Using topic models to describe disturbance & quantify responses to environmental change

Gavin L. Simpson & Emma Wiik ▪ University of Regina

ISEC 2018 St Andrews ▪ July 5th 2018

Acknowledgements

NSERC
CRSNG

Slides: **bit.ly/isectopicmodels**

Community response to environmental change

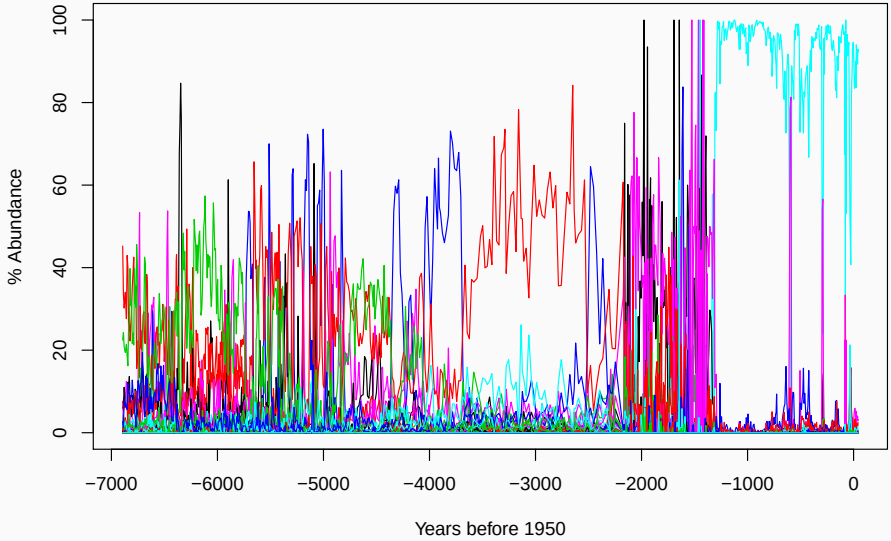
By Mauricio Antón CC BY 2.5, via Wikimedia Commons

Community response to environmental change

By Credit: NOAA George E. Marsh Album [Public domain], via Wikimedia Commons

How have aquatic communities responded to these rapid changes over the last few millennial?

Complex multivariate species data



Dimension reduction

Typically we can't model all 100+ taxa in data sets like this

- (M)ARSS-like models don't like the large n

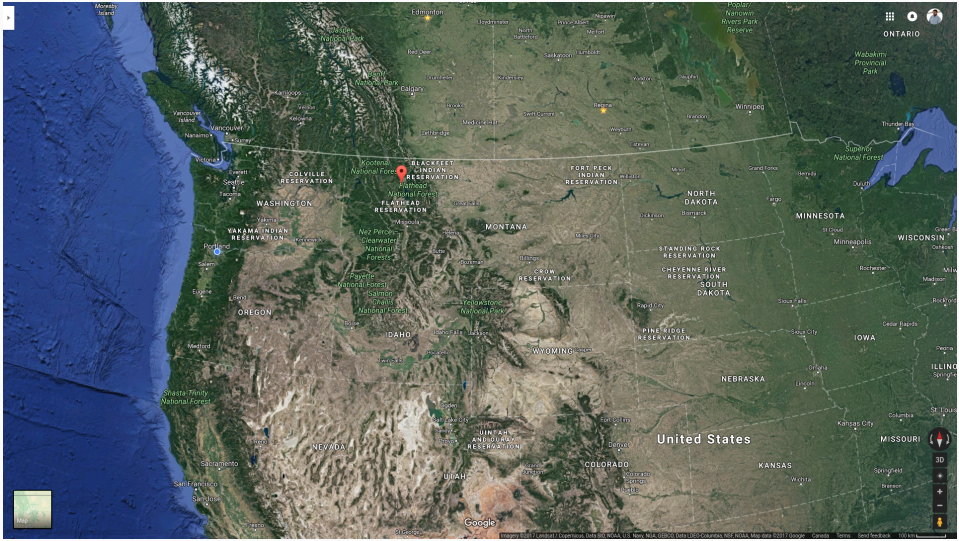
Seek a reduced dimensionality of the data that preserves the signal

Existing dimension reduction methods aren't appropriate for questions we want to ask

- Interpretation of latent factors is complex (PCA, CA, Principal Curves)

Can we group species into J associations and soft cluster samples as compositions of these associations?

Foy Lake — Montana



Foy Lake — Montana

Topic models

Machine learning approach for organizing text documents

- Latent Dirichlet Allocation (LDA) — (Blei, Ng, & Jordan, *J. Mach. Learn. Res.* 2003)

Generative model for word occurences in documents

- Valle, Baiser, Woodall, & Chazdon, R. (2014) *Ecology Letters* **17**
- Christensen, Harris, & Ernest. (2018) *Ecology* `doi:10.1002/ecy.2373`

Community of Skittles

Individual skittles from one of four flavoured packs

What are the proportion of flavours in each pack?

How many of each pack comprise the skittle community?

Latent Dirichlet Allocation

Aim is to infer the

- distribution of skittle **flavours** within each pack β_j, and
- distribution of skittle **packs** within each community (*sample*)

Achieve a soft clustering of samples — mixed membership model

Achieve a soft clustering of **species** (flavours) into **associations** of taxa (packs)

User supplies J — the number of associations *a priori* — $j = \{1, 2, \ldots, J\}$

J chosen using AIC, *perplexity*, CV, ...

Latent Dirichlet Allocation

1. Flavour distribution for jth type of Skittle

$$\beta_j \sim \text{Dirichlet}(\delta)$$

2. Proportions of each type in the Skittle community

$$\theta \sim \text{Dirichlet}(\alpha)$$

3. For each skittle s_i
   - Choose a pack in proportion

$$z_i \sim \text{Multinomial}(\theta)$$

   - Choose a flavour from chosen pack with probability

$$p(s_i|z_i, \beta_j) \sim \text{Multinomial}(\delta)$$

Correlated Topic Model

LDA assumes associations of species are **uncorrelated**

Potentially more *parsimonious* & *realistic* if associations were correlated

2. Proportions of each type in the Skittle community — draw

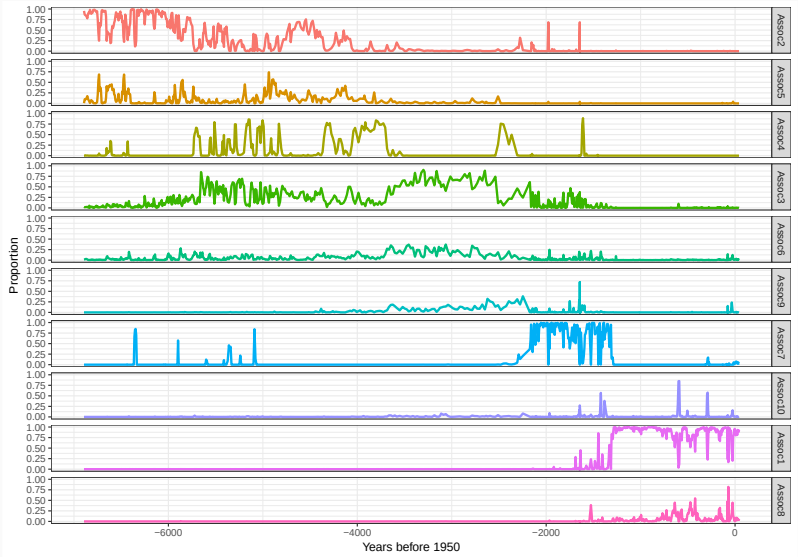
$$\eta \sim N(\mu, \Sigma)$$

with $\eta \in \mathbb{R}^{J-1}$ and $\Sigma \in \mathbb{R}^{(J-1)\times(J-1)}$
Then transform η_J to proportional scale

Σ controls the correlation between *associations*

Blei & Lafferty (2007) A correlated topic model of Science. *Ann. Appl. Stat.* **1**, 17–35.

Correlated Topic Model



Latent Dirichlet Allocation — Trend estimation

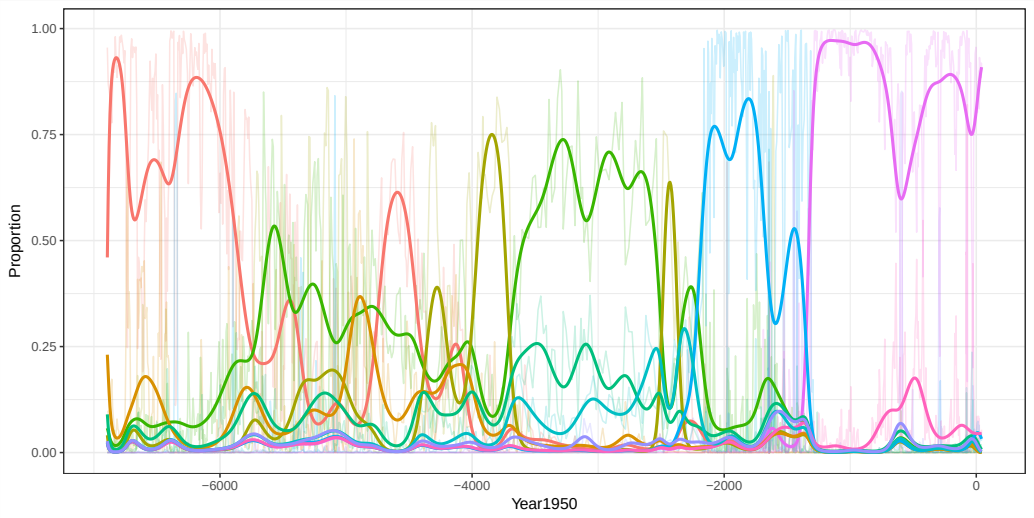
LDA knows nothing about the temporal ordering of the samples

Estimate trends in proportions of species associations using a GAM

- use adaptive spline to allow for rapid adaptation to changing data
- model each association as $\sim$ Beta(μ, ϕ)

Other methods would also be appropriate: eg. Bayesian change point model or Dirichlet regression

Correlated Topic Model — Trend estimation (Dirichlet Regression)



Summary

LDA & CTM proved well-capable of summarizing the complex community dynamics of Foy Lake

- Reduced 113 taxa to 10 associations of species
- Species associations match closely the expert interpretation of the record
  - make autecological sense also
- The CTM was more parsimonious — removed one rare association
- Estimated trends in proportions of species associations capture mixture of
  - smooth, slowly varying trends, and
  - rapid (regime shift?) state change $\sim$ 1.3 ky BP

Future directions

Choosing J is inconvenient

Address this via **Hierarchical Dirichlet Processes** and Bayesian Nonparametrics

- assume J is infinite & put a prior distribution over J

Associations in LDA & CTM are static — distributions are fixed for all samples

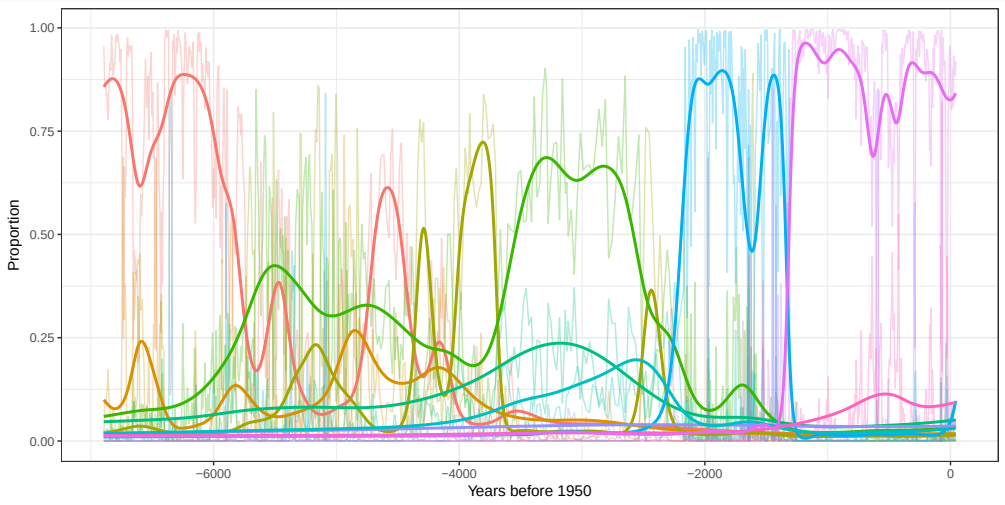
- dynamic & structural topic models allow distributions to vary smoothly with time

Many developments in this field:

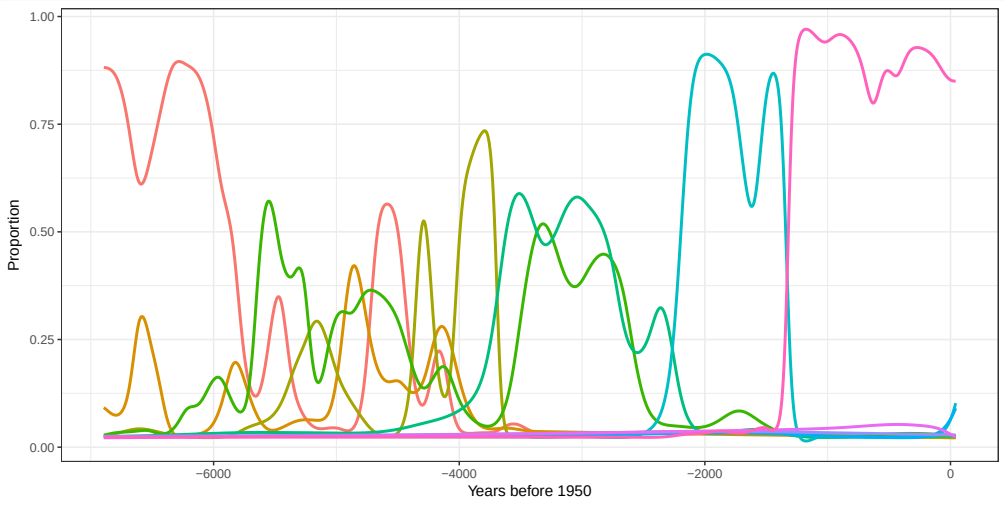
- *Chinese Restaurant Process*,
- *Indian Buffet Process*, &
- ...

Extra slides. . .

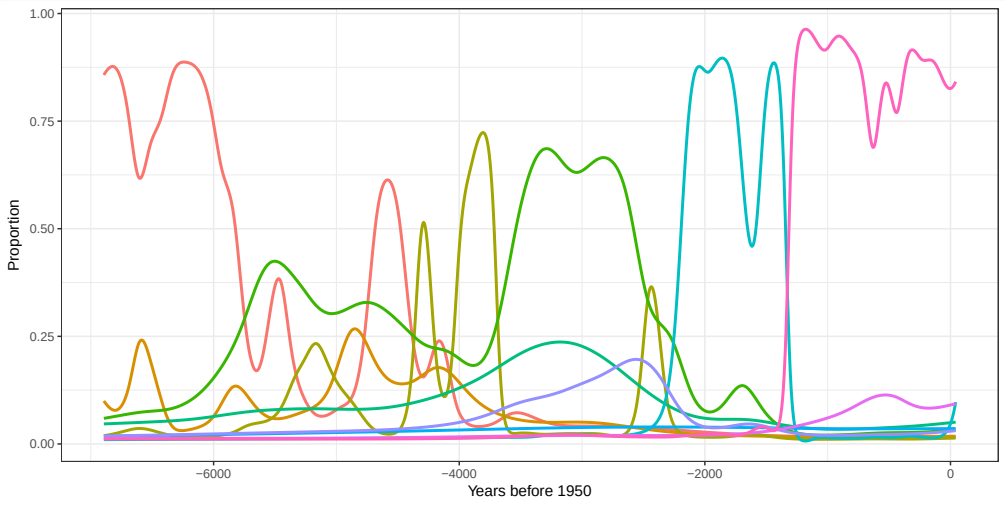
Correlated Topic Model — Trend estimation (Adaptive GAM)



Latent Dirichlet Allocation — Trend estimation (AdaptiveGAM)



Correlated Topic Model — Trend estimation (Adaptive GAM)



Intuition behind LDA

Latent Dirichlet allocation represents a trade-off between two goals

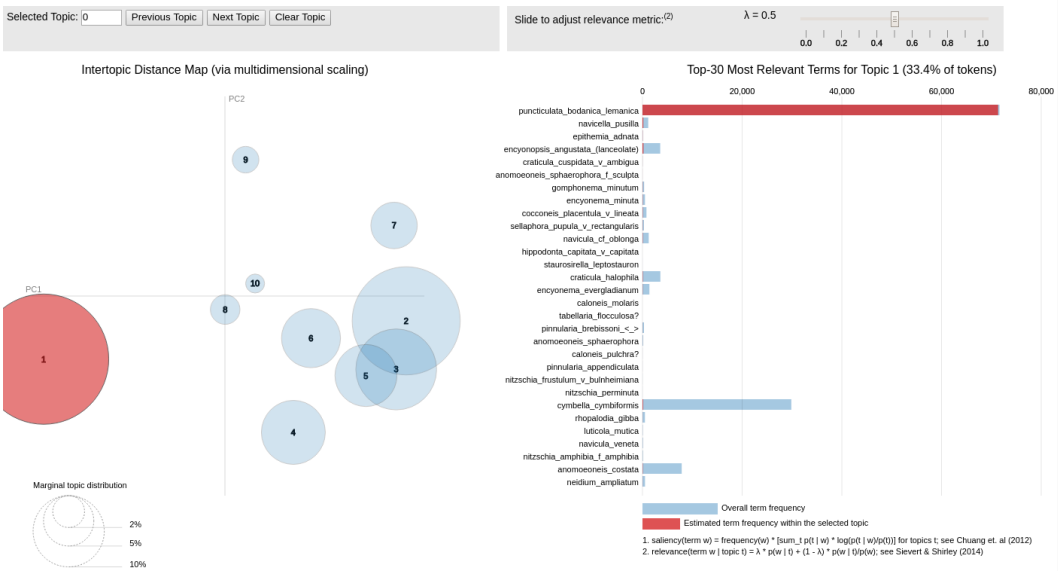
1. for each sample, allocate its individuals to a **few associations of species**
2. in each association, assign high probability to a **few species**

These are in opposition

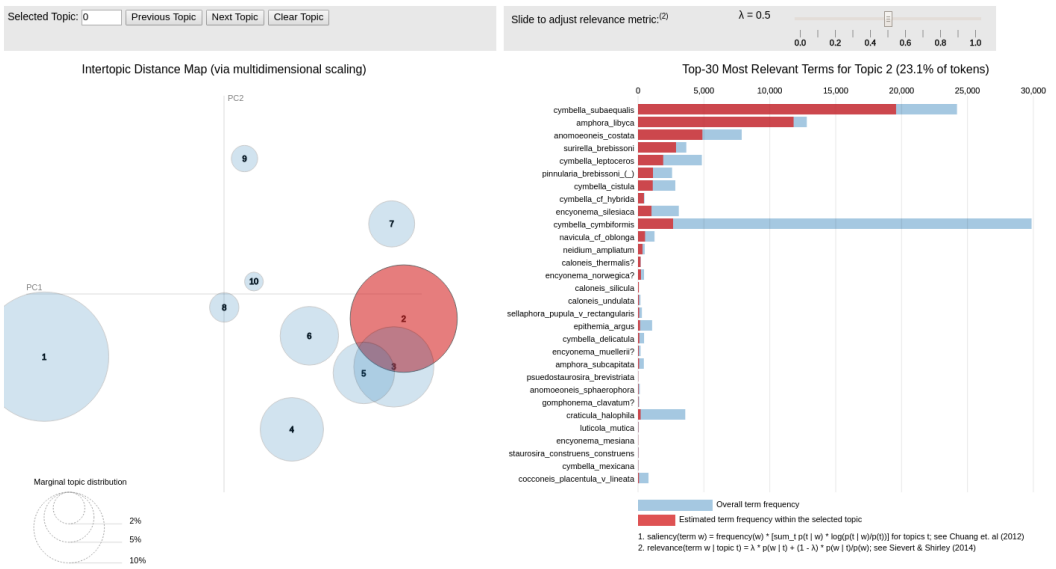
- assigning a sample to a single association makes *2* **hard** — all its species must have high probability under that one topic
- putting very few species in each association makes *1* **hard** — to cover all individuals in a sample must assign sample to many associations

Trading off these two goals therefore results in LDA finding tightly co-occurring species

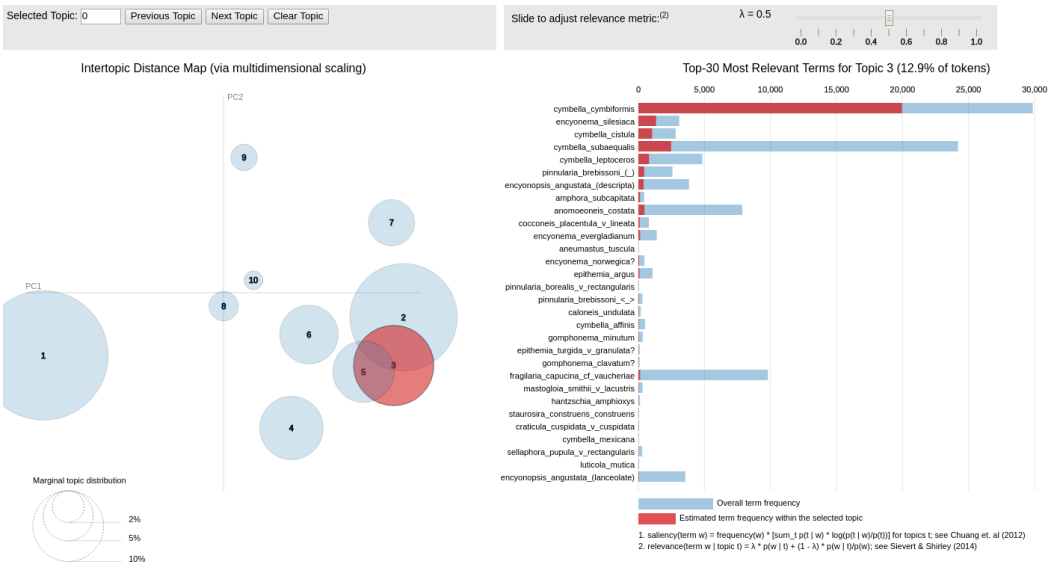
Latent Dirichlet Allocation — Association 1



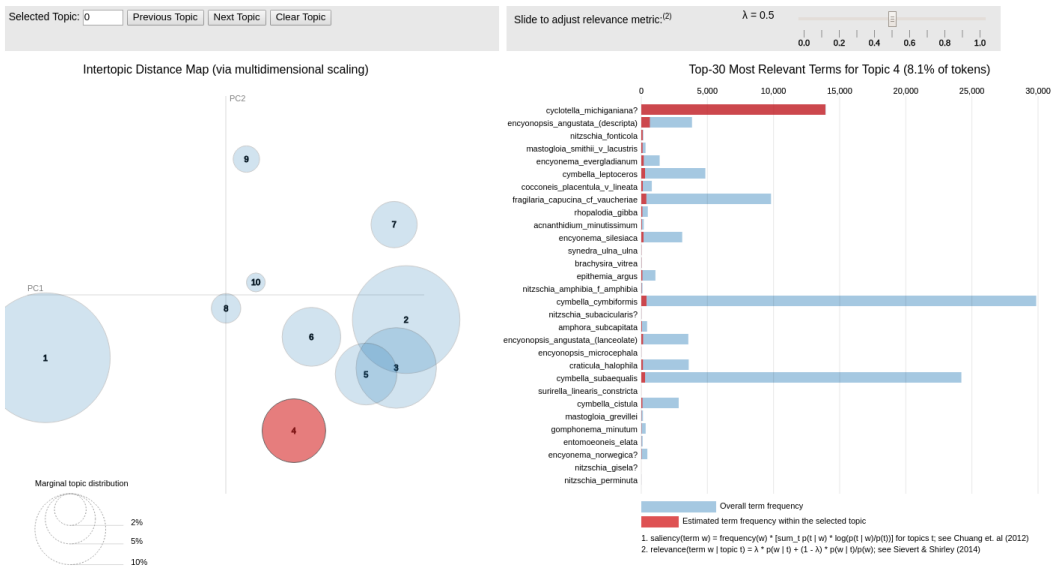
Latent Dirichlet Allocation — Association 2



Latent Dirichlet Allocation — Association 3



Latent Dirichlet Allocation — Association 4



Foy Lake

- Deep, freshwater lake
- Drought-sensitive Flathead River Basin
- Diatom assemblages sensitive to lake depth variation
- Related to variability in effective moisture

Regime shift ~1.3ka BP

Spanbauer *et al* PLOS ONE **9**(10) e108936