2018年6月15日　統計数理研究所　オープンハウス

計量政治学における統計的テキスト分析

持橋大地　数理・推論研究系 准教授　*daichi@ism.ac.jp*

本研究は, 佐々木智也氏・江島舟星氏 (東京大学 法学政治学研究科 D3・D1) との共同研究です。

Bayesian Changepoint Model for Text Data

Tomoya Sasaki
with Daichi Mochihashi

January 8th 2018

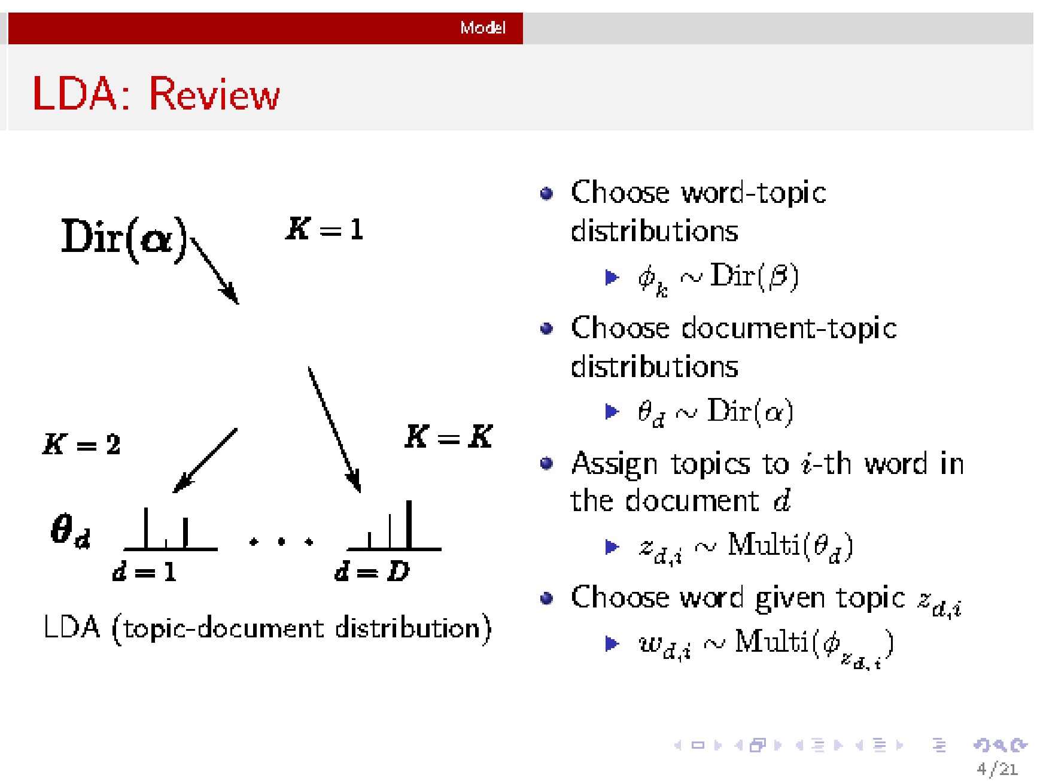
Motivation

- Changepoint analysis: identify the changes or shifts in preferences or behaviors of political actors
  - For count data, binary data, continuous data
- Some changes can be identified throught text-data
  - Party position from manifesto
  - Relationships between countries from treaties
- Proposed model: Bayesian changepoint model for text data, based on topic model

Motivation

The proposed model ...

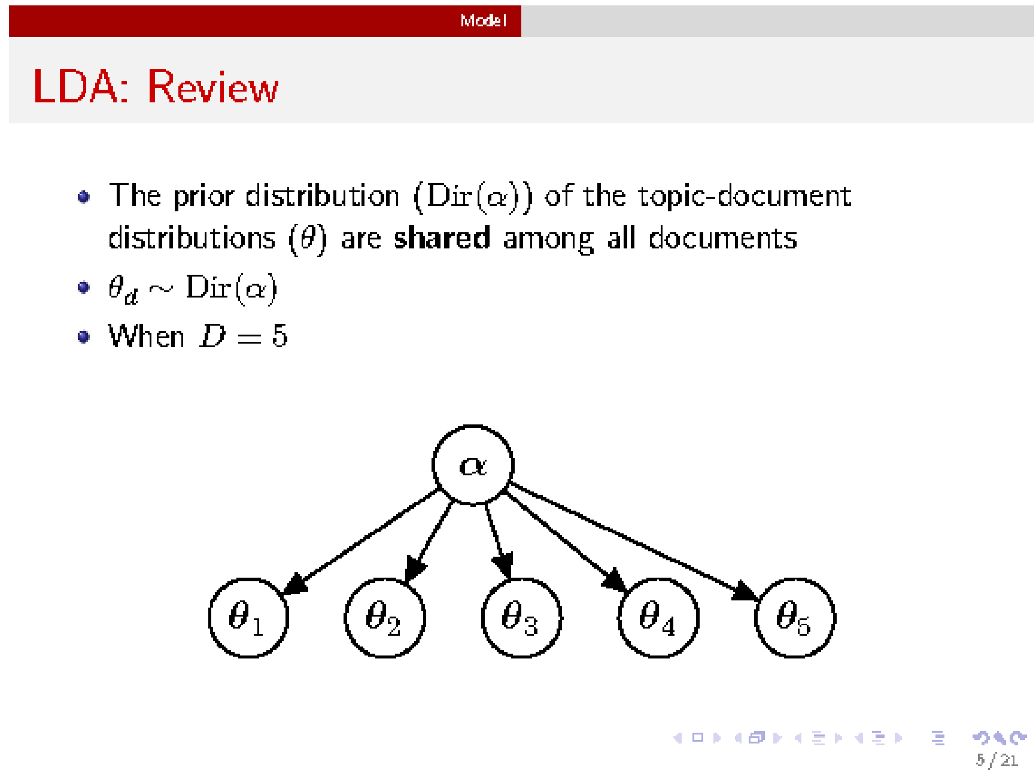
- simultaneously estimates
  - topics
  - the location of changepoints
  - the number of changepoints
- considers multiple latent traits in given documents
- shows which topics (characteristics) contribute to changes



LDA: Review

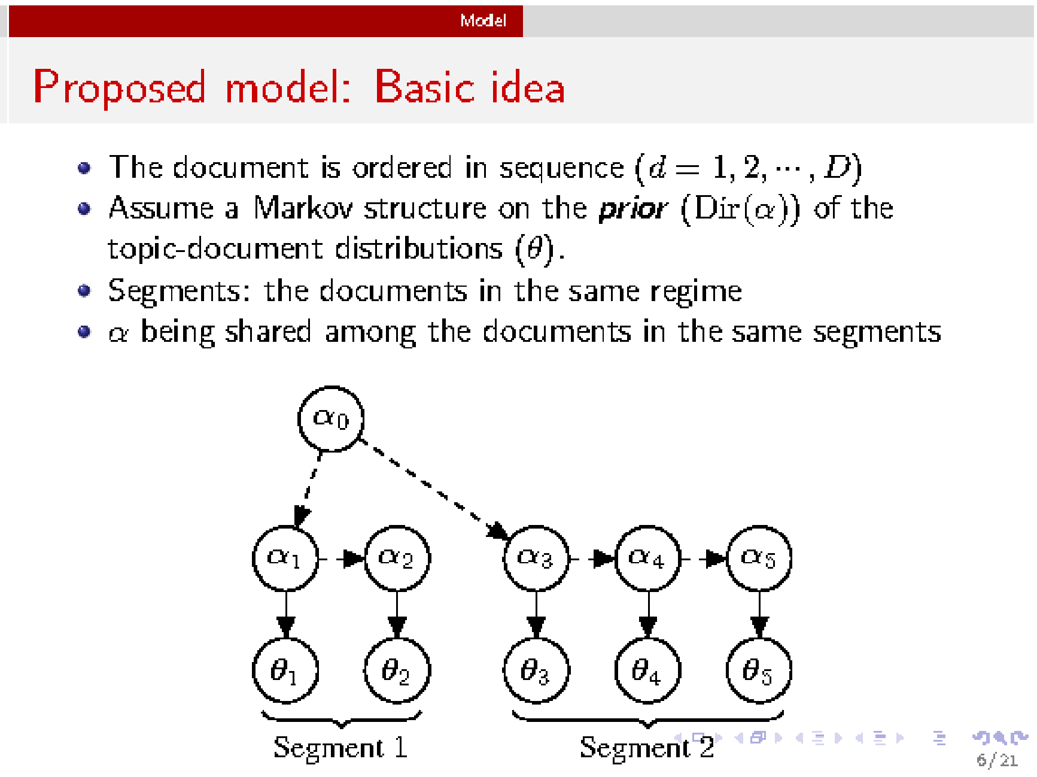
LDA (topic-document distribution)

- Choose word-topic distributions
  - $\phi_k \sim \mathrm{Dir}(\beta)$
- Choose document-topic distributions
  - $\theta_d \sim \mathrm{Dir}(\alpha)$
- Assign topics to i-th word in the document d
  - $z_{d,i} \sim \mathrm{Multi}(\theta_d)$
- Choose word given topic $z_{d,i}$
  - $w_{d,i} \sim \mathrm{Multi}(\phi_{z_{d,i}})$



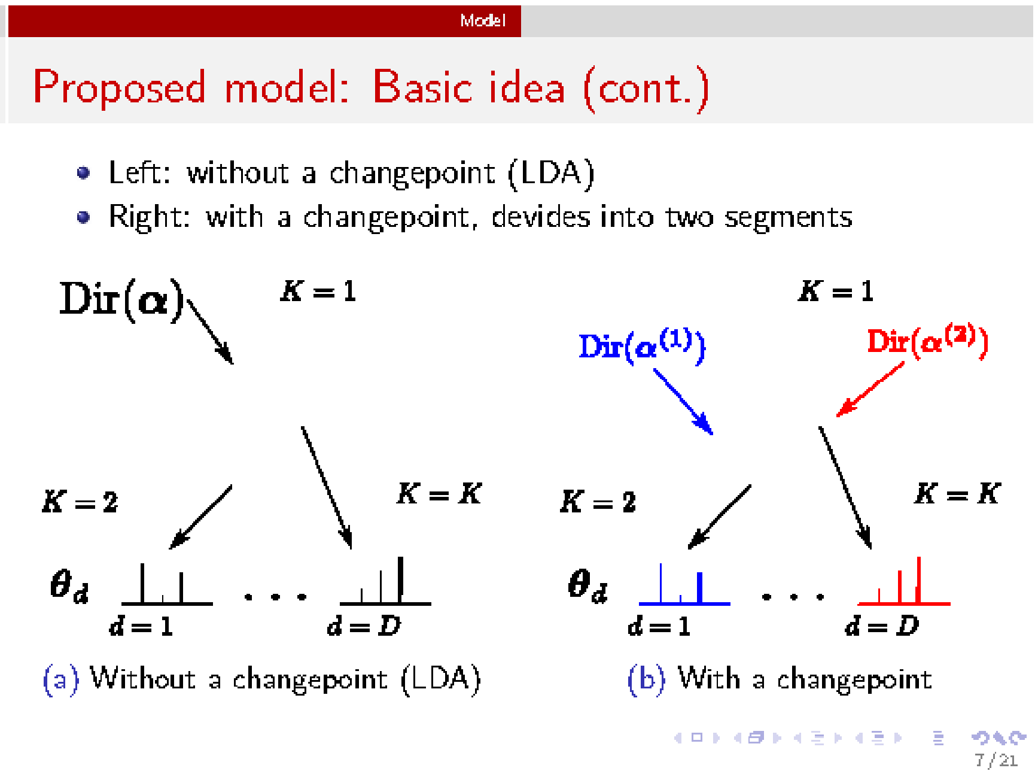
LDA: Review

- The prior distribution ($\mathrm{Dir}(\alpha)$) of the topic-document distributions (θ) are **shared** among all documents
- $\theta_d \sim \mathrm{Dir}(\alpha)$
- When $D = 5$



Proposed model: Basic idea

- The document is ordered in sequence ($d = 1, 2, \cdots, D$)
- Assume a Markov structure on the ***prior*** ($\mathrm{Dir}(\alpha)$) of the topic-document distributions (θ).
- Segments: the documents in the same regime
- α being shared among the documents in the same segments



Proposed model: Basic idea (cont.)

- Left: without a changepoint (LDA)
- Right: with a changepoint, devides into two segments

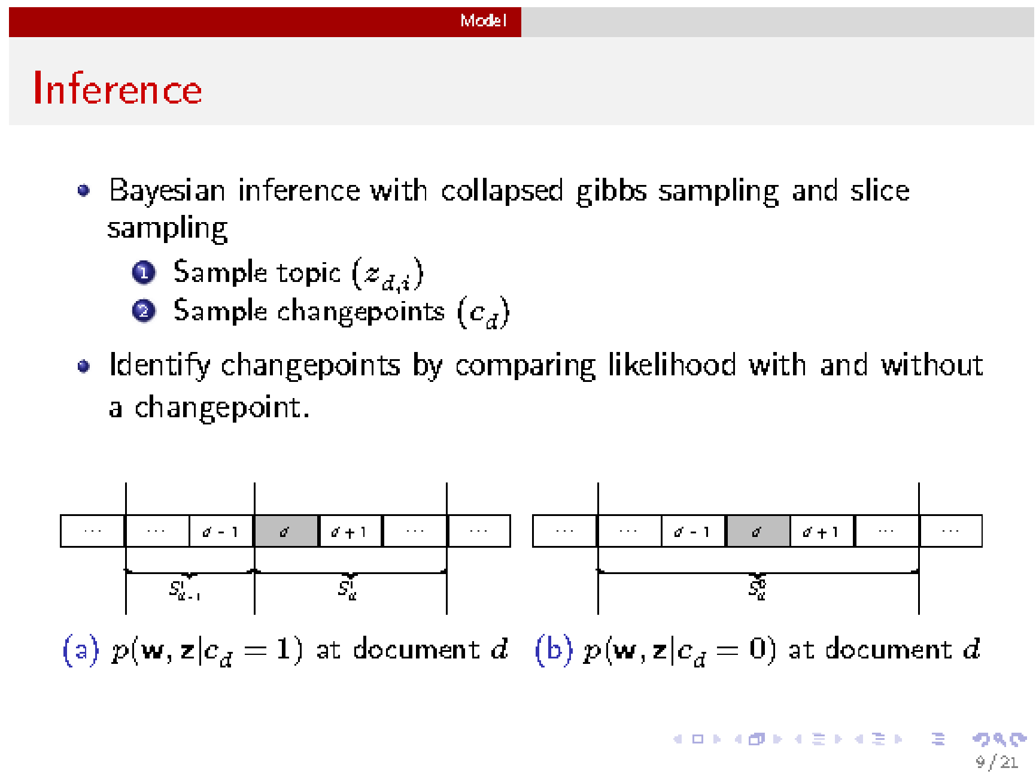
(a) Without a changepoint (LDA)　(b) With a changepoint

Proposed model: Basic idea (cont.)

The data generating process (specific to our model)

- Draw the probability of obtaining changepoints
  - $\pi \sim \mathrm{Beta}(\gamma_1, \gamma_2)$.
- For each document $d = 1, \cdots, D$,
  - draw an indicator of changepoint $c_d \sim \mathrm{Bern}(\pi)$,
  - if $c_d = 1$, select hyper-parameter $\alpha_{d,k} \sim \mathrm{Gamma}(\eta_1, \eta_2)$,
  - if $c_d = 0$, select hyper-parameter $\alpha_d = \alpha_{d-1}$,
  - choose a topic-document distribution $\theta_d \sim \mathrm{Dir}(\alpha_d)$.

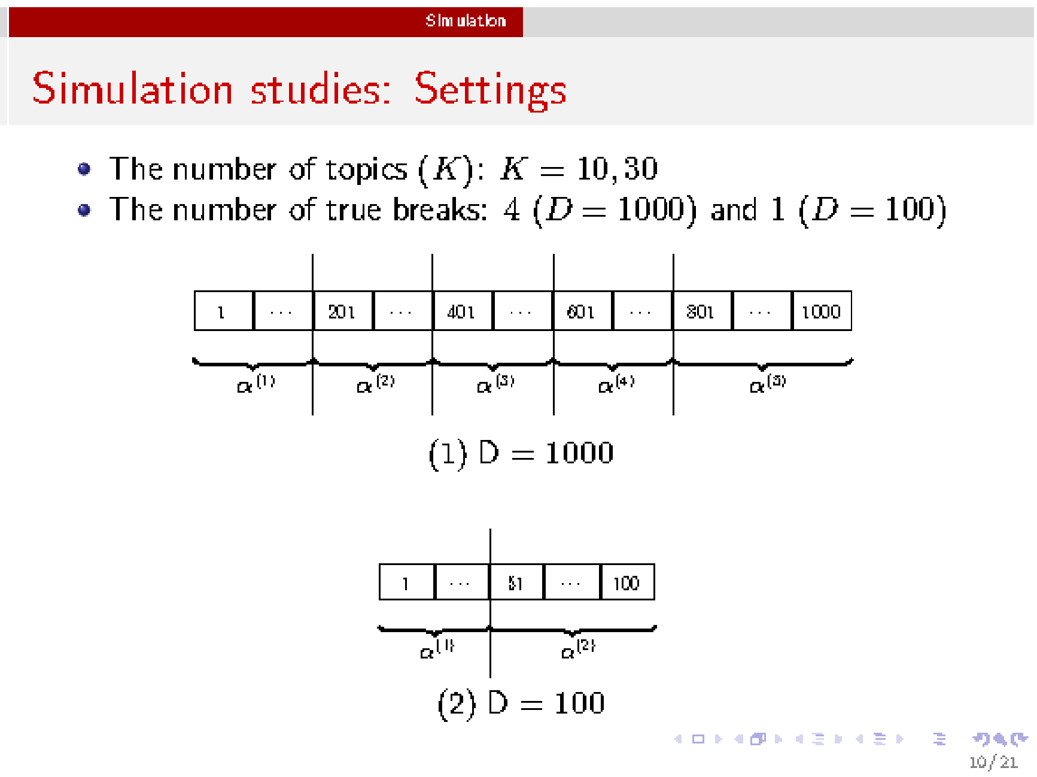
The rest is the same as LDA (word assignment part).



Inference

- Bayesian inference with collapsed gibbs sampling and slice sampling
  1. Sample topic ($z_{d,i}$)
  2. Sample changepoints (c_d)
- Identify changepoints by comparing likelihood with and without a changepoint.

(a) $p(\mathbf{w}, \mathbf{z} | c_d = 1)$ at document d　(b) $p(\mathbf{w}, \mathbf{z} | c_d = 0)$ at document d

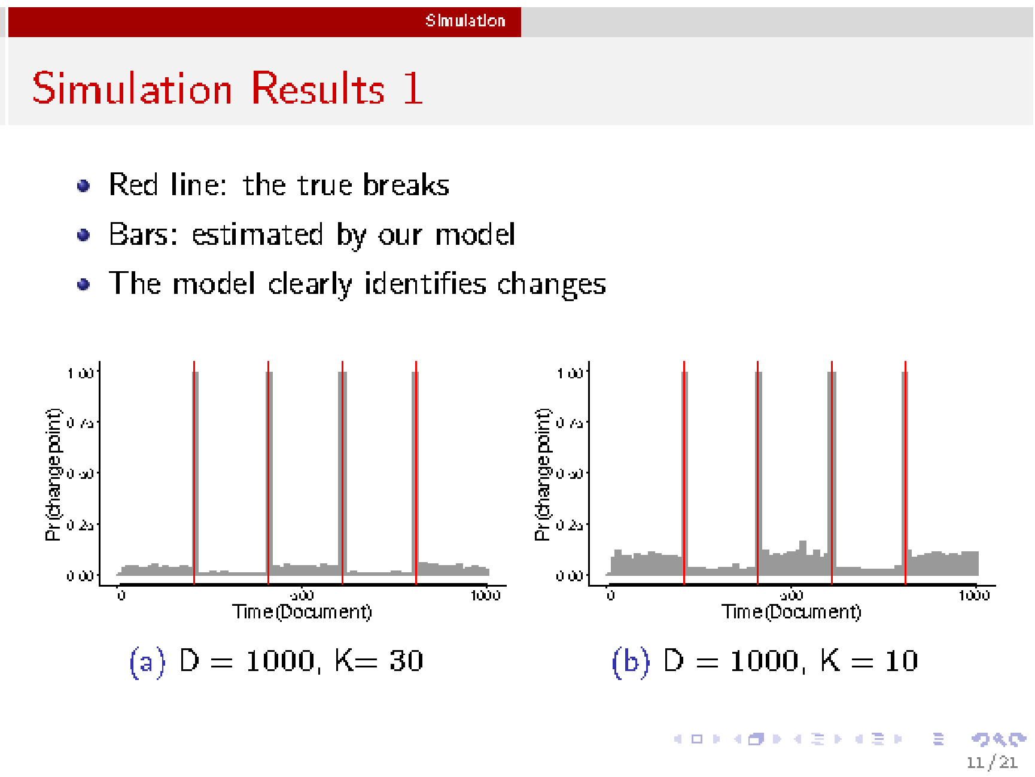


Simulation studies: Settings

- The number of topics (K): $K = 10, 30$
- The number of true breaks: 4 ($D = 1000$) and 1 ($D = 100$)

(1) D = 1000

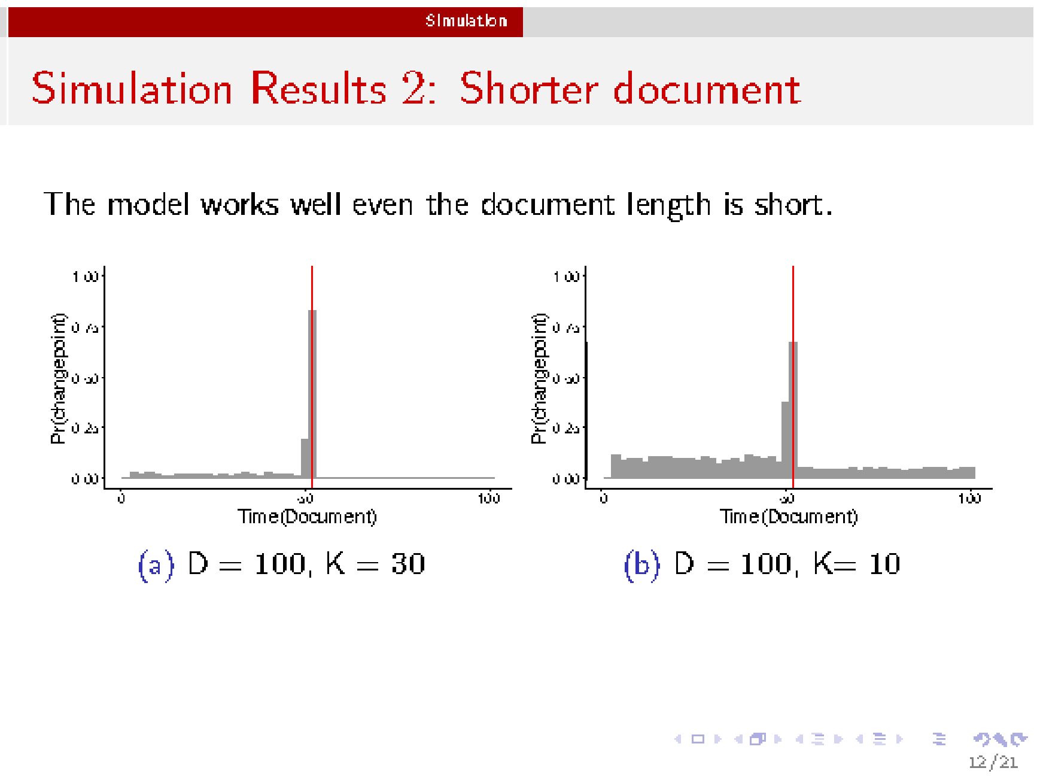
(2) D = 100



Simulation Results 1

- Red line: the true breaks
- Bars: estimated by our model
- The model clearly identifies changes

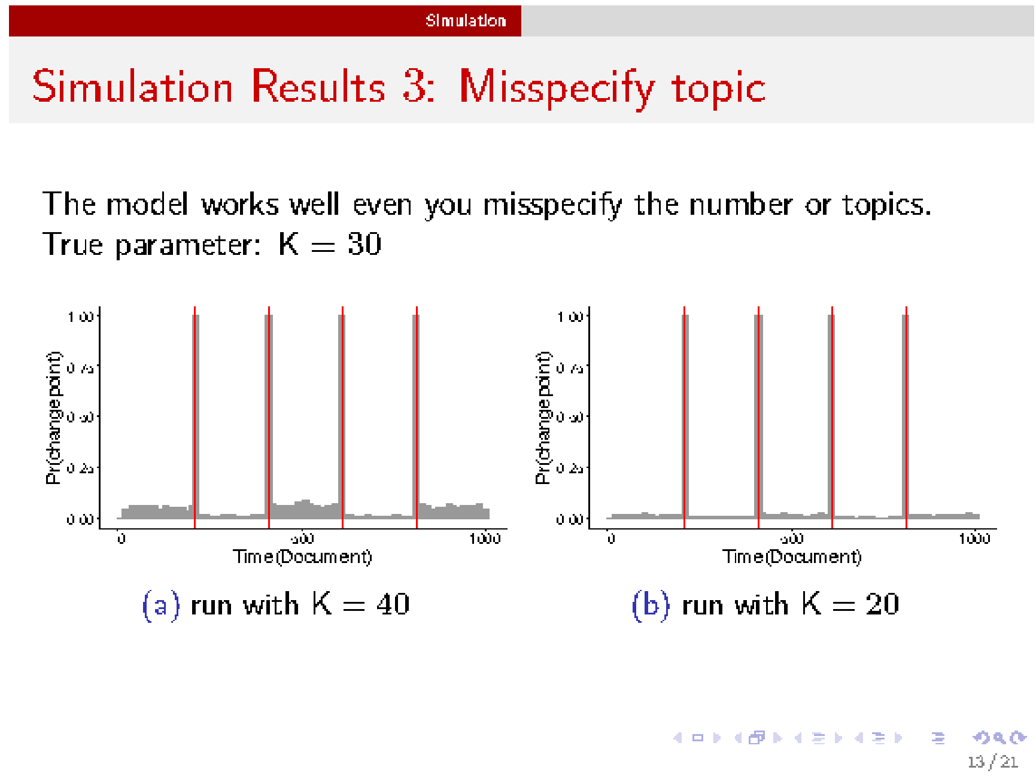
(a) D = 1000, K= 30　(b) D = 1000, K = 10



Simulation Results 2: Shorter document

The model works well even the document length is short.

(a) D = 100, K = 30　(b) D = 100, K= 10



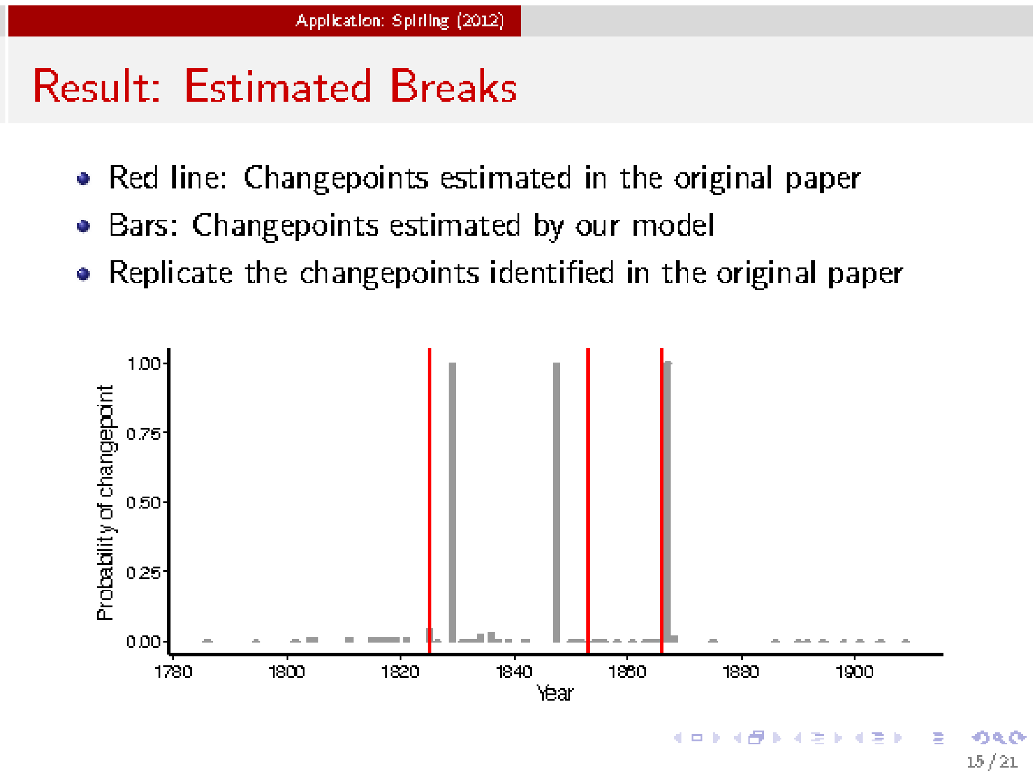
Simulation Results 3: Misspecify topic

The model works well even you misspecify the number or topics.
True parameter: K = 30

(a) run with K = 40　(b) run with K = 20

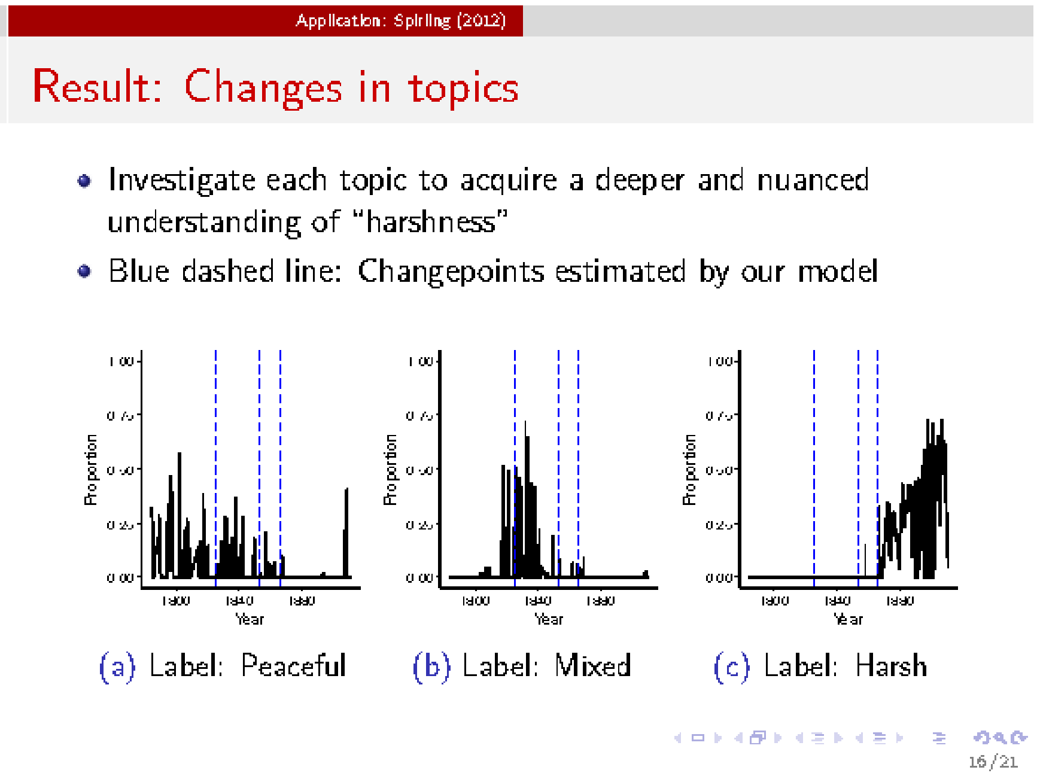
Application: Spirling (2012)

- Changes in the relationship between the Native Americans and the US government
- Employ treaties from 1784 to 1911
- Original paper: estimate "harshness" in a single dimension space
- Identify three changes
- The "harshness" against the Native Americans have gradually intensified



Result: Estimated Breaks

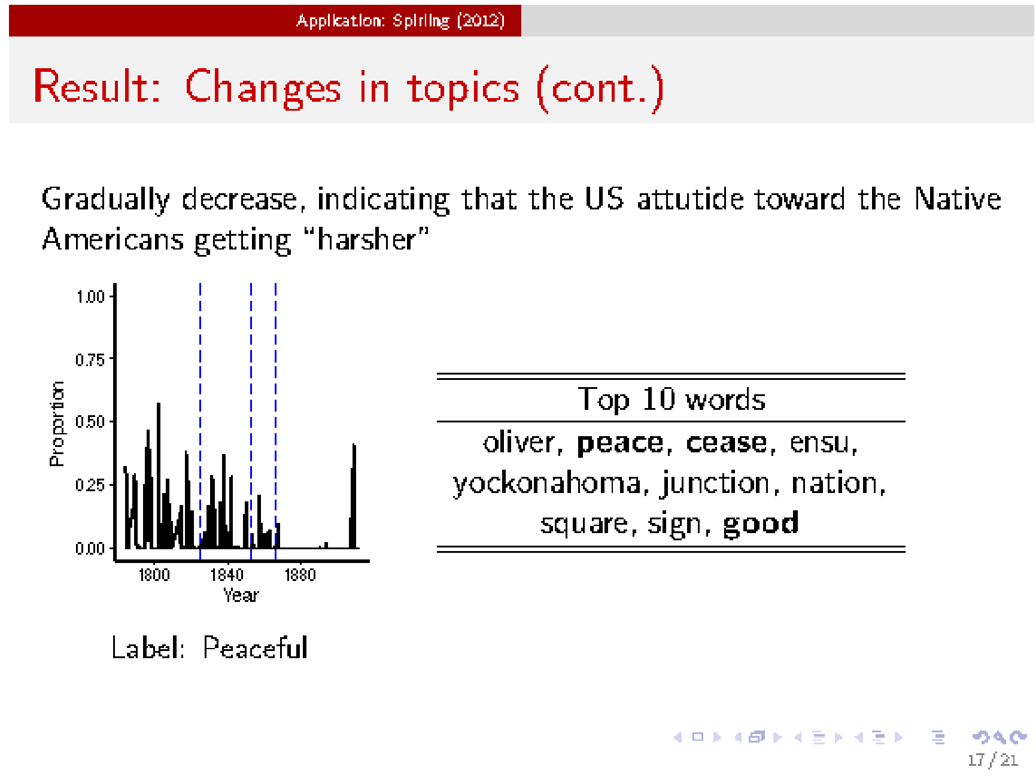
- Red line: Changepoints estimated in the original paper
- Bars: Changepoints estimated by our model
- Replicate the changepoints identified in the original paper



Result: Changes in topics

- Investigate each topic to acquire a deeper and nuanced understanding of "harshness"
- Blue dashed line: Changepoints estimated by our model

(a) Label: Peaceful　(b) Label: Mixed　(c) Label: Harsh

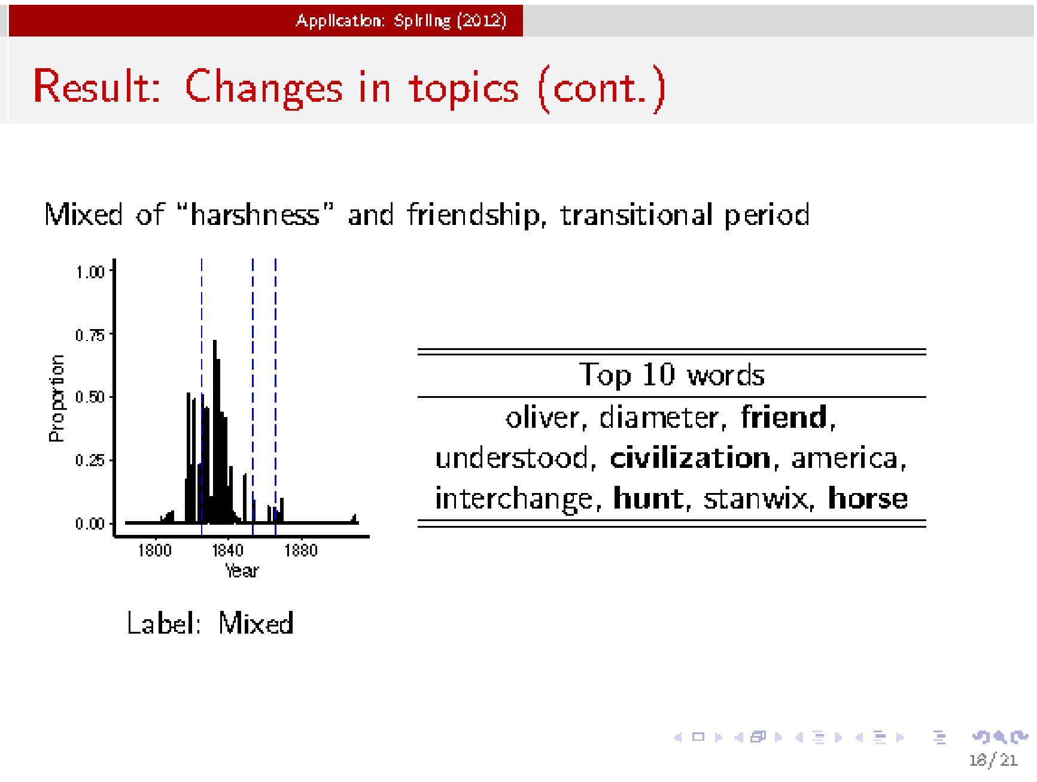


Result: Changes in topics (cont.)

Gradually decrease, indicating that the US attutide toward the Native Americans getting "harsher"

Top 10 words
oliver, **peace**, **cease**, ensu, yockonahoma, junction, nation, square, sign, **good**

Label: Peaceful

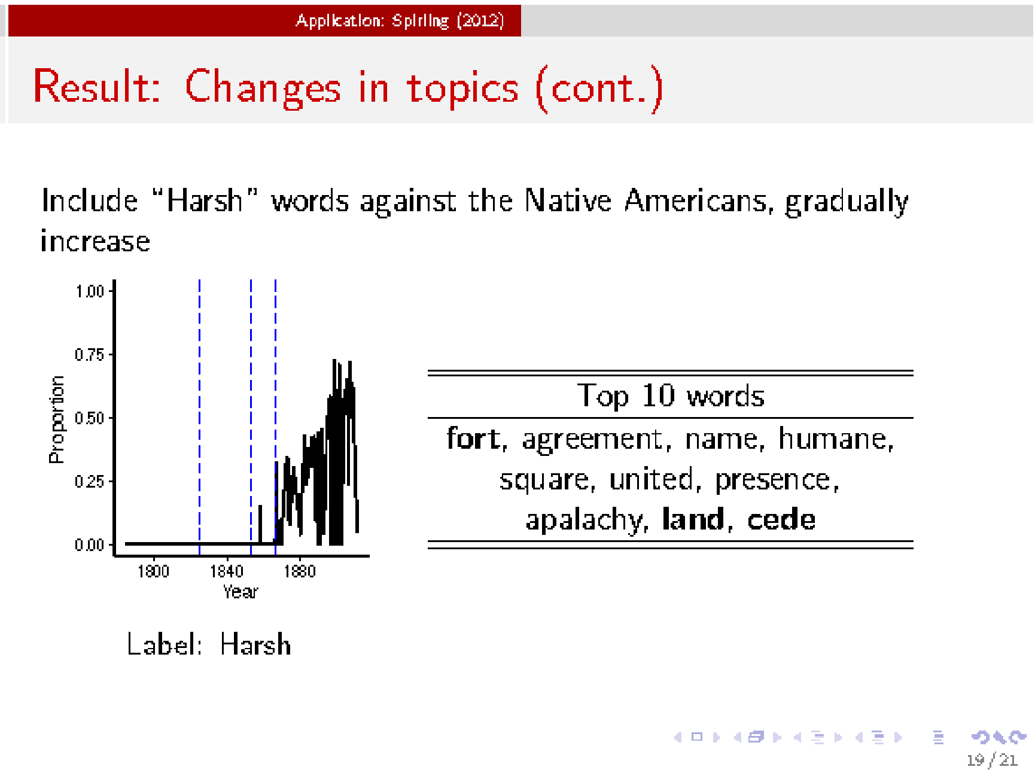


Result: Changes in topics (cont.)

Mixed of "harshness" and friendship, transitional period

Top 10 words
oliver, diameter, **friend**, understood, **civilization**, america, interchange, **hunt**, stanwix, **horse**

Label: Mixed



Result: Changes in topics (cont.)

Include "Harsh" words against the Native Americans, gradually increase

Top 10 words
fort, agreement, name, humane, square, united, presence, apalachy, **land**, **cede**

Label: Harsh

Next steps

- Incorporate hierarchical structure to understand the "gravity" of changes
- Add covariates to understand the relationships between document level meta-data and breaks

発表論文:

Eshima, Shusei (The University of Tokyo) and Daichi Mochihashi (The Institute of Statistical Mathematics), "Tree-Structured Topic Model: a nonparametric Bayesian approach to model texts in a continuous space", Asian Polmeth V (Poster), Seoul National University, 2018.

Sasaki, Tomoya (The University of Tokyo) and Daichi Mochihashi (Institute of Statistical Mathematics), "Detecting Topic Changes among Texts", Asian Polmeth V (Poster), Seoul National University, 2018.

大学共同利用機関法人 情報・システム研究機構
統計数理研究所
The Institute of Statistical Mathematics