Prior Thoughts on Mixed-Membership Models in Linguistics

Chundra Cathcart [1] Gerd Carling [2]

[1] University of Zurich

[2] Lund University

7 April 2019

github.com/chundrac/Bayes-Lund2019

Talk Outline

Goals of talk:

- Give overview of some applications of mixed-membership models to linguistic questions
- Give high-level and technical description of mixed-membership models
- Discuss consequences of different model assumptions concretely using posterior predictive checks

Presentation with full references plus Jupyter notebooks available at URL in footer of slides

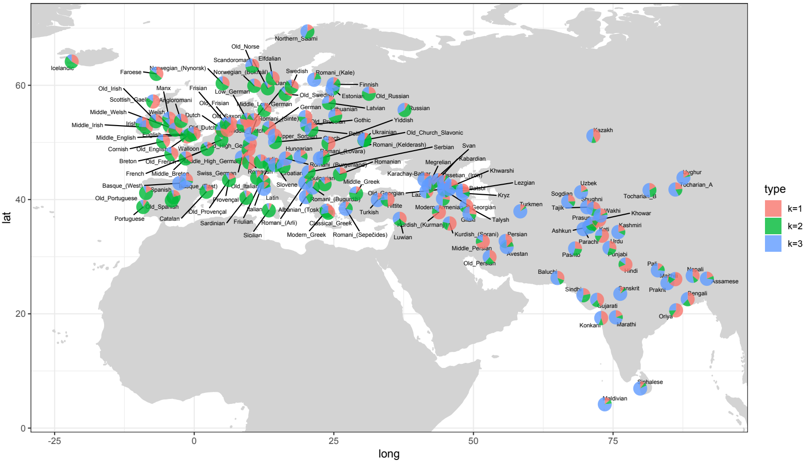
Computational Biology and Linguistics

- Like biological taxa, languages evolve and diversify over time
    - Phylogenetic inference methods can be applied to groups of languages to better understand relatedness
- Like biological organisms, languages display admixture, transferring linguistic features laterally
    - Mixed-membership models from population genetics can aid us in understanding historical contact between languages
- These methods have aided linguists in casting various questions in a probabilistic framework; at the same time, not all biological assumptions generalize well to linguistics

Mixed-membership models in Linguistics

- Structure (Pritchard et al. 2000): popular biological mixed-membership model designed to model genetic admixture between populations on the basis of allele frequencies at genetic loci
- Almost identical to Latent Dirichlet Allocation, designed for topic modeling in NLP

- This approach has been extended to problems in linguistics in order to disentangle relationships between different languages
- Applied to morphosyntactic data, i.e., data concerning grammatical patterns such as word order (Reesink et al. 2009)
- Applied to very diverse data sets (Syrjänen et al 2016)
- Cathcart forthcoming: uses a similar hierarchical model to analyze sound change patterns in Indo-Aryan dialects
- Basic idea: languages are analogous to individuals; linguistic features (such as word order) are analogous to genetic loci; linguistic feature variants (e.g., subject-object-verb word order) are analogous to alleles



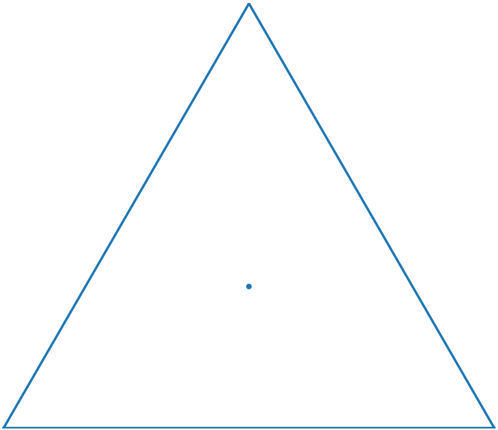
github.com/chundrac/Bayes-Lund2019

Multinomial probability distributions

- Linguistic data sets of the sort analyzed with structure contain CATEGORICAL DATA
- A single linguistic feature (e.g., word order) has two or more variants that can be expressed (e.g., SOV, SVO, VSO, V2, etc.)
- It makes sense that the occurrence of feature variants in languages should be modeled with the multinomial distribution, but what type of multinomial distribution?
- The 3-simplex provides an intuitive visual representation of this question
    - A single coordinate partitions the 3-simplex into regions of probability mass corresponding to different events

Smooth multinomial distribution

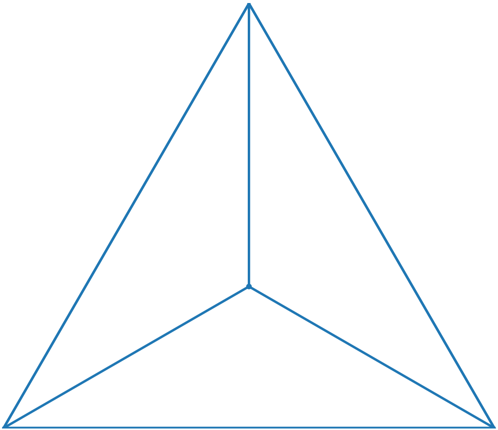
$P(A) = P(B) = P(C) = 1/3$



github.com/chundrac/Bayes-Lund2019

Smooth multinomial distribution

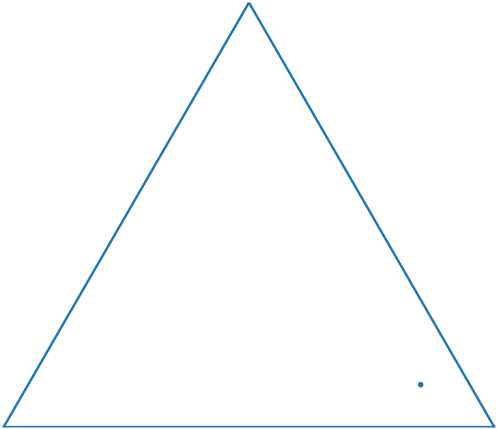
$P(A) = P(B) = P(C) = 1/3$



`github.com/chundrac/Bayes-Lund2019`

Sparse multinomial distribution

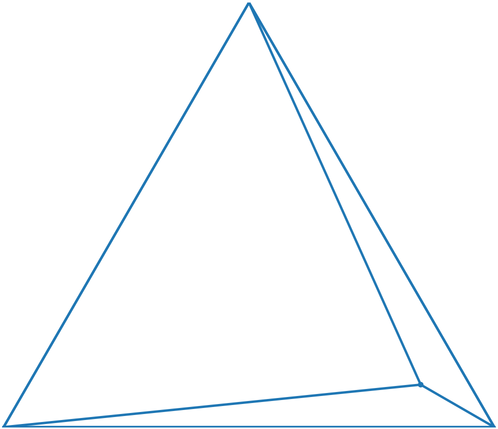
One event dominates in terms of probability mass



github.com/chundrac/Bayes-Lund2019

Sparse multinomial distribution

One event dominates in terms of probability mass



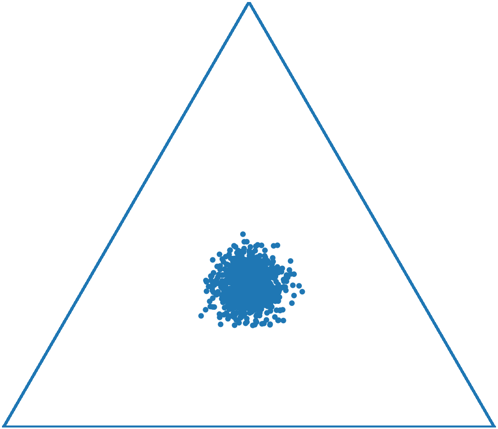
github.com/chundrac/Bayes-Lund2019

Dirichlet distribution

- Popular prior over multinomial distributions
- Parameterized by CONCENTRATION PARAMETERS
- Symmetric Dirichlet has only one concentration parameter α
- Crucially, α doesn't say WHERE probability mass is concentrated, but HOW it is allocated across outcomes

Dirichlet distribution: smooth draws

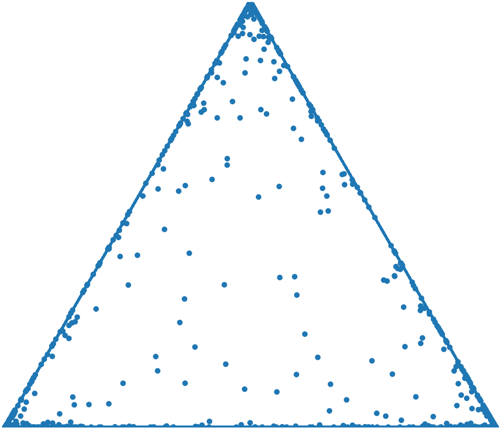
- If $\alpha > 1$, a symmetric Dirichlet distribution generates SMOOTH MULTINOMIAL DISTRIBUTIONS
- 1000 samples from $\mathrm{Dir}(50, 50, 50)$:



github.com/chundrac/Bayes-Lund2019

Dirichlet distribution: sparse draws

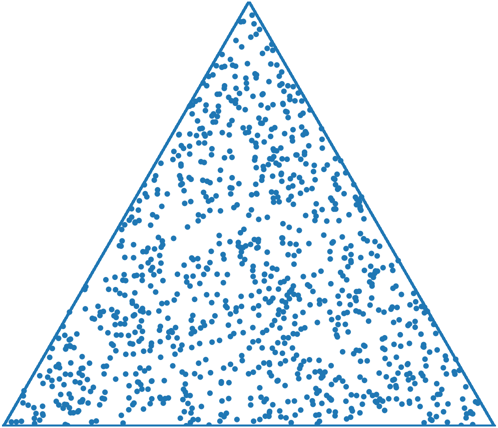
- If $\alpha < 1$, a symmetric Dirichlet distribution generates SPARSE MULTINOMIAL DISTRIBUTIONS
- 1000 samples from Dir$(.1, .1, .1)$:



github.com/chundrac/Bayes-Lund2019

Uniform Dirichlet distribution

- If $\alpha = 1$, a symmetric Dirichlet distribution generates both sparse and smooth multinomials with equal probability
- 1000 samples from $\text{Dir}(1, 1, 1)$:



github.com/chundrac/Bayes-Lund2019

Assumptions of mixed-membership models (with respect to languages)

- Each linguistic feature in each language is inherited from one of K LATENT ANCESTRAL POPULATIONS, CLUSTERS or COMPONENTS
- Each component has a distribution over linguistic features associated with it
- This distribution usually takes the form of a COLLECTION OF MULTINOMIALS, i.e., a set of ragged multinomial distributions of different length each governing how a particular part of the grammatical domain expresses itself, e.g.,
  - Main clause word order: SVO, SOV, V2, or VSO?
  - Order of noun and relative clause: Rel-N or N-Rel?

Generative process under mixed-membership model

- Fix K, the number of latent components
- Draw θ, each language's distribution over latent components, from some prior distribution
- Draw ϕ, each component's distribution over linguistic features, from some prior distribution
- For each language l
    - For each linguistic feature f
        - Draw $z_{l,f} \sim \text{Cat}(\theta_l)$, the component label associated with the current feature
        - Draw $y_{l,f} \sim \text{Cat}(\phi_{z_{l,f},f})$, the observed feature variant from the feature distribution associated with the label sampled in the previous step

- The generative story describes how we THINK the data were generated.
- We observe the data, but we don’t know the parameter values
- We need to invert the generative process to infer the relevant unknown quantities

- Three unknown quantities: $\phi, \theta, \boldsymbol{z}$
- In general, if we know the continuous variables ϕ, θ, we can reconstruct $\boldsymbol{z}$, and if we know $\boldsymbol{z}$, we can sample ϕ, θ
- We don't need to infer all three parameters. Two approaches:
    - Marginalize out $\boldsymbol{z}$: required for many probabilistic programming languages (including Stan) which are gradient based; discrete parameters are not differentiable
    - Marginalize out ϕ, θ: allows for Collapsed Gibbs Sampling, if Dirichlet priors are used

Priors on ϕ, θ

- Structure uses Gibbs Sampling, places symmetric Dirichlet priors over ϕ, θ
  - $\theta \sim \text{Dirichlet}(\alpha)$
  - $\phi \sim \text{Dirichlet}(\lambda)$
- The hyperparameter $\alpha \sim U(0, 10)$ is inferred from the data
- The hyperparameter λ is fixed
- The default value for λ is 1, though the authors emphasize that this setting is merely a default

Output: posterior distribution over component label assignment configurations

How many K?

- For a given implementation, K, the number of components assumed is fixed, but in reality this is an unknown
- Standard procedure: model selection according to model marginal likelihood
- However, no clear consensus among practitioners regarding some details of model selection (cf. Evanno et al. 2005)
- Model selection of this sort is inimical to the spirit of “continuous model expansion” advocated by Gilman and Shalizi (2013)
- Non-parametric alternatives (e.g., the Hierarchical Dirichlet Process) address this issue, but make problematic assumptions of their own
- We sidestep this issue in today’s talk...

Key Issues

- Researchers applying the Structure method to linguistic questions don't say much about the hyperparameters they use
- Presumably they are using the default settings
- But is it appropriate to assume that $\lambda = 1$?
- As researchers go further with real-world interpretations of results, there is an increasing need to consider the consequences of prior choices

Case in point

- Honkola et al. (2018) build off of the results from Syrjänen et al. (2016)
- They use the inferred clusters (ICs) for all of the Finnish dialects in their survey as a proxy for linguistic information
- Key issue: “transitional” dialects (i.e., speech varieties with a flatter distribution over components) are discarded
- However, the overall uncertainty regarding a speech variety’s component makeup may be highly dependent on the value of λ

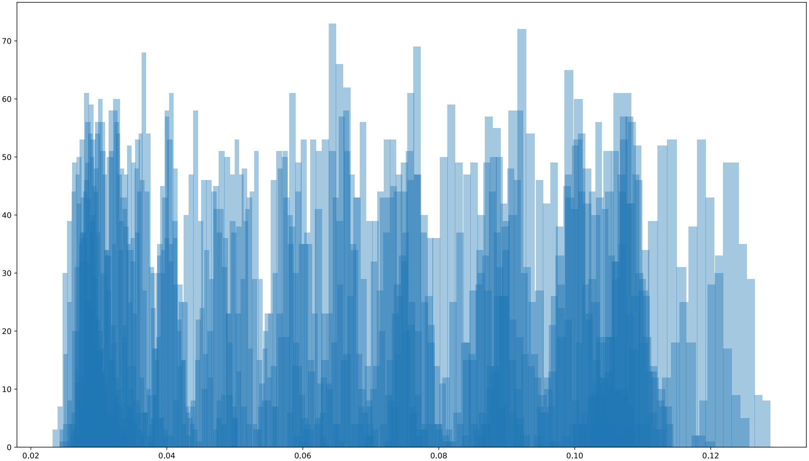
Experiment

- We carry out some experiments designed to assess the consequences of varying λ
- We use the DiACL Eurasia data set (Carling 2017), which contains fine-grained grammatical data from languages of Eurasia; we exclude ancient and medieval languages
- For $K \in \{2, ..., 20\}$, we run three inference regimes:
  - Uniform: $\lambda = 1$
  - Sparse: $\lambda = .1$
  - Inferred: λ inferred from data
- Inference carried out using Tensorflow Probability's Hamiltonian Monte Carlo
- 4 chains, 10000 iterations, first 2000 samples discarded

`github.com/chundrac/Bayes-Lund2019`

Posterior distributions of inferred λ

For $K \in \{2, ..., 20\}$

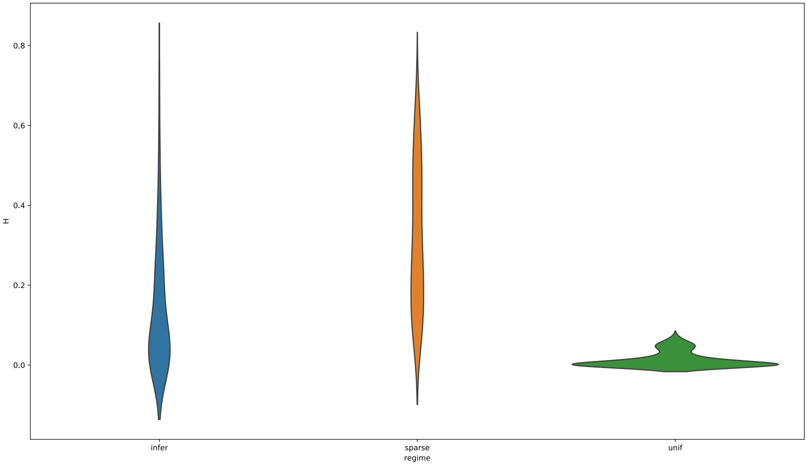


github.com/chundrac/Bayes-Lund2019

Model Criticism

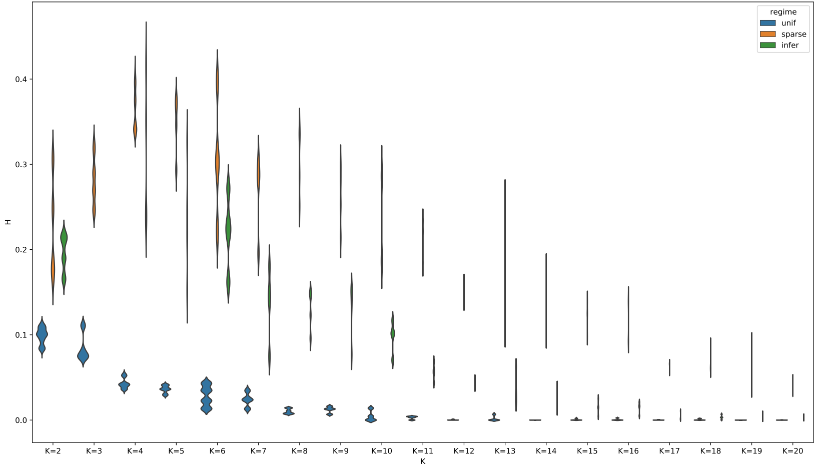
- We carry out model criticism using the following posterior predictive checks:
    1. We assess the average uncertainty (i.e., entropy) of $\hat{\theta}$, the posterior language-level component distributions, for each regime
    2. We compute the average uncertainty (i.e., entropy) in assignment of component labels to each data point, i.e., $P(\boldsymbol{z})$ (cf. Mimno et al. 2015)
    3. We compute the accuracy of data simulated with posterior parameters

Uncertainty over $\hat{\theta}$



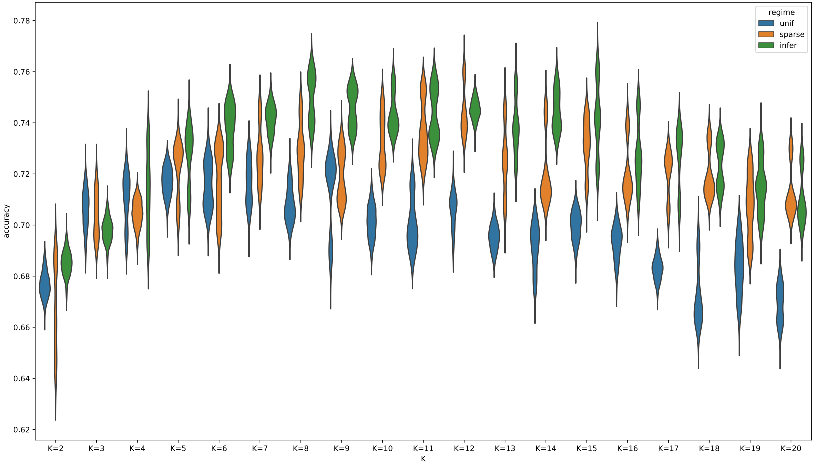
github.com/chundrac/Bayes-Lund2019

Uncertainty over $P(\boldsymbol{z})$



github.com/chundrac/Bayes-Lund2019

Accuracy



github.com/chundrac/Bayes-Lund2019

Concluding remarks

- Fixing λ to generate uniform or smooth draws from the Dirichlet distribution results in lower entropy for $\hat{\theta}$, and reduces uncertainty in component label assignment
- However, inferred values for λ tend to be lower than 1
- Furthermore, sparse and inferred values for λ outperformed uniform/smooth λ in terms of accuracy
- More work is needed to determine exactly which hyperparameter specifications are needed to truly capture the vagaries of different linguistic data sets
- Moving from biological software packages to probabilistic programming languages will allow linguists to fit more flexible models, and can help expand the inventory of prior distributions used (e.g., to the logistic normal distribution)

`github.com/chundrac/Bayes-Lund2019`

Acknowledgements

Thanks to contributors to the DiACL data set, and to Erich Round for helpful discussion!

github.com/chundrac/Bayes-Lund2019