Latent Variable Models

Kevin McAlister

Department of Political Science
University of Michigan

October 21, 2016

Latent Variable Models

- In the social sciences, models rely on observable data.
  - Survey Data
  - Experimental Data
  - Cross-Sectional Data
- Observable data allows us to model directly observable phenomena.
- What if the variable of interest is not directly observable?
  - Liberal-Conservative Ideology
  - Topic of a speech
  - Political Knowledge
- Measuring these concepts directly is difficult.
- Latent variable models provide an approach for measuring these concepts.

What Are Latent Variable Models?

- Latent variable models assume that the observed covariates, $X_i \forall i \in [1, p]$, are explained in a smaller set of unobserved latent variables (factors, hidden variables, blind sources).
- For example, imagine trying to model the probability that a person votes for Donald Trump.
- Covariates could include income, party ID, opinions of issues, etc.
- Given a training set, imagine a logistic model:

$$Y = X\beta + \epsilon$$

  where Y is an indicator of vote choice and X is the matrix of covariates.
- The coefficients on X would tell us how changes in the covariates influence the probability that a person votes for Trump.
- Changes in the variables only indirectly tell us what we want to know. We want to know how ideology influences probability of voting for Trump.
- If this is out variable of interest, why not attempt to measure it and apply it in the regression?
- Latent variable models allow us to attempt to indirectly measure ideology which can be applied to this model.

What Are Latent Variable Models?

- Similar to the above, what if we are interested in figuring out which groups of voters vote for Trump?
- How do we determine groups?
- Creating a classification structure would allow us to directly answer this question.
- Given high dimensional data, it is difficult to parse through all the data and create meaningful classifications.
- LVMs provide algorithmic methods for measuring the latent class structures that define the data set.

What Are Latent Variable Models?

- Note that LVMs are applied to a matrix of covariates.
- LVMs help us extract and understand the commonalities that are shared by our observed variables.
- LVMs can be used for many purposes:
  - Dimensionality reduction of large-P data
  - Indirect measurement of latent concepts
  - Group identification given a set of observations
- Much social science research revolves around the notion of extracting latent dimensions within the data, then applying names to said dimensions.

A Probabilistic View

- Given a matrix of covariates, X, we want to model a set of latent variables, Z, along with a set of model parameters, θ.
- θ can include information about the groups dictated by the latent variable. For example, Gaussian mixture models:
  - X is the matrix of covariates
  - Z is the matrix of K class labels, $z \in [1, K]$
  - θ includes the K means and variances associated with each of the classes.
- The joint distribution of Z and θ conditional on X:

$$p(Z, \theta | X)$$

where $p(.|.)$ is a probability density or mass function.

A Probabilistic View

- We want to find Z^* and θ^* that maximize the marginal probability of X, or the probability that we would observe X given the other variables.

$$P(X|\theta^*) = P(X, Z^*|\theta^*)$$

- Note that Z is associated with X. We can think of Z as additional covariates.
- But, Z is unobserved! How can we maximize if we don't know its structure?
- The structure of Z must be defined by the researcher.
- In general, the structure of Z is defined according to the desired level of measurement of the latent variable.
  - If we want to put observations in groups, Z is defined to be discrete.
  - If we want to measure a continuous latent variable, Z will be defined to be continuous.
- Given the level of measurement, we must also determine the distributional structure of the latent variable.
  - Discrete latent variables will generally follow a multinomial/categorical distribution.
  - Continuous latent variables have many more options. In general, latent variables are assumed to follow a normal distribution for convenience purposes.

A General Estimation Method

- Remember that we want to maximize the marginal probability of X.
- In terms of likelihoods, we want to find the values of θ and Z that maximize the log-likelihood of the N data points:

$$ll(\theta|X) = \log[P(X|\theta)] = \sum_{i=1}^{N} \log[P(x_i|\theta)]$$

- However, we want to calculate the likelihood of the complete data set, latent variable and all:

$$ll_c(\theta|X, Z) = \log[(P(X, Z|\theta)] = \sum_{i=1}^{N} \log[P(x_i, z_i|\theta)]$$

- Using the definition of joint probabilities, we can expand the above likelihood formulation to:

$$\sum_{i=1}^{N} \log[P(x_i|\theta, z_i)P(z_i|\theta)]$$

A General Estimation Method

- Thus, an estimation procedure should first find the probability that we would see the latent variable value for each observation given θ then find the probability that we would see x_i given z_i and θ.
- Each z_i should be imputed conditional on θ, then used in the marginal probability calculation for X.
- If N is large, the number of parameters to be estimated can be very large.
- Analytical solutions to the likelihood equations are generally not feasible. We must rely on computational methods.
- A frequentist approach to estimation of the likelihoods would utilize an EM algorithm.
- However, these algorithms tend to fail when N is large because inversion of the information matrix is computationally infeasible.
- A feasible alternative is a Markov Chain Monte Carlo algorithm. However, this requires that we approach this problem in a Bayesian way.

A Bayes Primer (with minimal evangelizing)

- Shifting from a frequentist to a Bayesian point of view requires a fundamental change in the way we interpret estimators.
- Frequentists view probability as a description of randomness. Estimators are associated with errors and those errors are assumed to arise due to random fluctuations in the data that we observe.
- If we repeat a data collection activity a large number of times, we would expect that our estimator would follow the error structure given by our declarations of probability.
- Bayesians view probability as a description of uncertainty. Estimators are associated with errors and these errors are assumed to arise due to uncertainty associated with the estimator. There is not notion of repeated experiments in the Bayesian point of view.
- An estimator is assumed to have a single uncertain value and we are only able to interpret the relative probability of the true value.

A Bayes Primer (with minimal evangelizing)

- What is the probability that a 95% confidence interval includes the true value of our parameter of interest?
- Our sample doesn't tell us anything about this.
- A Bayesian point of view is that the interval gives us credible information about the true value of the parameter of interest.
- What is the probability that a 95% credible interval gives includes the true value? .95!
- Bayesians interpret uncertainty in the way that we all want to interpret it anyways.

The Bayesian Recipe

- Assess hypotheses by calculating their probabilities, $p(H_i|...)$ conditional on known and/or presumed information using the rules of probability theory.
- The grammar:
  - $P(H_1 + H_2|\theta) = P(H_1|\theta) + P(H_2|\theta) - P(H_1, H_2|\theta)$
  - $P(H_1, D|\theta) = P(H_1|\theta)P(D|H_1, \theta)$
- The vocabulary:
  - Certainty: If A is certainly true given B, then $P(A|B) = 1$.
  - Falsity: If A is certainly false given B, then $P(A|B) = 0$.

The Bayesian Recipe

- The results:
  - Normalization:
    - For exclusive, exhaustive H_i: $\sum_i P(H_i|...) = 1$
  - Marginalization:
    - $\sum_i P(A, B_i|\theta) = \sum_i P(B_i|\theta)P(A|B_i, \theta) = P(A|\theta)$
- The law:
  - $P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$

Parametric Estimation Using Bayes

- Using these rules, we can develop a method for estimating the true values of a parameter set, θ, given some observed data, X:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

- $P(X|\theta)$ is the likelihood of our data given some values for θ. Note that this implies that we are making an assumption about the structure of $P(X|\theta)$.
- We're learning about θ. What is the $P(\theta)$?
- $P(\theta)$ is the incorporation of our prior information about the true value of θ.
- $P(X)$ is the marginal likelihood of the data. In other words, $P(X)$ is the $\sum_{\theta} P(X|\theta_i)$.
- Bayesian estimation is a lot like maximum likelihood, but with priors and a different interpretation of error.

Parametric Estimation Using Bayes

- We can summarize the Bayesian estimation structure as:

$$Posterior \propto Likelihood \times Prior$$

- Using the rules of probability, we get a posterior distribution which is the joint distribution of the parameters which we are trying to estimate.
- To get marginal probability distributions, we just marginalize over the posterior.
- The marginal PDF associated with each parameter is the distribution of the uncertainty of the true value of the parameter.
- What is the best estimate of the true value? Pick a value that minimizes Bayes' risk:
  - MSE? Posterior Mean
  - Linear loss? Posterior Median
  - Minimum a posterior loss? Posterior Mode

Parametric Estimation Using Bayes

- How do we quantify our uncertainty about our best estimate? Credible intervals!
  - Mean: $C\%$ density with the mean in the middle
  - Median: $[l, h] = \left[\int_{l}^{M} P(\theta_i) = \frac{C}{2}, \int_{M}^{h} P(\theta_i) = \frac{C}{2}\right]$
  - Mode: Smallest interval which includes $C\%$ of the density of $P(\theta_i)$

Get Out The Pitchforks!

- This is the point where we discuss priors. (If subjectivity in statistics is not something you accept, cover your ears now.)
- Bayesian estimation requires specifying priors for all elements of θ.
- Where does a prior come from? Your mind! A prior is your guess at the true value of θ.
- A prior also incorporates your uncertainty about your guess.
- The choice of distribution of your prior can be informed by your knowledge of the parameter of interest.
- Most times, though, priors are chosen to be *conjugate* to your likelihood functions for convenience reasons.
- Your prior choice will influence the location and spread of the marginal posterior distribution.

Get Out The Pitchforks!

- Your prior choice will influence the location and spread of the marginal posterior distribution.
- How much? It depends:
  - If the prior is close to the distribution of the likelihood, then the posterior doesn't change much.
  - If the prior has a stronger certainty than the likelihood, then the posterior will favor the prior.
  - If the prior has little certainty (high spread), then the posterior will favor the likelihood.
- A low variance ML estimate arises when N is large. This implies that when N is large, the prior specification doesn't matter much.
- There is a lot of discussion about uninformative priors. Even a flat prior over a subset of $\mathbb{R}$ has information in some respect. In general, to let the likelihood speak as loudly as possible in the posterior, a flat prior is used when the parameter space is finite and an extremely diffuse prior is used in the infinite case.

The Duality (Any Covered Ears Can Be Uncovered)

- When models are linear in the parameters and have additive Gaussian noise, frequentist results are identical to Bayesian results with flat priors.
- For models that have frequentist analogues, as $N \to \infty$, frequentist results will be virtually indiscernible from the corresponding Bayesian results.
- This duality is an important fact in modern statistics. As big data becomes more prevalent, the frequentist vs. Bayesian debate becomes less and less important.

Bayesian Model Comparison

- Another critical distinction between frequentist and Bayesian approaches is the hypothesis space.
- Frequentists assume that the hypothesis space is infinite and test against a known, "uninteresting" hypothesis. Rejecting the null lets us know that the parameter of interest is "interesting".
- Bayesians assume that the hypothesis space is finite. By the normalization result, Bayesians define the hypothesis space as a discrete number of possibilities and calculate the probability that each hypothesis is correct.
- Assume that each model tests a unique hypothesis. We can compare models and select the best one given the data.
- The marginal likelihood, $P(X)$, is used as the assessment of the quality of a model. As $P(X)$ increases, the model does a better job of fitting the data.
- The key here is that $P(X)$ marginalizes the estimated parameters over the priors.

Bayesian Model Comparison

- Model comparison utilizes a Bayes Factor. When comparing two models (M_1 and M_2), the Bayes Factor is:

$$BF_{1:2} = \frac{P(X|M_1)}{P(X|M_2)}$$

- A Bayes Factor greater than 1 favors M_1. A bigger value has a higher probability of being correct.
- Given that a Bayes Factor expresses the odds that M_1 is better than M_2, we can extend the Bayes Factor to probabilities.
- Similarly, we can do multiple pairwise comparisons and compute probabilities for more than 2 models.

Estimation of Posterior Distributions

- Sometimes, marginalizing the joint posterior distribution is very simple. We care about individual parameters, not the effect of the full system of parameters.
- When the system is low-dimensional and mathematically simple, we can calculate marginal posterior densities analytically.
- Most of the time, however, this is not possible:
  - As the number of parameters being estimated increases, the level of dependence between each estimated parameter is generally increasing. This dependence makes integration over the joint posterior difficult.
  - Asymptotic approximations require using frequentist tools that we are trying to avoid.
- There are a number of methods for numerical integration of the joint posterior.
- When the number of parameters of interest is higher than ≈ 5, posterior sampling is the best approach.

MCMC Methods

- Sampling from the posterior distribution is relatively simple utilizing Markov Chain - Monte Carlo methods.
- Given the specification of our likelihood functions and priors, we can utilize the converging properties of Markov Chains to start from any point in the space of θ and after many draws from the proposal kernel, we eventually end up in the posterior space.
- The most common method of posterior sampling is called Gibbs Sampling.

MCMC Methods

- Gibbs Sampling uses the following recipe:
  - Determine the structure of the join posterior distribution.
  - Specify the set of full conditionals:

$$P(\theta_i|\theta_{-i})$$

  - Specify starting points for each parameter in θ.
  - For the first parameter, take a single sample from the conditional distribution holding θ_{-i} at their respective starting values. Record this value.
  - For all other parameters, set θ_i equal to the most recently drawn sample. Draw from the conditional. Record these values.
  - Repeat this process many times.
  - At the stopping point, find the point where the posterior samples converge to a stationary distribution and lose any samples before that.
  - Boom. Samples from the marginal posteriors.

MCMC Methods

- How do we assess convergence?
- Markov Chains (like Gibbs Sampling) will always converge to a stationary distribution, but we can't really calculate how long it will take to get there.
- That's a tough question. It's really more of an art than a science.
- See Geweke diagnostic, Gelman's PSRF, etc.
- Minimize the burn-in time by selecting good initial values.

MCMC Methods

- Gibbs Sampling is a very powerful tool in machine learning and Bayesian inference.
- As long as we can specify the joint posterior and calculate full conditionals, we can avoid having to do an integral for each parameter in θ.
- Gibbs Sampling and Metropolis-Hastings are very flexible and can be used to calculate probability distributions for very complicated models.
- A benefit of this procedure is that the full conditionals are very predictable given reasonable distributional assumptions. This led to distribution of canned Bayesian estimation software. See WinBUGS and JAGS.

The Evangelizing Slide

- Bayesian inference provides a lot of niceties for statistical analysis:
  - Provides probabilities for hypotheses
  - Simple interpretation
  - Explicit assumptions
  - Marginalizes nuisance parameters
  - Model comparisons for more than 2 nested or non-nested models
  - Automatic overfitting penalties via Occam's factors
  - Valid for all sample sizes
  - Handles multimodality
  - Accounts for prior information and tests
  - Does not suffer from early stopping of experiments
  - Provides consistent, calibrated estimators
  - Good coverage (frequently better than frequentist analogues)

Back to LVMs

- Why go through all of that Bayes stuff?
- With MCMC methods and a Bayesian point of view, we have a general method of estimating latent variable models of any form:
  1. Determine level of measurement and distribution of the latent variable
  2. Define priors for θ
  3. Calculate joint posterior, $P(\theta|Z, X)$
  4. Work out full conditional set for θ
  5. Figure out $P(Z|X, \theta)$
  6. Use Gibbs Sampling to simulate from marginal posteriors.
- Of course, this is more difficult than this list makes it seem.
- Fortunately, this is a very rich area of research in statistics.

What LVMs Exist?

- A number of common LVMs are utilized in research across disciplines.
- The type of LVM utilized is conditional on the level of measurement of the desired latent variable and the level of measurement of the covariates, X.
  1. Continuous X, Continuous Z: Factor Analysis
  2. Discrete X, Continuous Z: Item-Response Models
  3. Continuous X, Discrete Z: Mixture Models
  4. Discrete X, Discrete Z: Latent Class Analysis
- While each type of LVM deserves a week in this lecture series, this week we will be focusing on a hierarchical version of Latent Class Analysis called Latent Dirichlet Allocation.

LDA Intro

- Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives of text:
  1. Discover the hidden themes in the collection
  2. Annotate the documents according to these themes
  3. Use annotations to organize, summarize, and form predictions.
- Due to the size and complexity of text data (speeches, books, papers, etc.), we need unsupervised probabilistic models.

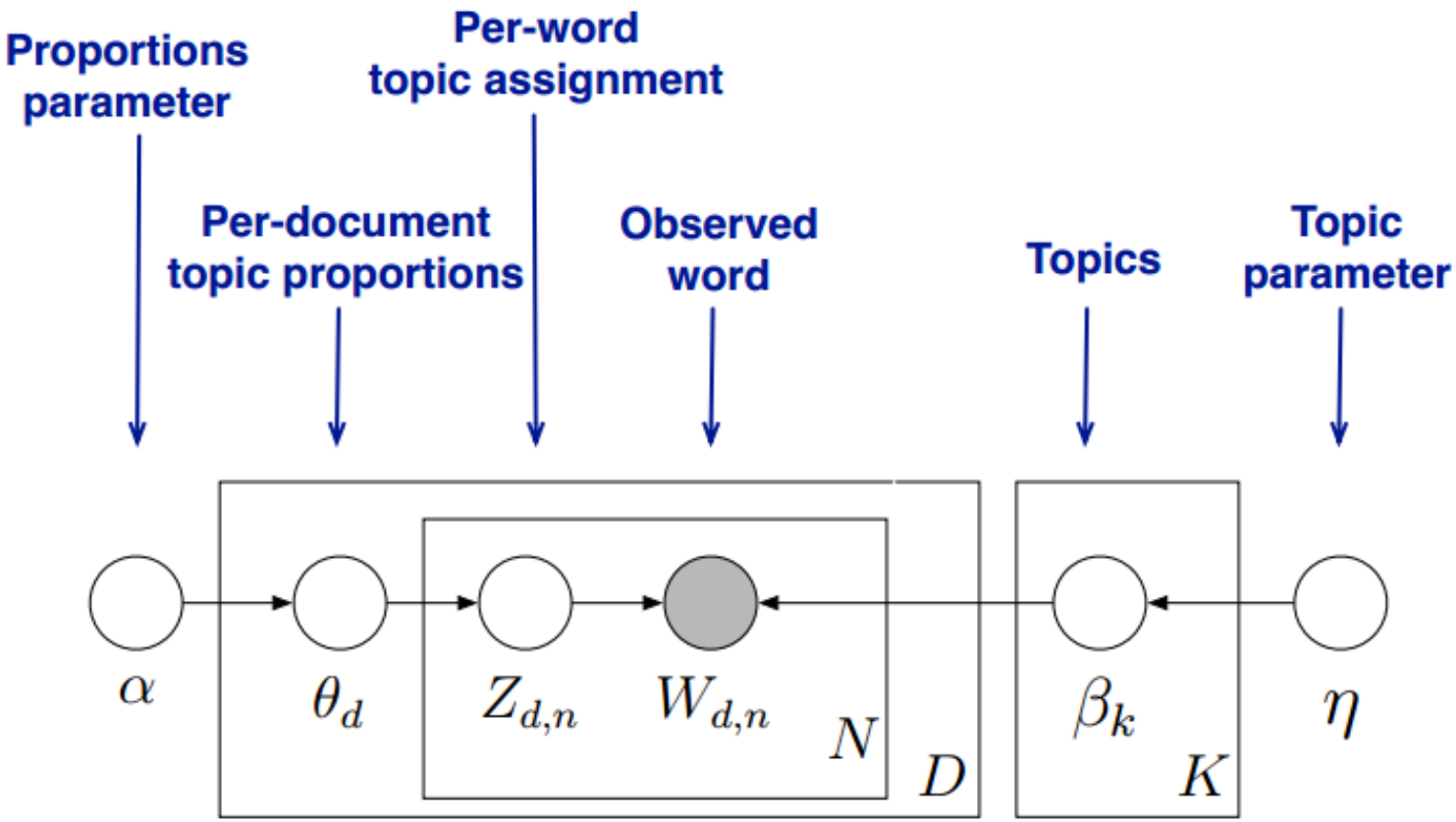
LDA Assumptions

- We have a set of documents $D_1, D_2, ..., D_D$.
- Each document is just a "bag of words". The order of the words and the grammatical role of the words are not considered in the model.
- Common words that carry no substantive meaning (e.g. am, is, are, of, the, but, etc.), can be eliminated eliminated from the documents in a preprocessing step.
- Similarly, we should eliminate words that appear in most of the documents in the collection. Words that $> 90\%$ of the documents carry no disriminative power.
- Each document, $d \in [1 : D]$, is composed of N important words.
- We fix the number of topics to be modeled at K.

Model Definition

- Each topic is a distribution over words.
- Each document is a mixture of topics.
- Each word is drawn from one of these topics.
- We only observe the words within the documents and the other structures are latent variables.

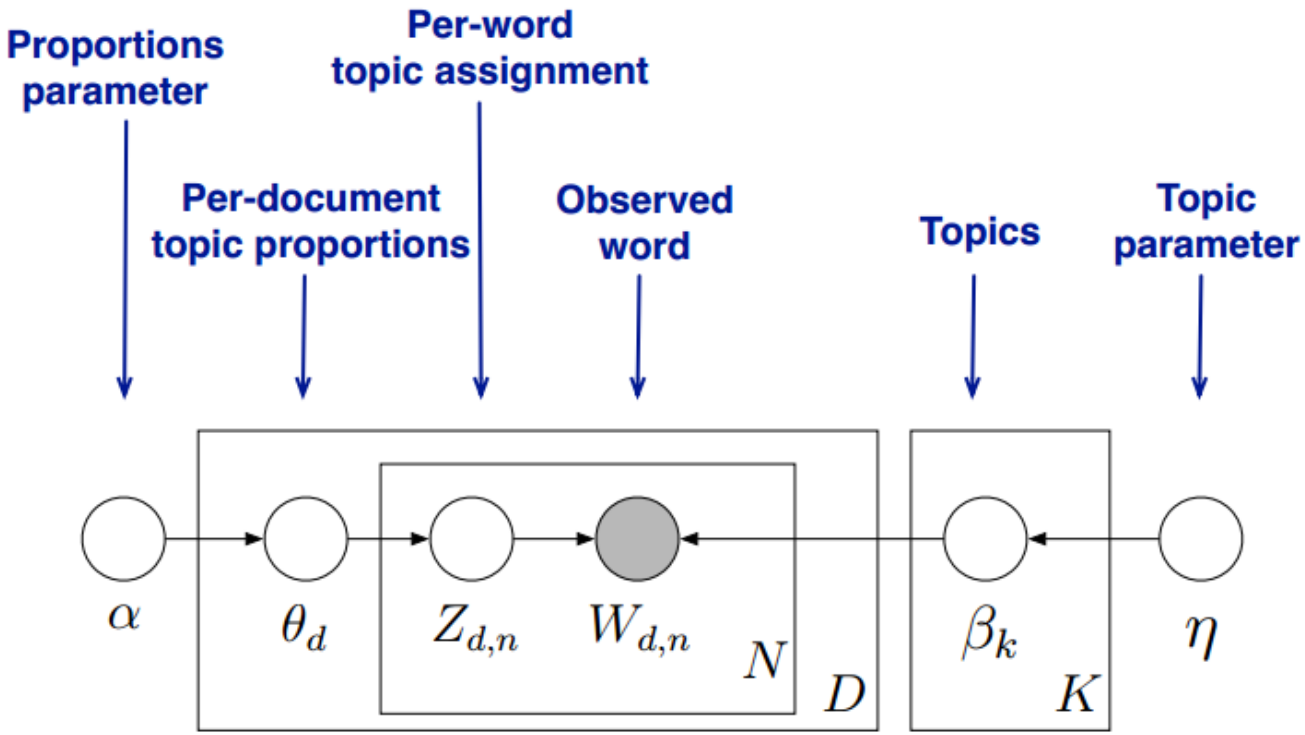
Model Specification



Model Specification

- Out goal is to infer or estimate the latent variables, i.e. computing the distribution conditioned on the dosuments.
- $P(\textit{topics}, \textit{proportions}, \textit{assignment} | \textit{documents})$
- Nodes are random variables, edges indicate dependence.
- Shaded nodes are observed.
- Unshaded nodes are latent.
- Plates indicate replicated variables.

Model Specification



$$p(\beta, \theta, \mathbf{z}, \mathbf{w}) = \left(\prod_{i=1}^{K} p(\beta_i \mid \eta) \right) \left(\prod_{d=1}^{D} p(\theta_d \mid \alpha) \prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right)$$

Model Specification

- Draw each topic $\beta_k \sim Dir(\eta)$ for $i = [1, k]$.
- For each document:
  1. First, draw topic proportions $\theta_d \sim Dir(\alpha)$
  2. For each word within the document:
     - Draw $Z_{d,n} \sim Multi(\theta_d)$
     - Draw $W_{d,n} \sim Multi(\beta_{z_{d,n}})$
- This defines the joint posterior, $p(\theta, z, \beta|w)$.
- We have to infer:
  1. Per-word topic assignment $z_{d,n}$
  2. Per-document topic proportions θ_d
  3. Per-corpus topic distributions β_k

Model Estimation

- From a set of N_d documents and the observed words within each document, we want to infer the posterior distribution, $P(\theta, z, \beta|w)$.
- Bayes! Use Gibbs Sampling.
- Denote ϕ as the collection of model parameters and X as the observed data. This gives us:

$$p(\phi|X) = \frac{p(x|\phi)p(\phi)}{p(X)}$$

- Note that calculating $p(X)$ is really difficult and that the number of inferred parameters is potentially very large. We need simulational methods to make this work.

Model Estimation

- Applying Gibbs Sampling to estimating $p(\theta, z|w, \beta)$:
  1. $p(\theta|z_{1:N}) = Dir(\alpha + c(z_{1:N}))$ where $c()$ is a vector of the count of each topic.
  2. $p(z_i|z_{-i}, \theta, w_{1:N}) \propto p(z_i|\theta)p(w_i|\beta_{1:K}, z_i)$
- Note that we have assumed that $\beta_{1:K}$ is fixed in the above Gibbs Sampler. We can use more complex Gibbs to infer β, as well.

LDA in R

- Text Mining Package in R: library("tm")
- Topic Models in R: library("topicmodels")
- We can perform preprocessing steps using tm_map() in R. This removes unimportant words and can create frequency tables to remove very low discrimination words from the documents.
- LDA function in R is in the topicmodels package. It can fit the LDA models for a specific number of topics. K is predefined before running the code.

LDA Example

- All issues of *Science* from 1990 - 2000
- 17k documents
- 11 millions words
- 20k unique meaningful words
- Model: 100-topic LDA model

LDA Example

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations