# PPOL 628: Text as Data – Computational Linguistics for Social Scientists

Class 10: Topic Model Extensions and Validation

# Today

- Lecture: Extensions, Evaluation and Validation.
  - The structural topic model (STM) as a natural extension of LDA.
  - Advice and metrics to assess performance.
  - Validation strategies.

- Lab: topic_models_2.R

- Website: github.com/matthewjdenny/PPOL_628_Text_As_Data

# The Structural Topic Model (STM)

- When social scientists first began to use topic models, they would often take inferred document topic proportions, topic top-words, etc. and **look at their relationship to document level covariates**.

- STM is a model that combines both parts of this process into a single model.
  - Better description of generative process, more complex inference.

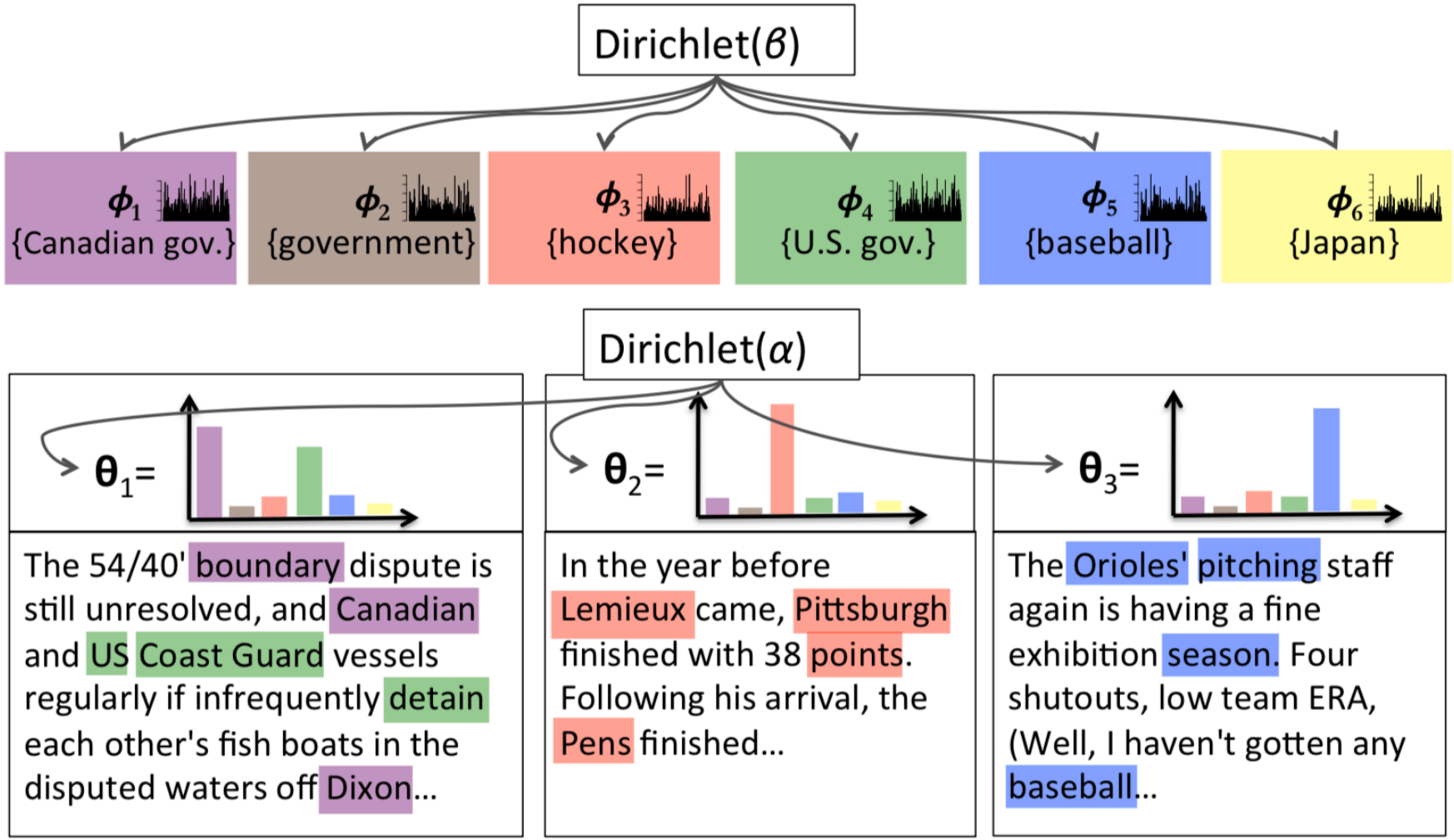
- Many papers (which I screenshot from liberally in these slides):
  - First Paper: https://scholar.princeton.edu/files/bstewart/files/stmnips2013.pdf
  - Derivation: https://scholar.princeton.edu/sites/default/files/bstewart/files/ajpsappendix.pdf
  - Paper we read for class: Roberts et al. (2014). Structural topic models for open-ended survey responses.
  - Package Vignette: https://cran.r-project.org/web/packages/stm/vignettes/stmVignette.pdf
  - Slides: http://ica-cm.org/wp-content/uploads/2017/05/roberts_topicmodels_combo.pdf

# The Structural Topic Model

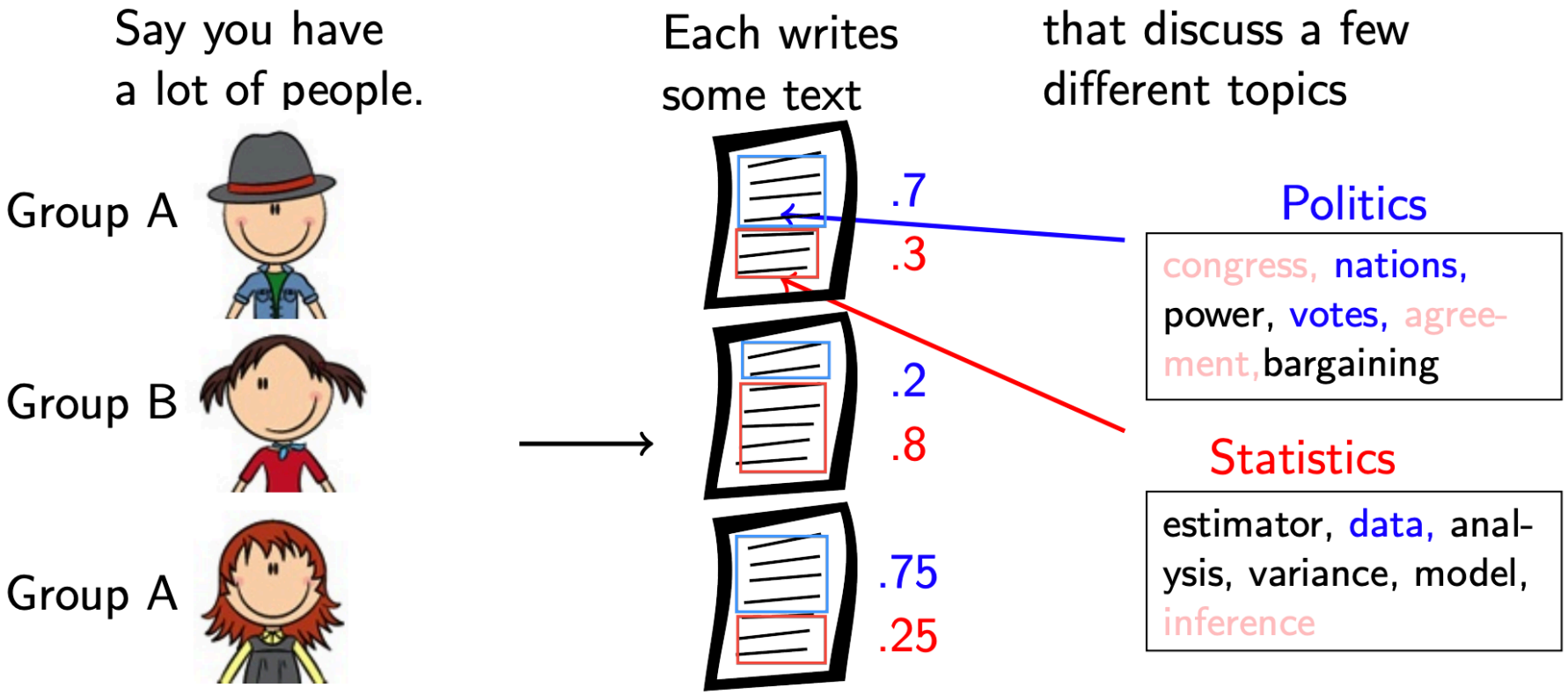
- Built on top of LDA, but allows us to incorporate document-covariates into model.

- STM provides two ways to include contextual information
  - Topic prevalence can vary by metadata
    - ★ e.g. Democrats talk more about education than Republicans
  - Topic content can vary by metadata
    - ★ e.g. Democrats are less likely to use the word "life" when talking about abortion than Republicans
- Including context improves the model:
  - more accurate estimation
  - better qualitative interpretability

# LDA Generative Process



# The Structural Topic Model




The STM Allows for:

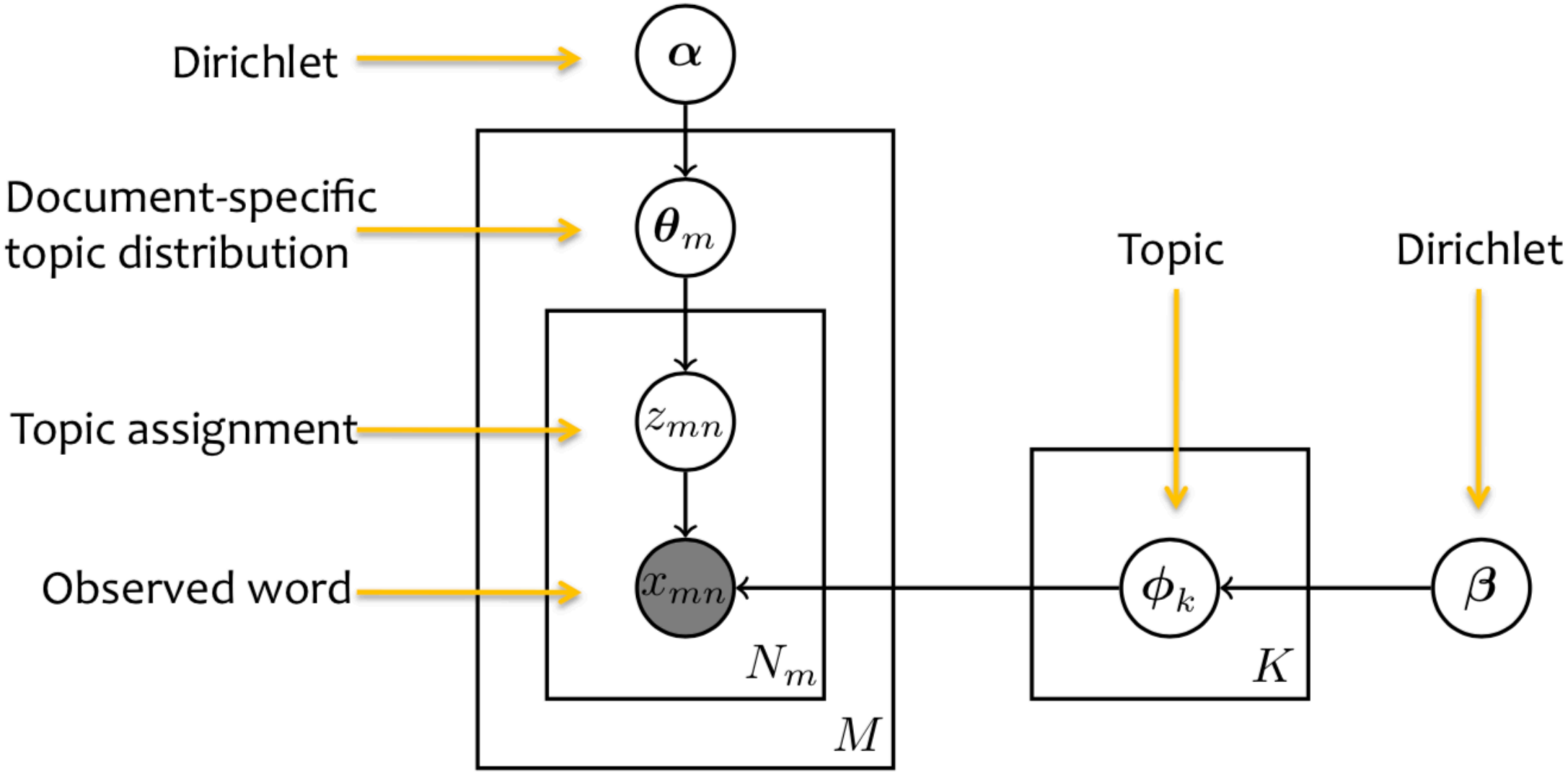
1. The words in each topic to vary by gender
2. The topic proportions to vary by group

# The Structural Topic Model

More formal terminology:

- User specifies the number of topics: $K$
- Observed data for standard topic models
  - Each document ($i \in 1 \ldots D$) is a collection of $M_i$ tokens
- Additional data for STM
  - Topic prevalence covariates: $D \times P$ matrix $\mathbf{X}$
  - Topical content groups: $D$ length vector $\mathbf{Y}$
- Latent variables
  - $D \times K$ matrix $\theta$: proportion of document on each topic.
  - $K \times V$ matrix $\beta$: probability of drawing a word conditional on topic.

# LDA Plate Model

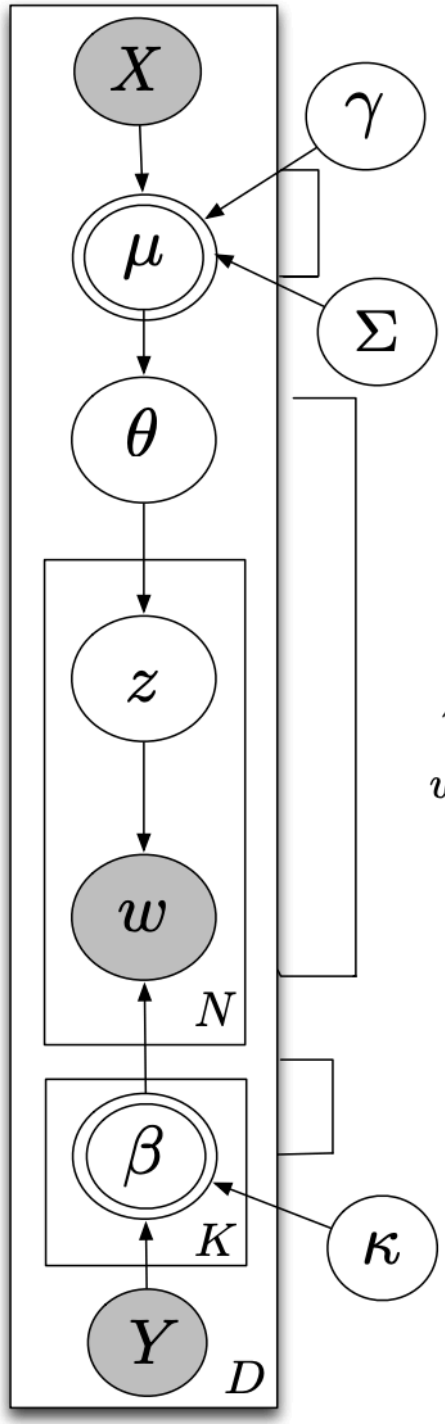


# STM Plate Diagram

**Logistic-normal GLM with covariates**

**“Normal LDA”**

**Multinomial Logit**




**Topic Prevalence:**

$$\mu_{d,k} = X_d\gamma_k$$
$$\gamma_k \sim \mathcal{N}(0, \sigma_k^2)$$
$$\sigma_k^2 \sim \text{Gamma}(s^\gamma, r^\gamma)$$

**Language Model:**

$$\theta_d \sim \text{LogisticNormal}(\mu_d, \Sigma)$$
$$z_{d,n} \sim \text{Mult}(\theta_d)$$
$$w_{d,n} \sim \text{Mult}(\beta_d^{k=z_{d,n}})$$

**Topical Content:**

$$\beta_{d,v}^k \propto \exp(m_v + \kappa_v^{\cdot,k} + \kappa_v^{y,\cdot} + \kappa_v^{y,k})$$
$$\kappa_v^{y,k} \sim \text{Laplace}(0, \tau_v^{y,k})$$
$$\tau_v^{y,k} \sim \text{Gamma}(s^\kappa, r^\kappa)$$

# The Structural Topic Model

- $\theta$, $D \times K$ document-topic matrix $\Leftarrow$ logistic normal glm with covariates
  - Covariate-specific prior with global topic covariance
  - $\theta_{i,\cdot} \sim \text{LogisticNormal}(X_i\gamma, \Sigma)$
- $\beta$, $K \times V$ topic-word matrix $\Leftarrow$ multinomial logit with covariates
  - Each topic is now a covariate-specific deviation from a baseline distribution.
  - $\vec{\beta}_{k,\cdot} \propto \exp\left(m + \kappa^{(\text{topic})} + \kappa^{(\text{cov})} + \kappa^{(\text{int})}\right)$
  - Thee parts: topic, covariate, topic-covariate interaction
- Each token has a topic drawn from the document mixture
  - Draw token topic $z_{i,m}$ from Multinomial($\theta_i$)
  - Draw observed word $w_{i,m}$ from Multinomial($\beta_{k=z,}$)

# STM: Inference

- **General Approach:** Same as LDA in that we want to "invert" the generative process we have posited for our model and use the data we observe to infer the latent variables that most likely generated our data.
- Unlike LDA, where we can derive an elegant and relatively simple (collapsed) Gibbs sampler STM requires a more complex, approximate inference technique.
  - No longer have conjugate priors.
  - More expressive model ←→ more difficult inference.
- Semi-collapsed variation expectation maximization.
  - Approximate true posteriors of interest to make inference feasible.

# Semi-collapsed variational EM

- **Variational Approximation**: take a distribution that is difficult to work with (intractable inference) and pick a distribution that behaves similarly but is easier to work with.
- **Expectation Maximization Algorithm**:
  - E-Step: Find an approximation of the log likelihood function for your model, conditional on you data and current latent variable estimates.
  - M-Step: Maximize this approximation of the log likelihood function with respect to your latent variables.
  - Iteratively repeat these steps until convergence.
- **Semi-collapsed**: Instead of changing one token-topic assignment at a time, update all of them for a document at once.

# STM: semi-collapsed variational EM

- For each document
  - Update $q(\eta)$ (document-topic proportions) by optimizing the collapsed objective
  - Solve in closed form for $q(z)$ (assignments to topic for each word)
- Update $\gamma$ (coefficients for topic prevalence)
- Update $\Sigma$ (global covariance matrix controlling correlation between topics)
- Update each $\kappa$. (topical-content parameters describing deviations from baseline word rate)
- Repeat until convergence.

# STM: Overall Thoughts

- **Widely Used, Flexible Model, Highly Interpretable**
  - If you care about how different authors use topics differently, how topics vary with covariates, etc. then you can neatly infer these relationships all at once.
  - Interpretability for document-covariate effects is just like a regression model.
  - Lots of nice built-in functionality in stm package for interpreting output.

- **Complex Model**
  - Can have poor performance where model assumptions do not hold. Like vanilla LDA, still want a relatively large number of documents (thousands) to feel confident in inference.
  - Time and space complexity are high – not suitable for application to millions of documents with an extremely large vocabulary.

# LDA Optimization and Validation

- **Hyper-parameter Optimization.**
  - Selecting $\alpha$ and $\beta$ -- let model choose.
  - Selecting number of topics – theory, let model choose (some applications).
  - Selecting number of iterations and assessing convergence (with Geweke test).
- **Assessing model performance**
  - High quality topics? Cohesiveness, Exclusivity, FREX, Perplexity
- **Validation**
  - Key point is to think about it, take steps to determine validity.
  - From Quinn et al. (2010) – Specification + Sensitivity Analysis, Reliability, Validity, Interpretability.

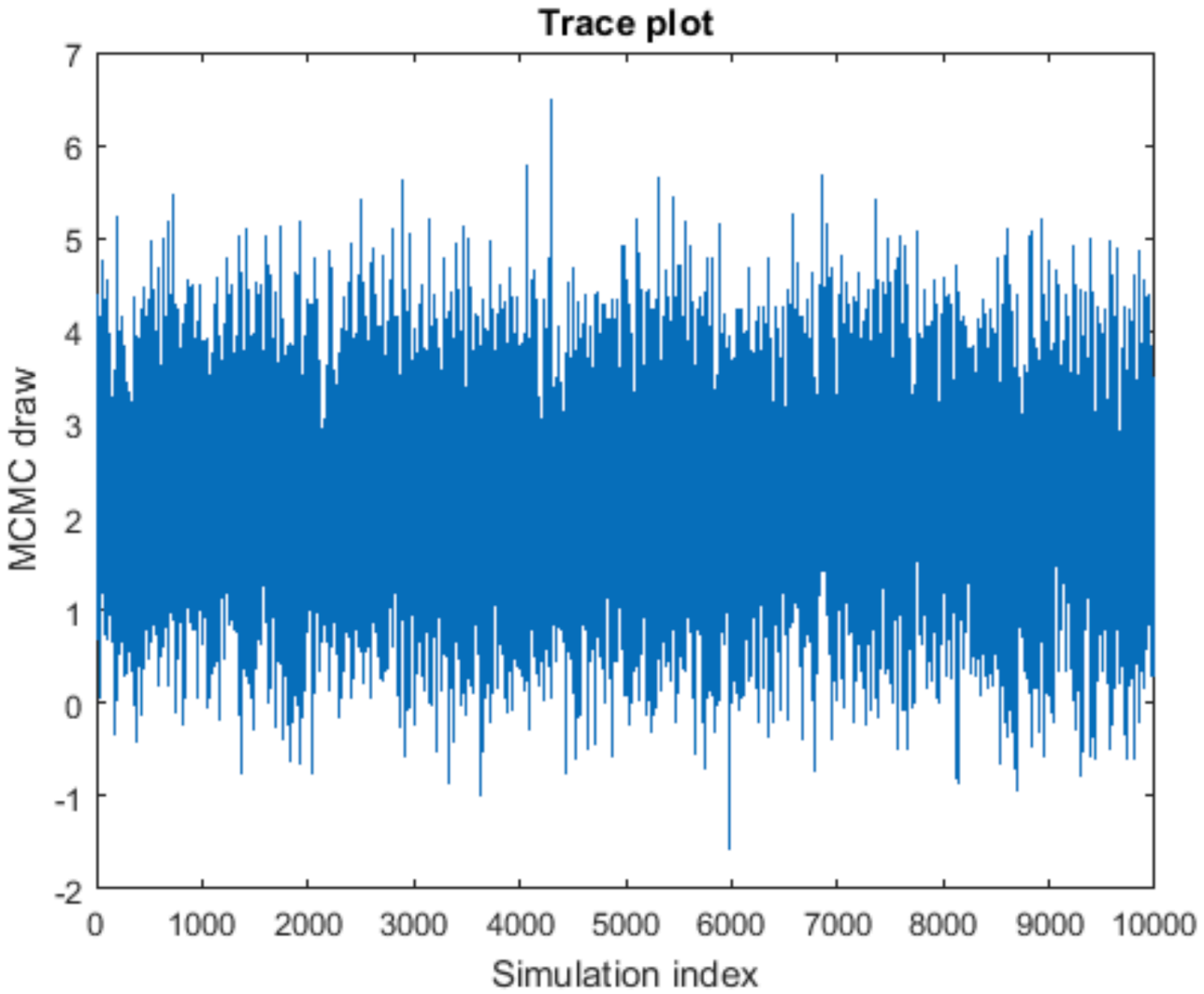
# Geweke Test for Convergence of MCMC

- Want to determine if we have reached the **stationary distribution** of our latent variables of interest.
- Way to determine if we have selected an appropriate number of iterations for inference.
- Geweke test compares the first 50% of samples (after burnin) to the last 10% and looks for a statistically significant difference.

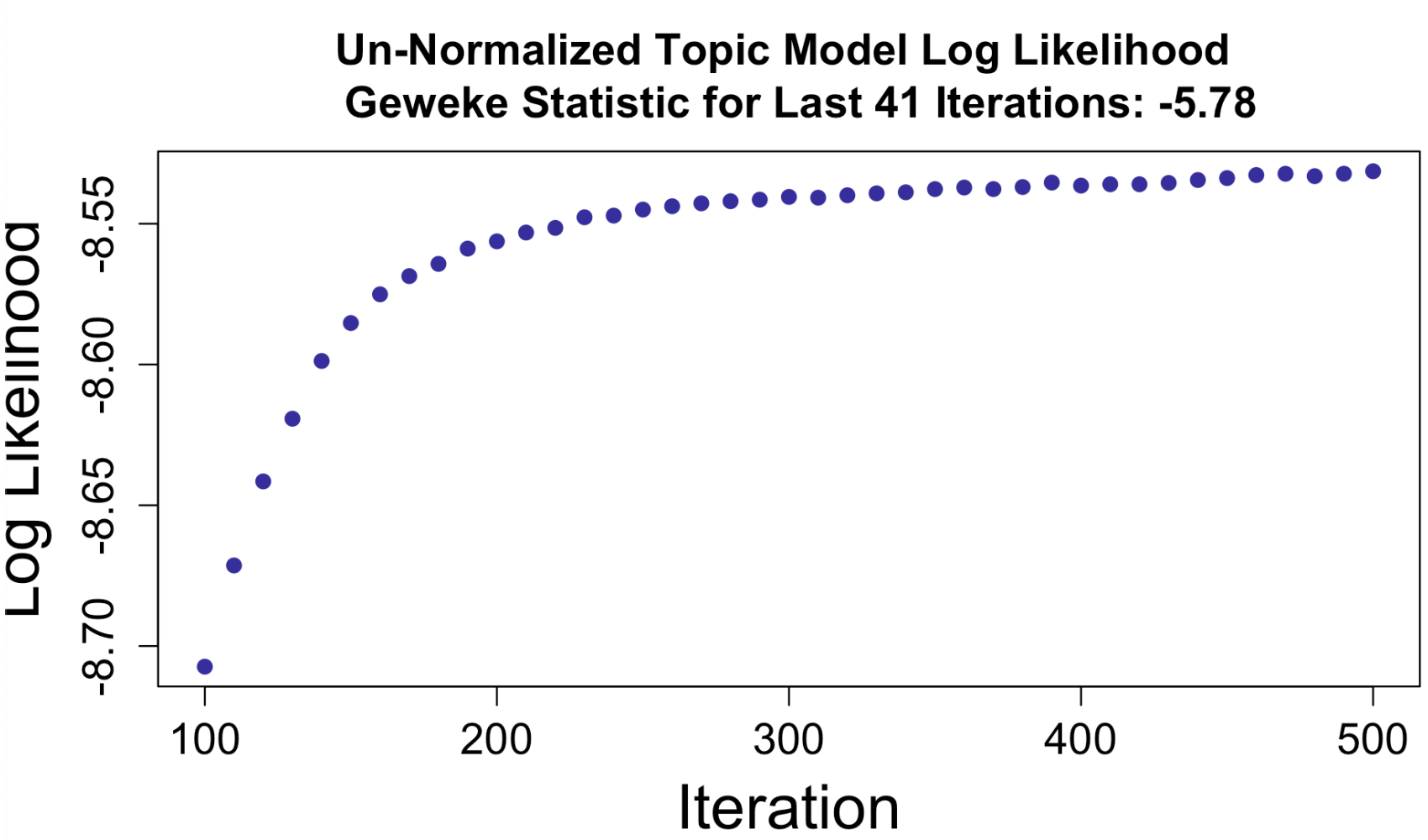
Two-sample ($X_1$ and $X_2$) T test of mean (unequal variance)

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}},$$

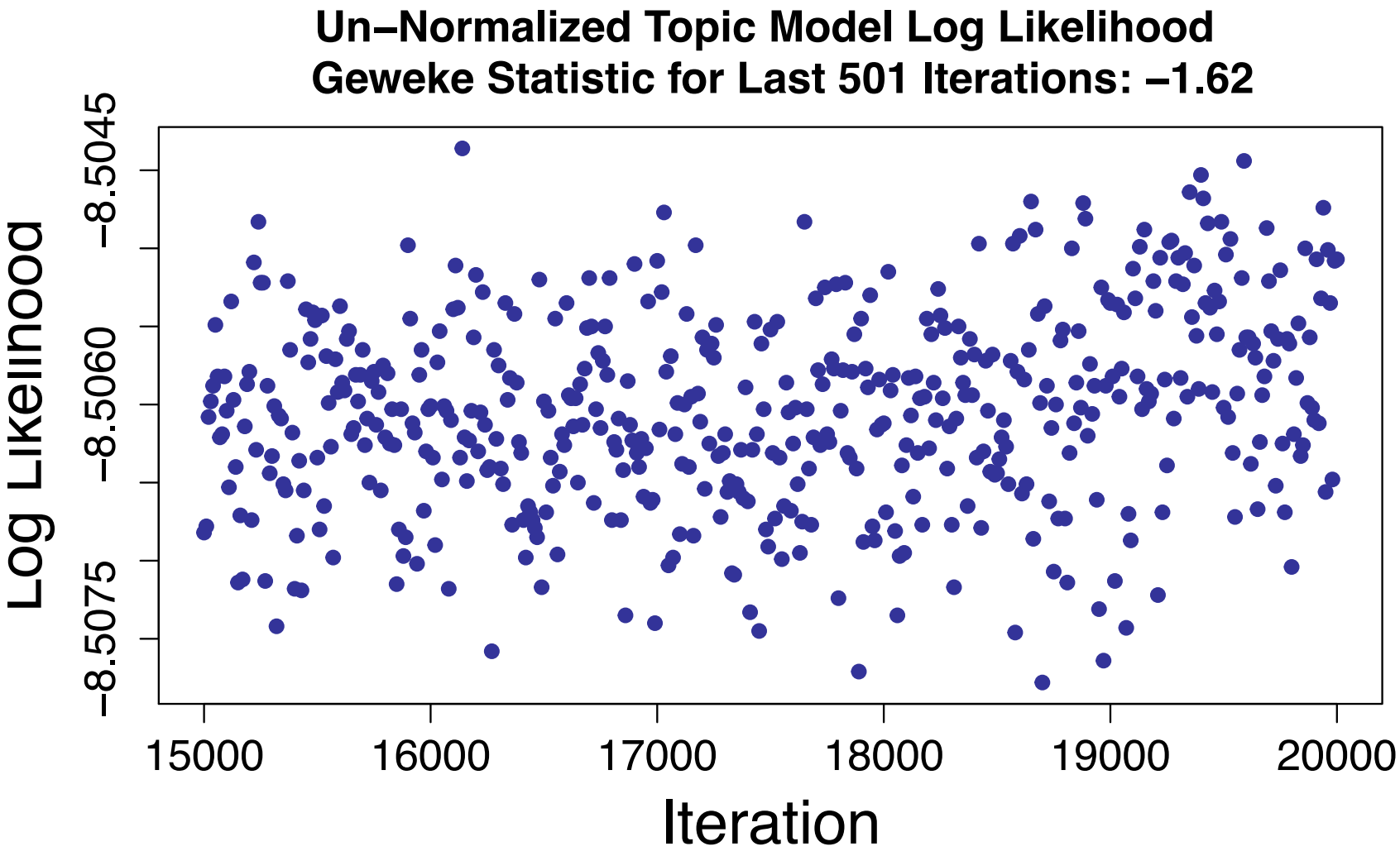
# Goal: reach the stationary distribution



# Model has not reached stationary distribution

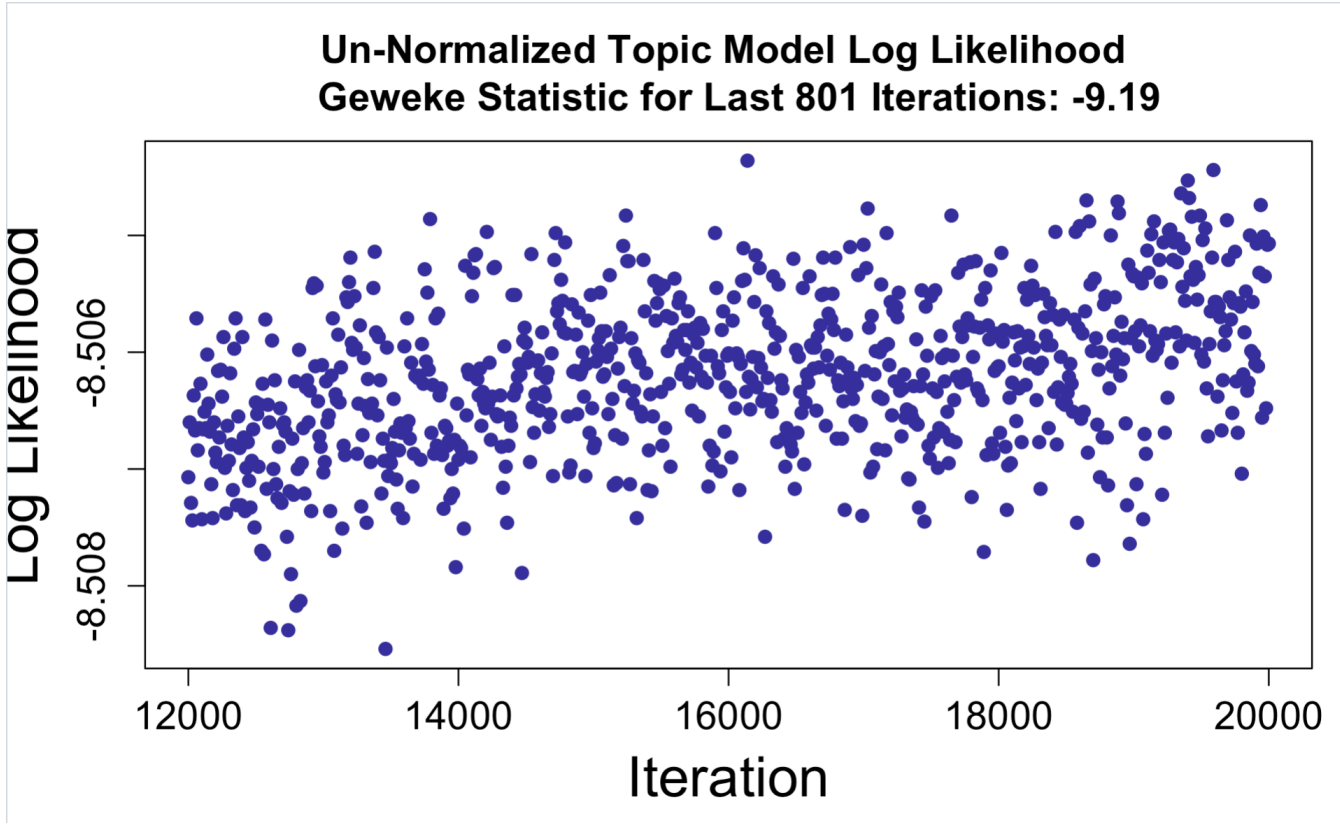


# Model has reached stationary distribution



# Geweke Statistic is not a Panacea

- Same chain as last slide, different window, different conclusion.
- Run model for longer than you think you need to, look at chain.
- Experience is important.



http://dirichlet.net/pdf/mimno11optimizing.pdf

# Topic Cohesiveness (Coherence)

- Consider the output of a topic model)
- We might select 5 top words for each topic

| | | | | | |
|---|---|---|---|---|---|
| Topic 1 | bill | congressman | earmarks | following | house |
| Topic 2 | immigration | reform | security | border | worker |
| Topic 3 | earmark | egregious | pork | fiscal | today |

- An ideal topic?$\rightsquigarrow$ will see these words co-occur in documents
- Define $\boldsymbol{v}_k = (v_{1k}, v_{2k}, \ldots, v_{Lk})$ be the top words for a topic
- For example $\boldsymbol{v}_3 =$ (earmark , egregious , pork , fiscal , today )

http://dirichlet.net/pdf/mimno11optimizing.pdf

# Topic Cohesiveness (Coherence)

Define the function $D$ as a function that counts the number of times its argument occurs:

$$\begin{aligned} D(\text{earmark}, \text{egregious}) &= \text{No. times earmark and egregious co-occur} \\ D(\text{egregious}) &= \text{Number of times Egregious occurs} \end{aligned}$$

Define cohesiveness for topic $k$ as

$$\text{Cohesive}_k = \sum_{l=2}^{L} \sum_{m=1}^{l-1} \log \left( \frac{D(v_{lk}, v_{mk}) + 1}{D(v_{mk})} \right)$$

Define overall cohesiveness as:

$$\begin{aligned} \text{Cohesive} &= (\sum_{k=1}^{K} \text{Cohesive}_k)/K \\ &= \left( \sum_{k=1}^{K} \sum_{l=2}^{L} \sum_{m=1}^{l-1} \log \left( \frac{D(v_{lk}, v_{mk}) + 1}{D(v_{mk})} \right) \right) /K \end{aligned}$$

http://dirichlet.net/pdf/mimno11optimizing.pdf

# Topic Coherence

- Measure of the degree to which top words in topic co-occur in documents.
- Maximized when all top terms always co-occur with each other.
- Minimized when top terms never co-occur.

Letting $D(v)$ be the *document frequency* of word type $v$ (i.e., the number of documents with least one token of type $v$) and $D(v, v')$ be *co-document frequency* of word types $v$ and $v'$ (i.e., the number of documents containing one or more tokens of type $v$ and at least one token of type $v'$), we define *topic coherence* as

$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})}, \quad (1)$$

where $V^{(t)} = (v_1^{(t)}, \ldots, v_M^{(t)})$ is a list of the $M$ most probable words in topic $t$. A smoothing count of 1 is included to avoid taking the logarithm of zero.

# Topic Exclusivity

We also want topics that are exclusive$\rightsquigarrow$ few replicates of each topic

$$\text{Exclusivity}(k, v) = \frac{\mu_{k,v}}{\sum_{l=1}^{K} \mu_{l,v}}$$

Suppose again we pick $L$ top words. Measure Exclusivity for a topic as for a topic as:

$$\begin{aligned}
\text{Exclusivity}_k &= \sum_{j:v_j \in \boldsymbol{v}_k} \frac{\mu_{k,j}}{\sum_{l=1}^{K} \mu_{l,j}} \\
\text{Exclusivity} &= \left( \sum_{k=1}^{K} \text{Exclusivity}_k \right) / K \\
&= \left( \sum_{k=1}^{K} \sum_{j:v_j \in \boldsymbol{v}_k} \frac{\mu_{k,j}}{\sum_{l=1}^{K} \mu_{l,j}} \right) / K
\end{aligned}$$

# Topic Coherence

- How well does your model fit held-out data?
- Wallach et al. (2009). Evaluation methods for topic models. *ICML*
  - **Left-to-right** method is preferred.

How well does our model perform? $\rightsquigarrow$ predict new documents?
Problem $\rightsquigarrow$ in sample evaluation leads to overfit.
Solution $\rightsquigarrow$ evaluate performance on held out data
For held out document $\boldsymbol{x}^*_{\text{out}}$

$$\text{Perplexity} \quad = \quad \exp\left(-\log p(\boldsymbol{x}^*_{\text{out}}|\boldsymbol{\mu}, \boldsymbol{\pi})\right)$$

# Hyper-parameters + Model Performance

- It is ok to refit your model, even if you are using it for measurement if you do not base decisions to refit on properties that will influence measurements (e.g. topic proportion in documents, etc.).
  - Can optimize for coherence + exclusivity to select # topics.
  - Use Geweke test + look at chain to determine convergence.
- Many automated metrics (esp. Perplexity) will prefer large number of topics.
- preText for sensitivity analysis once you have final specification.
- My take: no hard and fast rules. Want to combine domain expertise (READ DOCUMENTS) + theory + metrics to pick between a few specifications + sensitivity analysis.

# Sensitivity, Reliability, Interpretability

- **Specification + Sensitivity Analysis**
  - How does varying preprocessing, hyperparameters change results.
  - Do you have a strong theory about why you chose the parameters you did?
  - For those you did not have a strong theory for, sensitivity analysis.
- **Reliability**
  - Make sure to set your random seed + save replication materials so others can exactly replicate your results.
  - Intercoder reliability → analogue is same results across multiple seeds.
- **Interpretability**
  - Can you draw conclusions from your results? Do topics map on to constructs you care about?
  - **Do high probability documents in topics actually contain that topic?**

# Validity

- **Semantic validity:** the extent to which each category or document has a coherent meaning and the extent to which the categories are related to one another in a meaningful way.
- **Convergent construct validity**: the extent to which the measure matches existing measures that it should match.
- **Discriminant construct validity**: the extent to which the measure departs from existing measures where it should depart.
- **Predictive validity**: the extent to which the measure corresponds correctly to external events.
- **Hypothesis validity**: the extent to which the measure can be used effectively to test substantive hypotheses.

# Validation – My Take

- **Start with a plan:** document the steps you take in data preparation in analysis. Have reasons/theory for your modelling choices. Tell your reader when things did not work. Set your seed!
- **Use hyper-parameter optimization**: where possible, allow the model to learn optimal hyper-parameters to reduce researcher input.
- **Make sure model has converged before interpretation**.
- **Engage in some validation efforts**.
  - How you validate depends on your measurement goals, what metadata you have available. Do not need to do everything from Quinn et al., (2010).
- **Interpretation checks**: make sure you look at a sample of high probability documents in topics of interest to make sure they are actually about that topic. Read enough of corpus to come to qualitative judgement about model quality.