Application of LDA topic model to song lyrics

Silvia Ventoruzzo

Freie Universität Berlin

1. Motivation

- Topic models are unsupervised learning methods to extract topics from a corpus of documents
- Latent Dirichlet Allocation (LDA) is "generative probabilistic model for collections of discrete data such as text corpora" [1]
- Commonly used methods to estimate the latent parameters: VEM [1] and Gibbs Sampling (Griffiths and Steyvers, 2004). VEM will be used in this project
- LDA can be used for general discrete data, but it is often used in the context of text mining
- We will try to obtain topics

2. Latent Dirichlet Allocation [1]

Model

- The corpus D is a collection of M documents: $D = \{d_1, ..., d_M\}$
- A document of the corpus d_j is a sequence of N_j words: $d_j = (w_1, ..., w_{N_j})$
- A word present in the corpus w_i is an item from the vocabulary, that is the set of all words present in the corpus ponweiser2012latent: $w_i \in \{1, ..., V\}$

Generative process

- For all topics $k \in \{1, ..., K\}$:
  1. Choose a word distribution: $\beta_k \sim Dir(\delta)$
- For all documents d_j where $j \in \{1, ..., M\}$:
  1. Choose a topic distribution: $\theta_j \sim Dir(\alpha)$
  2. Assign topics for all words w_i where $i \in \{1, ..., N_j\}$: $z_{j,i} \sim Mult(\theta_j)$
  3. Choose a word w_i: $w_{j,i} \sim Mult(\beta_{z_{j,i}})$

Estimation

Optimize log-likelihood of the marginal distribution of the documents d_j with respect to the coefficients α and β:

$$p(d_j|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{i=1}^{N_j} \sum_{z_{j,i}} p(z_{j,i}|\theta) p(w_{j,i}|z_{j,i}, \beta) \right) d\theta$$

Inference

Since the function $p(d_j|\alpha, \beta)$ is intractable, variational Bayesian inference is used with the variational parameters γ and ϕ:

$$(\gamma^*, \phi^*) =_{(\gamma,\phi)} D_{KL}(q(\theta, \mathbf{z}|\gamma, \phi)||p(\theta, \mathbf{z}|\mathbf{d}, \alpha, \beta))$$

Variational expectation-maximization (VEM):

- **E-step**: Find $\{\gamma_j^*, \phi_j^*\}$, where $d_j \in D$, i.e. the optimal values of the variational parameters γ and ϕ for each document
- **M-step**: Maximize the resulting lower bound on the log-likelihood with respect to the latent parameters α and β

Number of topics

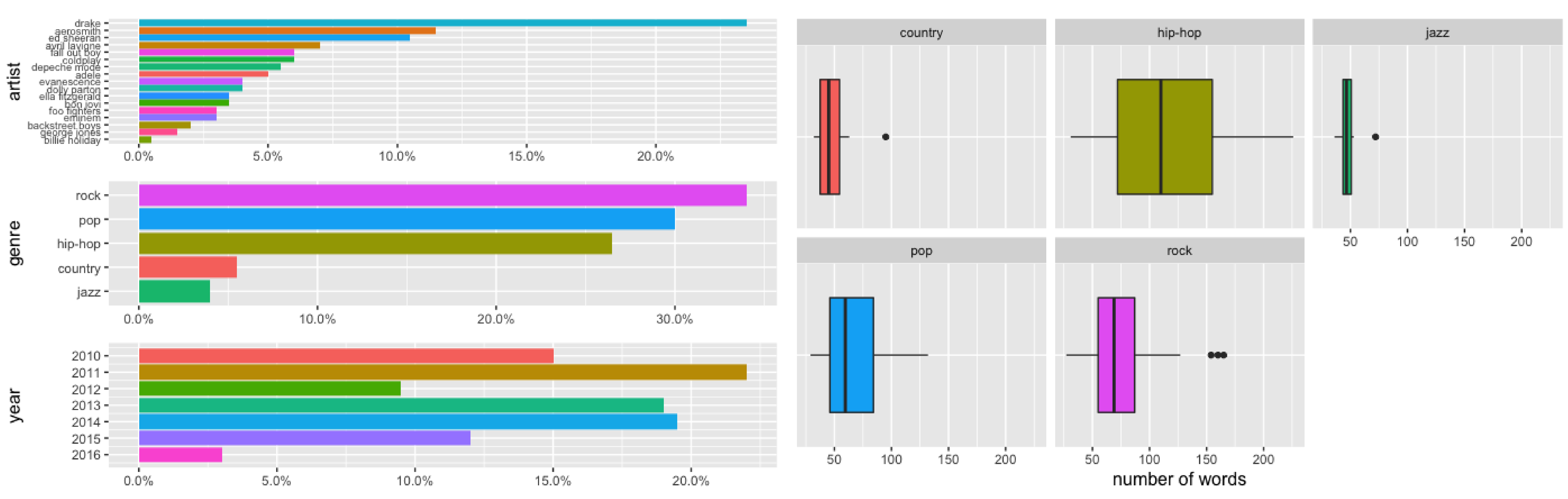
Methods to derive the appropriate number of topics:

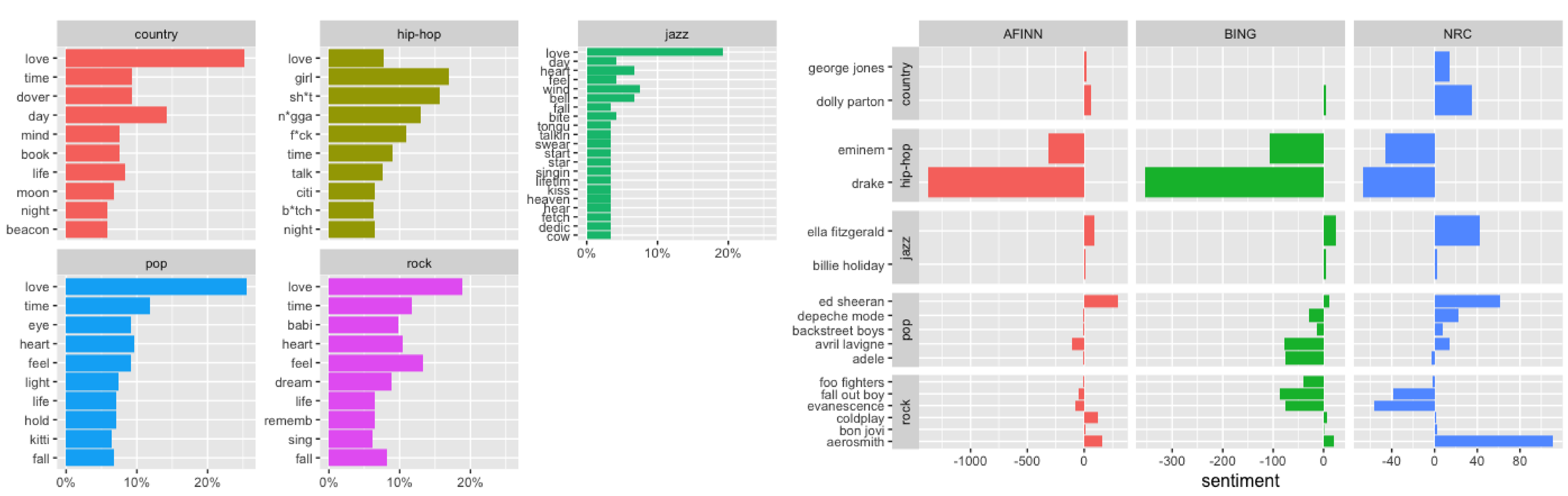
- Perplexity of hold-out set [1]
- Metric based on the average cosine distance among semantic clusters [2]
- Metric based on the information divergence between pairs of topics [3]
- Metric based on the divergence among stochastic matrices [4]
- Human judgment [5]

Interpretation

3. Dataset

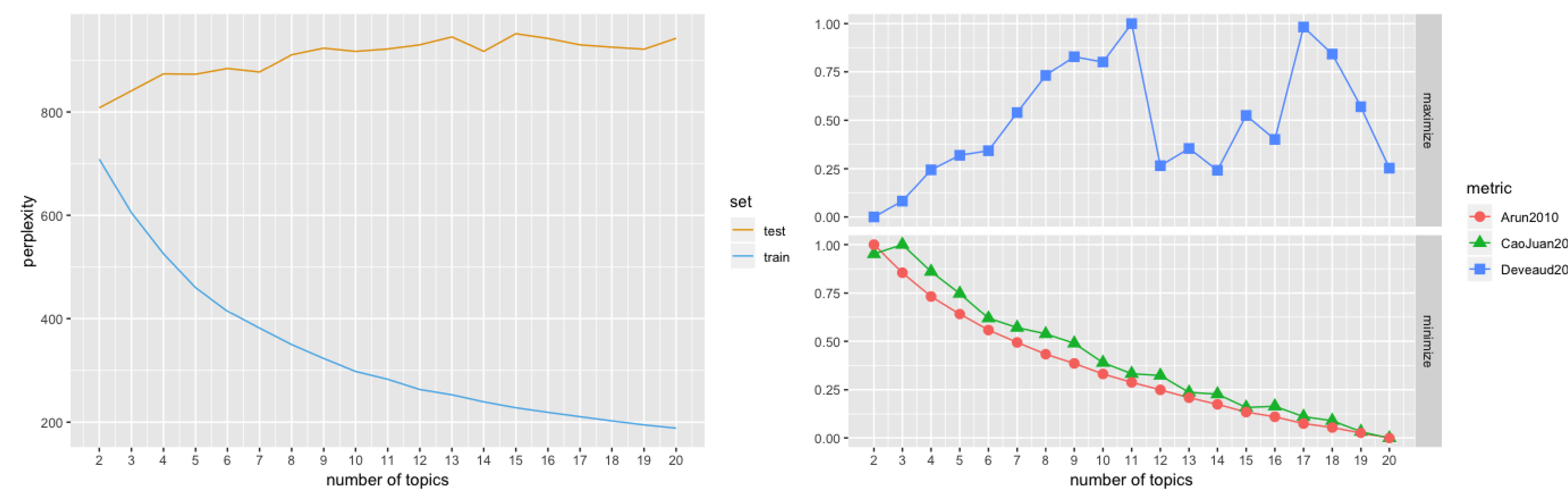
200 songs from 17 artists and 5 music genres between 2010 and 2016.



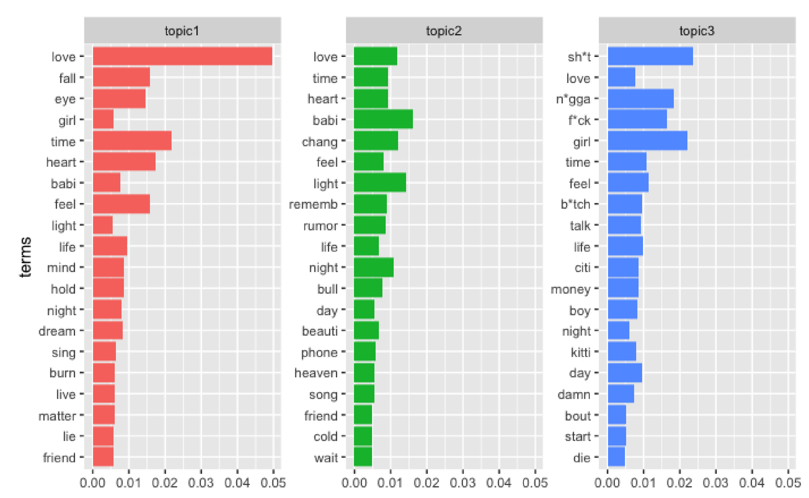


4. Implementation

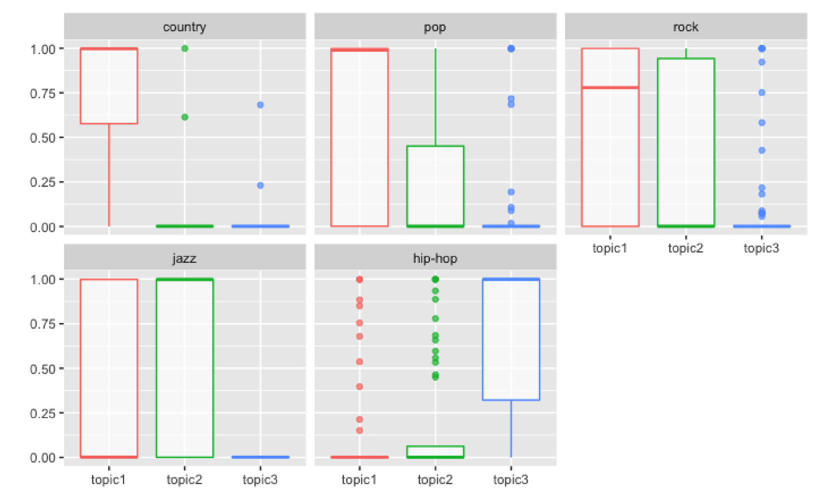
Number of topics



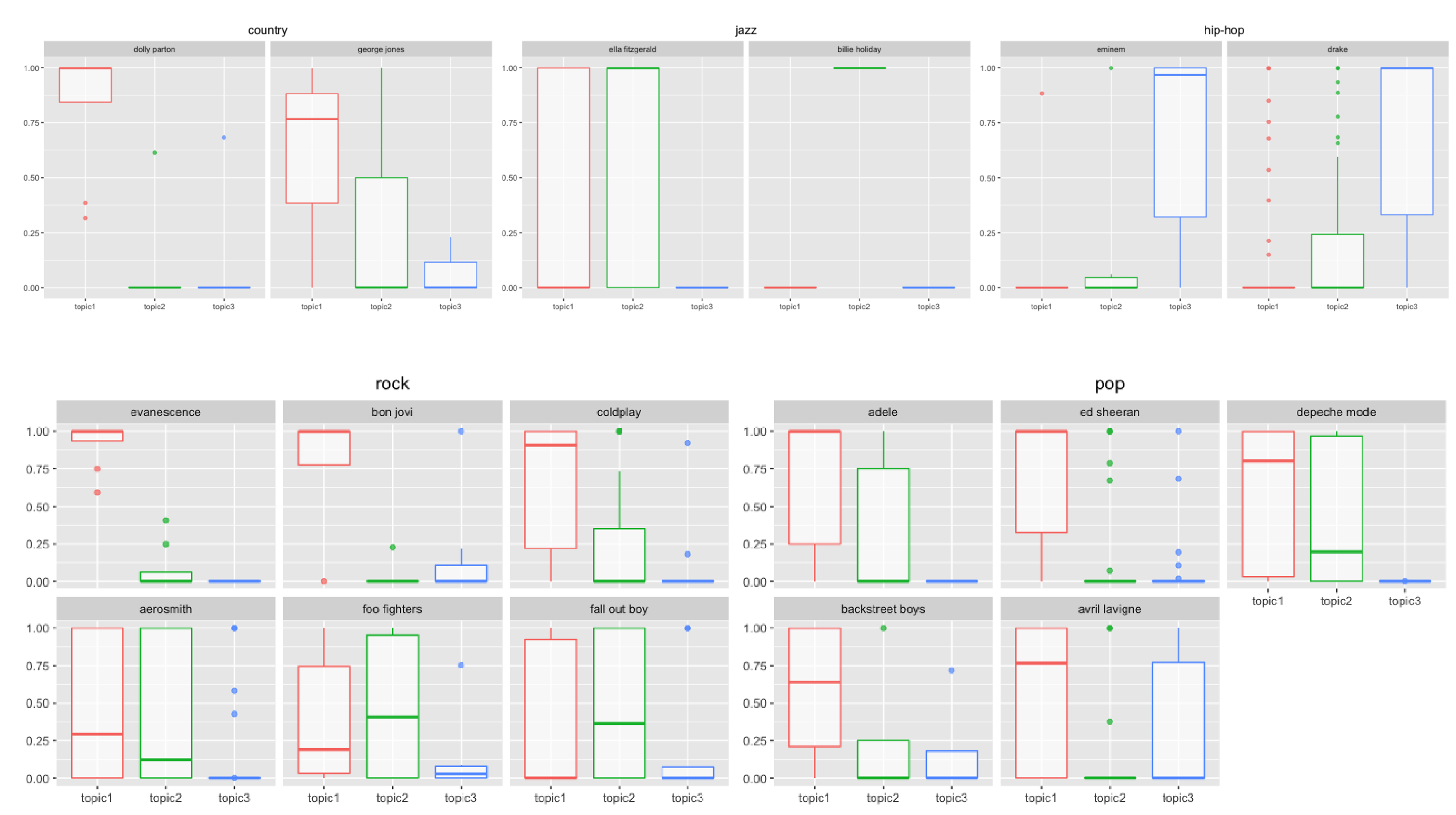
Per-topic-per-word probability



Per-genre-per-topic probability



Per-artist-per-topic probability



5. Results and conclusions

- According to LDA with VEM estimation, the dataset can be split into 3 topics:
  1. Romantic love: mostly Pop and Country songs, part of Rock ones
  2. Everyday life: primarily Jazz songs, part of Rock ones
  3. Race fight: mainly Hip-Hop songs [FIND BETTER WORD]
- There is, in general, no clear separation of topics inside genres and artists
- Increasing the number of topics improves the definition of topics for artists, but reduces the interpretability

References

[1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3 (Jan), 993–1022

[2] Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive lda model selection. Neurocomputing, 72 (7-9), 1775–1781.

[3] Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In Pacific-asia conference on knowledge discovery and data mining (pp. 391–402)

[4] Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. Document numérique, 17 (1), 61–84.

[5] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In Advances in neural information processing systems (pp. 288–296).

For further information

Silvia Ventoruzzo

silvia.ventoruzzo@gmail.com
Immatriculation number: 592252 (HU)

Freie Universität Berlin
Institute of Statistics and Econometrics
Seminar: Advanced Statistical Modeling