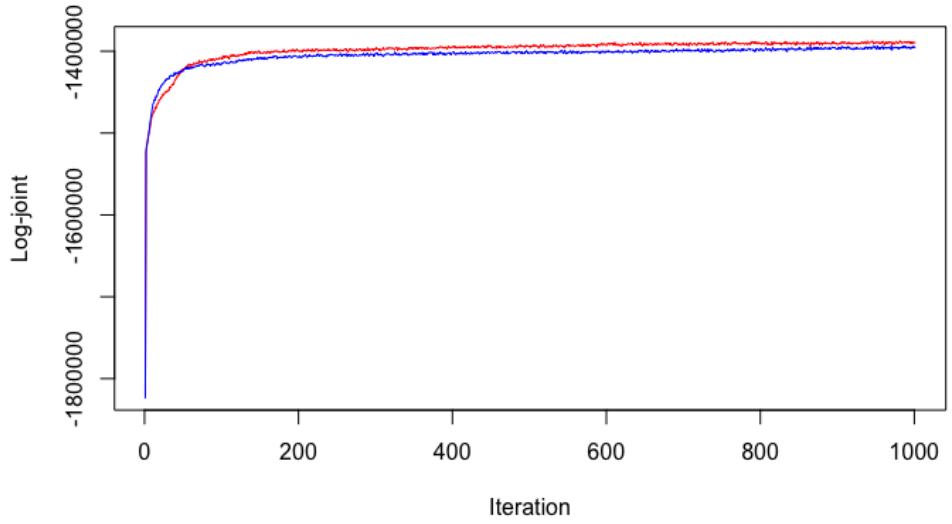

Figure 1: Deviance, or log-joints, of chain 1 (red) and chain 2 (blue) across all iterations. The deviance converges, but for each chain converges at a different value.

Within-chain convergence. Figure 2 plots sampled parameters of two chains. It suggests that each chain achieves stationarity. Visualization shows that the samples of λ parameters oscillate around a certain values, without being directional (increasing or decreasing). Because chains that are run here do not mix, diagnosis metrics cannot be reliable. A cursory computation and examination shows that, each chain individually satisfies convergence metrics such as number of effective sample size (Neff) and autocorrelation on each chain. Unfortunately, because the chains do not mix, these metrics' values do not yield many insights.

We compute Neff by the following:

$$\hat{n}_{eff} = \frac{n}{1 + 2\sum_{t=1}^{T} \hat{\rho}_t},$$

where n is number of actual samples, ρ_t is autocorrelation with lag t, and T is a stopping criteria, the first occurrence where ρ_{T+1} and $\rho_{T+2} < 0$ (Gelman, et al. 2013).

Across-chain convergence. While each chain appears stationary, the chains do not mix. Blei, et al. (2003) writes that different initializations on mixed-membership models on text data usually converge to different local modes. This phenomenon also applies here to the mixture model. This figure indicates that each chain is likely exploring its own mode. An interpretation is that, with different initializations, the sampler finds different ways to cluster these texts into five clusters. Indeed, as we shall see from 'topics' (Figure 3 and 4), these chains interpret different topics.

Visualizing topics. Our analysis is inspired by LDA (Blei, et al. 2003), and uses the following

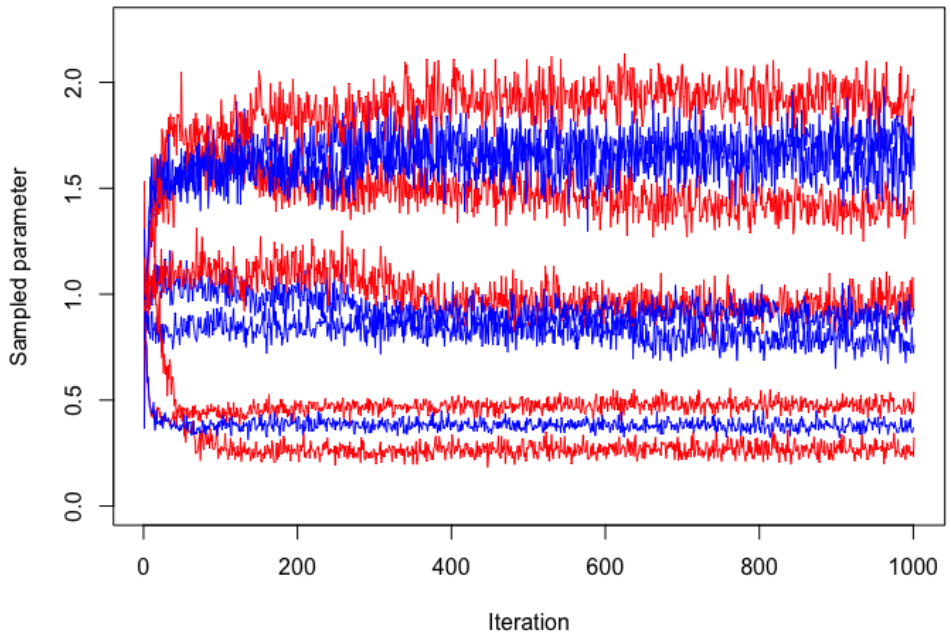

Figure 2: Samples of $\lambda_{:,1}$, i.e., the component-parameter across classes for the first word, from chain 1 (red) and chain 2 (blue). (Reasonably, there are 5 red lines and 5 blue lines.) While each chain eventually shows within-chain convergence, the figure demonstrates across-chain divergence. It is also worthwhile to note that, for $\lambda_{:,v}$ corresponding to less common words, posterior variance would be larger.

```
 "percent" "billion" "market"  "stock"   "index"   "prices"  "futures" "million"
 "cent"    "oil"     "cents"   "trade"   "year"    "stocks"  "higher"  "deficit"
 "trading" "lower"   "tax"     "sales"
"------------------------------"
 "dollar"   "yen"      "mecham"   "francs"   "ounce"    "late"     "fair"
 "troy"     "i"        "aspirin"  "bid"      "dukakis"  "dealers"  "trust"
 "gold"     "fbi"      "manville" "milken"   "london"   "lire"
"------------------------------"
 "soviet"     "gorbachev"  "party"      "bush"       "president"  "jackson"
 "dukakis"    "communist"  "south"      "union"      "government" "i"
 "political"  "aid"        "east"       "talks"      "reagan"     "republics"
 "united"     "leader"
"------------------------------"
 "police"    "killed"    "army"      "iraq"      "shot"      "wounded"   "death"
 "lebanon"   "hostages"  "iranian"   "military"  "people"    "iran"      "un"
 "israeli"   "fire"      "beirut"    "troops"    "kuwait"    "greyhound"
"------------------------------"
 "sothebys" "dresses"  "index"    "hubbert"  "art"      "school"   "ferret"
 "auction"  "teachers" "fe"       "grammer"  "magazine" "film"     "hunt"
 "diamond"  "films"    "survey"   "police"   "memorial" "trump"
"------------------------------"
```

Figure 3: 20 words with highest term-score from each class, using samples from chain 2. The first topic is (roughly) economics and stocks-trading; the second topic is international currency; the third topic is politics and US-Soviet affairs; the fourth topic is news from the Middle-East; the fifth topic is some day-to-day niceties.

term-score to visualize topics (Blei and Lafferty, 2009) as one form of model-checking:

$$\text{term-score}_{k,v} = \hat{\lambda}_{k,v} \log(\frac{\hat{\lambda}_{k,v}}{\prod_{j=1}^{K} \hat{\lambda}_{j,v}}).$$

We estimate $\hat{\lambda}_{k,v}$ by averaging over sampled parameters across the last 500 iterations. For each chain and for each cluster k, we print out the 20 words that have highest term-score. (We claim an article to be from cluster k by using sampled z at the last iteration.) Figure 3 and 4 show these words for chain 2 and 1, respectively. They are indeed very interpretable topics. For example, Figure 3 shows that the first topic is (roughly) economics and stocks-trading; the second topic is international currency; the third topic is politics and US-Soviet affairs; the fourth topic is news from the Middle-East; the fifth topic is some day-to-day niceties. Furthermore, the figures confirm that chain 1 and chain 2, which do not have across-chain convergence, explore different topics.

As reference, we also print out average of z over samples (across last 500 iterations), which indicate proportions of each of the five topics. For chain 1, it is $0.180, 0.091, 0.164, 0.223, 0.342$; for chain 2, it is $0.380, 0.154, 0.168, 0.142, 0.156$.

Further work. First, we can test this Gibbs Sampler on simpler, one-mode models to confirm the algorithm itself. Second, we can run more chains. Third, we can run the model on different values of K and compute probabilities on the heldout dataset. (Hypocritically, I have a 10% heldout dataset that I didn't use.) Fourth, we can further explore the model by adding complexity and flexibility.

```
 "police"        "fire"          "northern"      "injured"       "rain"
 "aoun"          "inches"        "beirut"        "army"          "wounded"
 "aouns"         "snow"          "winds"         "hezbollah"     "sinhalese"
 "moslem"        "richter"       "hospital"      "manila"        "thunderstorms"
"------------------------------"
 "iraq"       "military"   "kuwait"     "iraqi"      "army"       "defense"
 "abortion"   "i"          "government" "iran"       "noriega"    "contras"
 "mecham"     "saudi"      "police"     "israel"     "troops"     "people"
 "navy"       "trial"
"------------------------------"
 "dukakis"    "soviet"     "bush"       "gorbachev"  "party"      "i"
 "jackson"    "campaign"   "president"  "trade"      "political"  "democratic"
 "bentsen"    "republics"  "poll"       "reagan"     "gorbachevs" "delegates"
 "east"       "communist"
"------------------------------"
 "court"     "trial"     "judge"     "keating"   "prison"    "nordstrom" "case"
 "workers"   "convicted" "federal"   "fbi"       "jury"      "deconcini" "employees"
 "contract"  "attorney"  "company"   "million"   "police"    "sentence"
"------------------------------"
 "percent"  "yen"      "market"   "dollar"   "cents"    "prices"   "index"
 "billion"  "futures"  "stock"    "cent"     "trading"  "rose"     "rates"
 "stocks"   "higher"   "million"  "lower"    "analysts" "sales"
"------------------------------"
```

Figure 4: Same method, but from chain 1. It shows that chain 1 explores different clusterings from chain 2, demonstrating topics that chain 2 does not have (in particular, a court-trial-crime related topic). The chains also roughly share a few similar topics.

References

Blei, D. M. & Lafferty, J. D. (2009). Topic Models.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation (pp. 993-1022). Journal of Machine Learning Research.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian Data Analysis (3rd ed.). Chapman & Hall/CRC.

David M. Blei (2014). Build, Compute, Critique, Repeat: Data Analysis with Latent Variable Models. *Annual Review of Statistics and Its Application, 1, 203-232.* doi: 10.1146/annurev-statistics-022513-115657.