Text Mining for Economics and Finance
Bayesian Inference for Discrete Data

Stephen Hansen

Introduction

Models like pLSI have a huge number of parameters to estimate.

As we shall see, one popular approach to estimation in high-dimensional spaces is Bayesian inference in which parameters are treated as random variables drawn from a prior distribution.

In this lecture, we take the first step of looking at simple Bayesian models for discrete data.

Review

Recall the simple unigram model of a document in which

$$\Pr[\mathbf{x}_d \mid \boldsymbol{\beta}] = \prod_v \beta_v^{x_{d,v}}.$$

The maximum likelihood estimate for the vth categorical probability is $\widehat{\beta}_v = \frac{x_{d,v}}{N_d}$

To maximize the probability of the observed data, we get parameters that exactly match the observed frequencies.

Black Swan Paradox

What would the model predict is the probability of seeing an unobserved term?

This is sometimes called the *black swan paradox*. Europeans assumed that the fact that they had never observed a black swan implied black swans could not exist.

Since the document-term matrix is sparse, the black swan paradox will be particularly relevant for text mining.

Bottom line is we need to incorporate some additional uncertainty in our inference procedure to not drive beliefs about unobserved events to zero.

Bayesian Inference

One solution is to adopt a Bayesian inference approach, which treats $\boldsymbol{\beta}$ as a random variable rather than a fixed parameter.

Recall that Bayes' rule states that

$$\Pr[\boldsymbol{\beta} \mid \mathbf{x}_d] = \frac{\Pr[\mathbf{x}_d \mid \boldsymbol{\beta}] \Pr[\boldsymbol{\beta}]}{\Pr[\mathbf{x}_d]}$$

where

- $\Pr[\boldsymbol{\beta} \mid \mathbf{x}_d]$ is the posterior distribution.
- $\Pr[\mathbf{x}_d \mid \boldsymbol{\beta}]$ is the likelihood function.
- $\Pr[\boldsymbol{\beta}]$ is the prior distribution on the parameter vector.
- $\Pr[\mathbf{x}_d]$ is a normalizing constant sometimes called the evidence.

The prior distribution introduces initial uncertainty about the value of the parameter vector.

Dirichlet Prior

One way of ensuring Bayesian inference is tractable is to select a prior distribution from a family that ensures the posterior will be in the same family given the likelihood function. This is called a *conjugate* prior.

The Dirichlet distribution is conjugate to the categorical likelihood function, and so is a popular choice for the prior in Bayesian models of discrete data.

The Dirichlet distribution is parametrized by $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_V)$; is defined on the $V - 1$ simplex; and has probability density function

$$\text{Dir}(\boldsymbol{\beta} \mid \boldsymbol{\alpha}) \propto \prod_v \beta_v^{\alpha_v - 1}.$$

The normalization constant is $B(\boldsymbol{\alpha}) \equiv \prod_{v=1}^{V} \Gamma(\alpha_v) / \Gamma\left(\sum_{v=1}^{V} \alpha_v\right)$.

Interpreting the Dirichlet

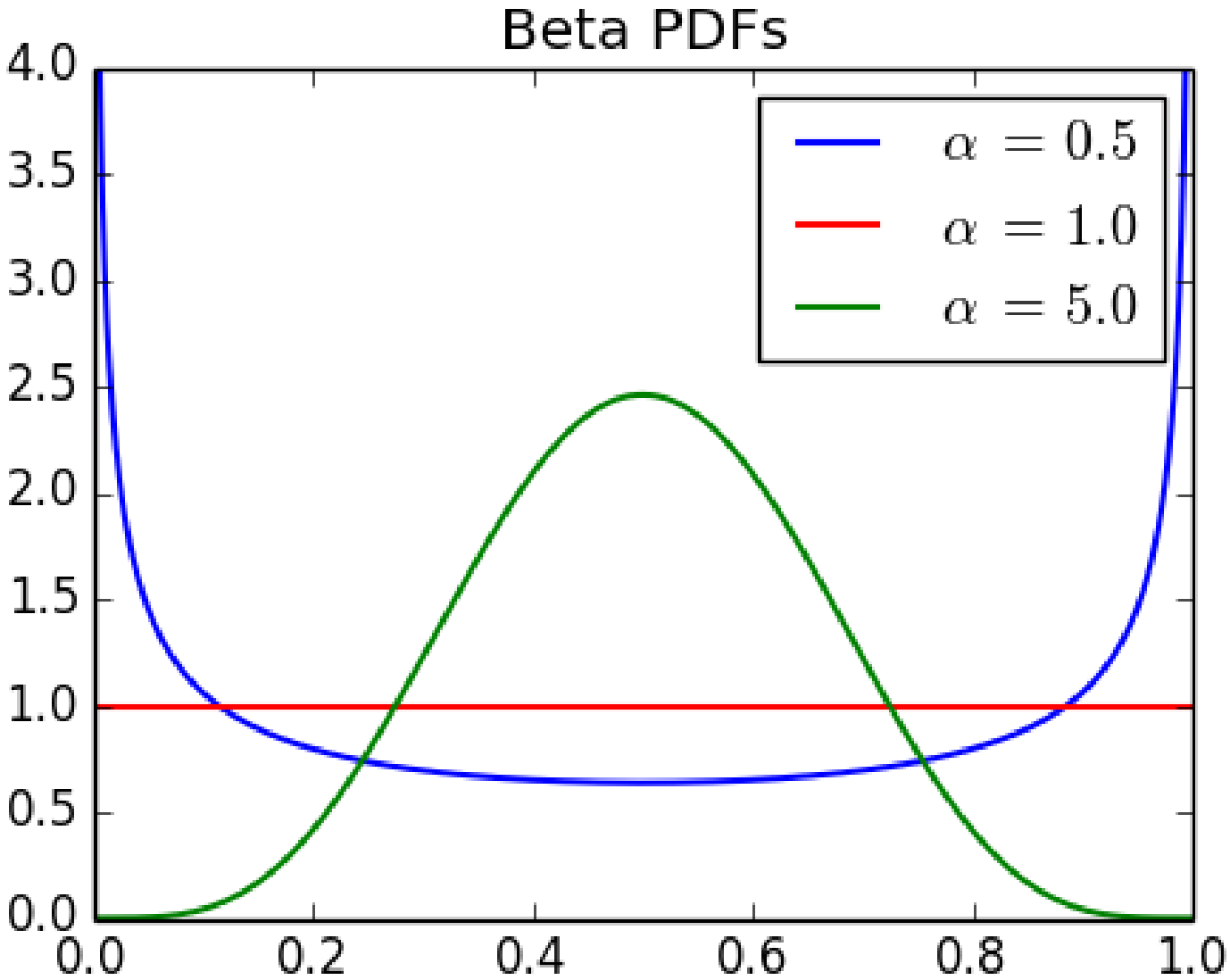
Consider a symmetric Dirichlet in which $\alpha_v = \alpha$ for all v. Agnostic about favoring one component over another.

Here the α parameter measures the concentration of distribution on the center of the simplex, where the mass on each term is more evenly spread out:

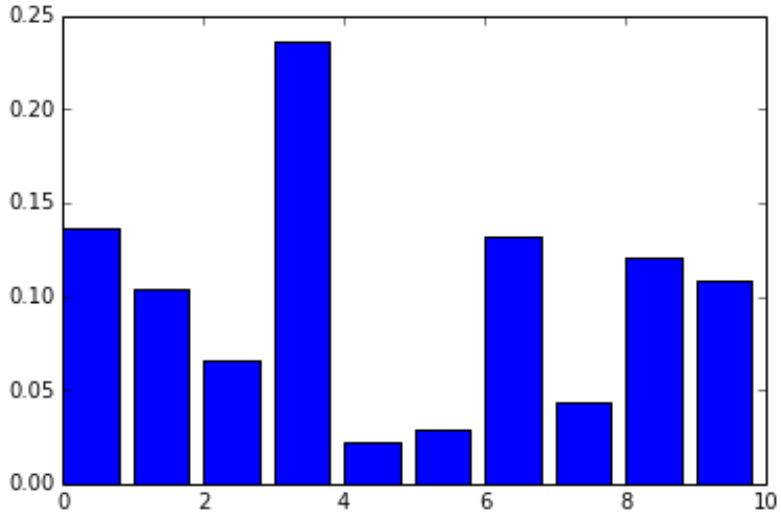
1. $\alpha = 1$ is a uniform distribution.
2. $\alpha > 1$ puts relatively more weight in center of simplex.
3. $\alpha < 1$ puts relatively more weight on corners of simplex.

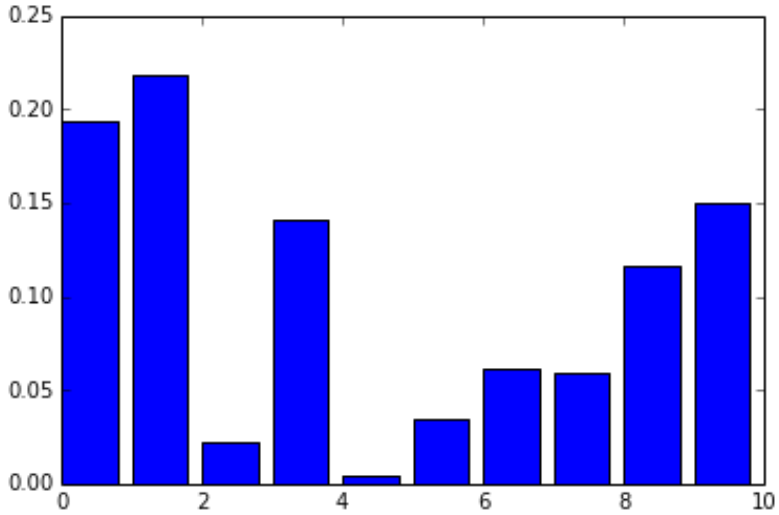
When $V = 2$, the Dirichlet becomes the beta distribution.

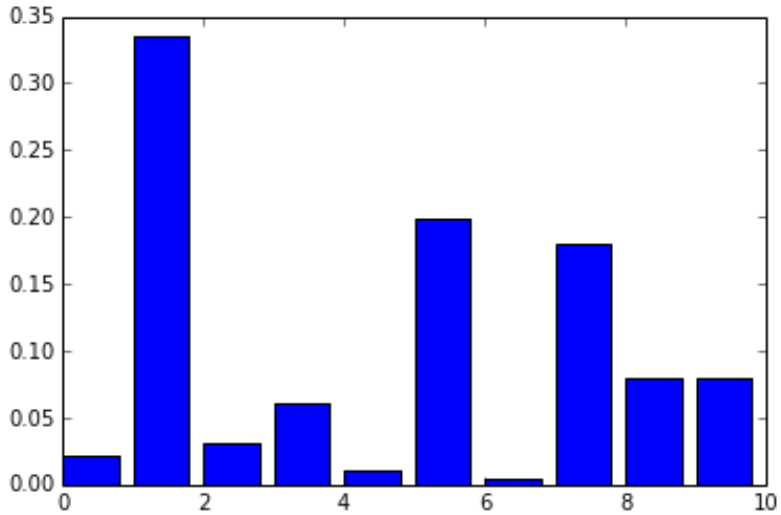
Beta with Different Parameters

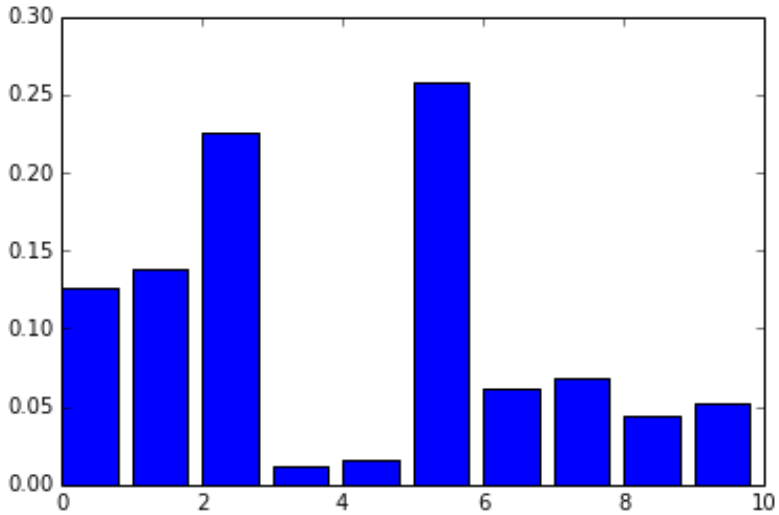


Draws from Dirichlet with $\alpha = 1$

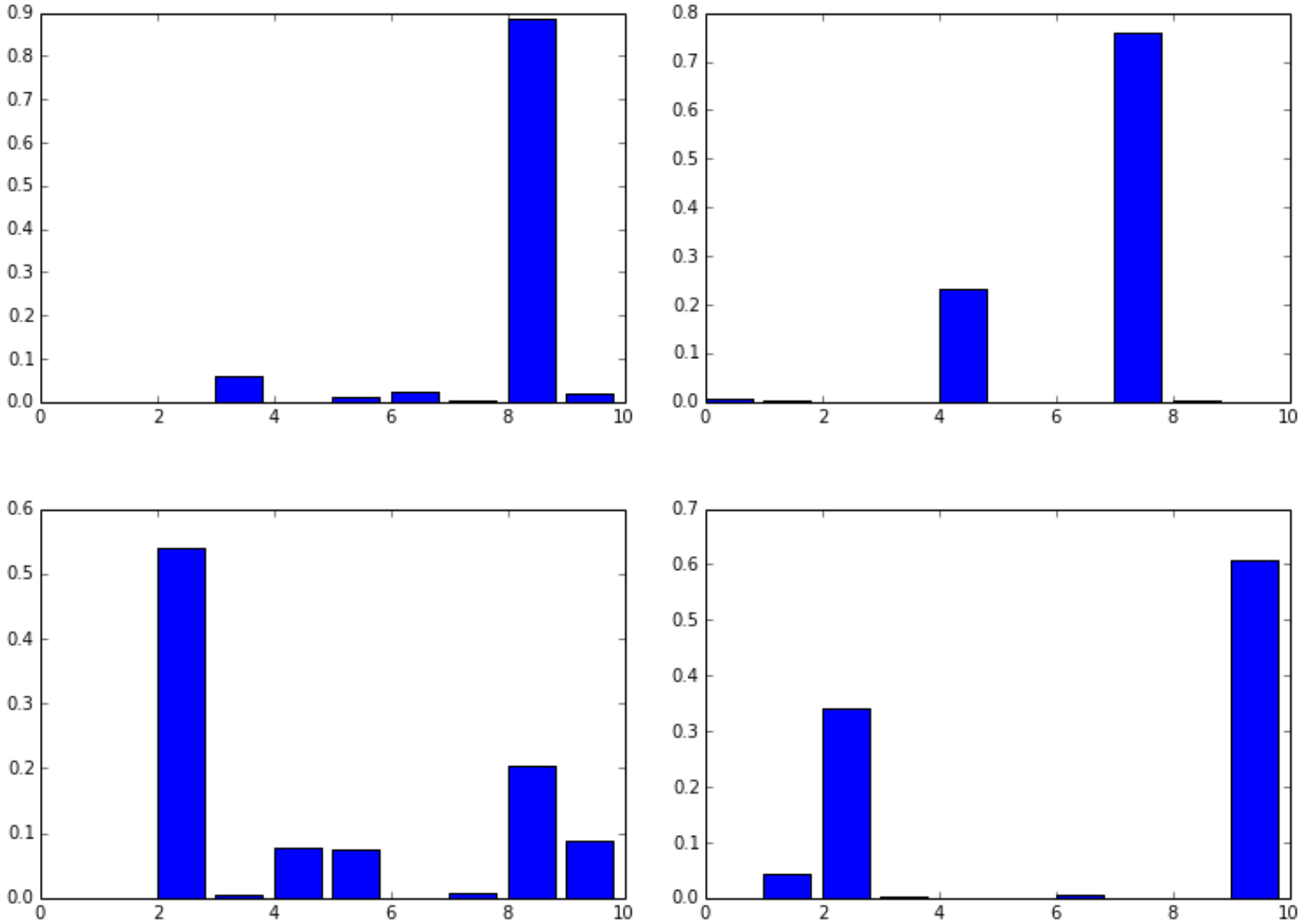




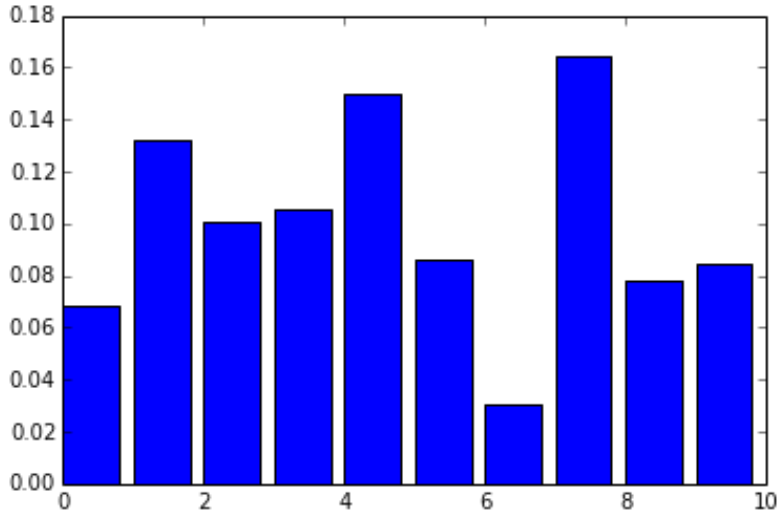


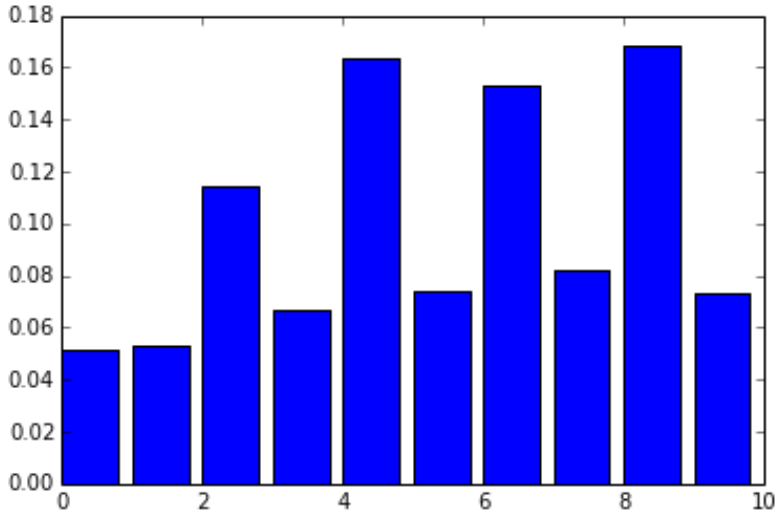


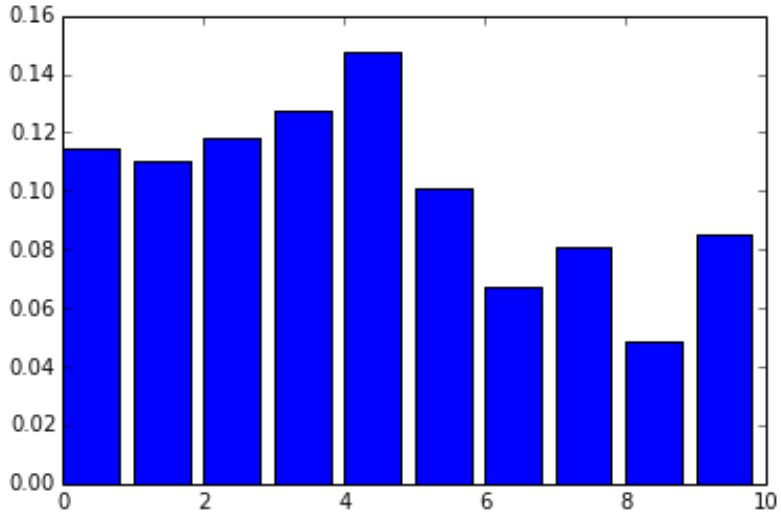
Draws from Dirichlet with $\alpha = 0.1$

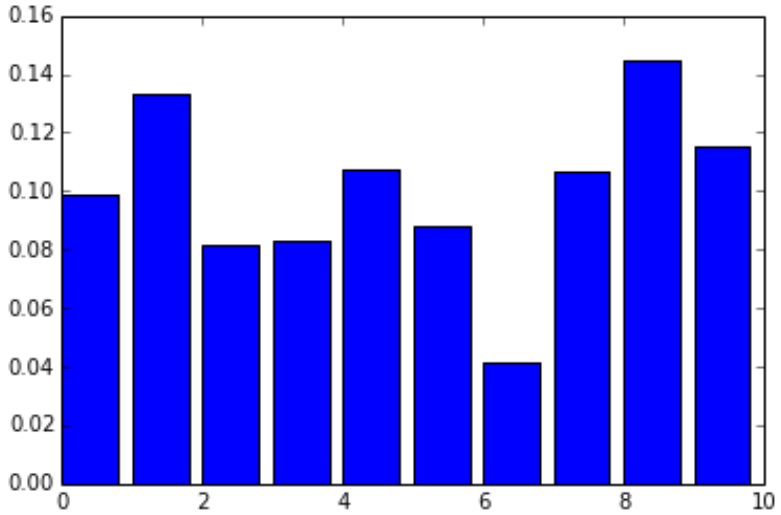


Draws from Dirichlet with $\alpha = 10$









Posterior Distribution

$$\Pr[\boldsymbol{\beta} \mid \mathbf{x}_d] \propto \Pr[\mathbf{x}_d \mid \boldsymbol{\beta}] \Pr[\boldsymbol{\beta}] \propto \prod_{v=1}^{V} \beta_v^{x_{d,v}} \prod_{v=1}^{V} \beta_v^{\alpha_v - 1} = \prod_{v=1}^{V} \beta_v^{x_{d,v} + \alpha_v - 1}.$$

Posterior is a Dirichlet with parameters $(\widehat{\alpha}_1, \ldots, \widehat{\alpha}_V)$ where $\widehat{\alpha}_v \equiv \alpha_v + x_{d,v}$.

Add term counts to the prior distribution's parameters to form posterior distribution.

The parameters in the prior distribution are sometimes called *pseudo-counts*, and can be viewed as observations made before $\mathbf{x}_d$.

Moments of the Posterior

One can show that a Dirichlet with parameter vector $\boldsymbol{\alpha}$ satisfies

$$\mathbb{E}[\,\beta_v\,] = \frac{\alpha_v}{\alpha} \text{ and } V[\,\beta_v\,] = \frac{\alpha_v(\alpha - \alpha_v)}{\alpha^2(\alpha+1)}, \text{ where } \alpha \equiv \sum_v \alpha_v.$$

If we apply these formulas to the posterior distribution we obtain

$$\mathbb{E}[\,\beta_v\,] = \frac{\alpha_v + x_{d,v}}{\alpha + N_d} \text{ and } V[\,\beta_v\,] = \frac{(\alpha_v + x_{d,v})(\alpha + N_d - \alpha_v - x_{d,v})}{(\alpha + N_d)^2(\alpha + N_d + 1)}.$$

One can also show that the mean corresponds to the predictive distribution for an additional word.

Data Overwhelming the Prior

Recall the MLE estimates for $\widehat{\beta}_v$ satisfies $N_d \widehat{\beta}_v = x_{d,v}$. We then have

$$\mathbb{E}[\,\beta_v\,] = \frac{\alpha_v + N_d \widehat{\beta}_v}{\alpha + N_d} \text{ and } V[\,\beta_v\,] = \frac{(\alpha_v + N_d \widehat{\beta}_v)(\alpha + N_d - \alpha_v - N_d \widehat{\beta}_v)}{(\alpha + N_d)^2(\alpha + N_d + 1)}.$$

If we take the limit as $N_d \to \infty$, we obtain a degenerate posterior distribution concentrated fully on the MLE parameter estimates.

Intuition: the more data we see, the less our priors should influence our beliefs.

Predictive Distribution

Recall the predictive distribution is a distribution over new data given observed data (rather than the unknown parameter $\boldsymbol{\beta}$).

What's the probability that a $N_d + 1$th word drawn for document d is term v?

$$\Pr[\, w_{d,N_d+1} = v \mid \mathbf{w}_d \,] = \int \Pr[\, w_{d,N_d+1} = v \mid \boldsymbol{\beta} \,] \Pr[\, \boldsymbol{\beta} \mid \mathbf{w}_d \,] d\boldsymbol{\beta} = \int \beta_v \Pr[\, \boldsymbol{\beta} \mid \mathbf{w}_d \,] d\boldsymbol{\beta},$$

which is simply the expectation of β_v computed under the posterior Dirichlet, or

$$\mathbb{E}[\, \beta_v \mid \mathbf{w}_d \,] = \frac{\alpha_v + x_{d,v}}{\alpha + N_d}.$$

Pseudo-counts act to smooth predictive likelihood and relax black swan paradox.

Conclusion

In MLE we treat parameters as constants, and choose them to maximize the log-likelihood function. In Bayesian estimation, we treat them as random variables and compute a posterior distribution given observed data.

In models with a large number of parameters, Bayesian inference can be more robust and avoids over-sensitivity to sparse data.

Given a large amount of data, MLE and Bayesian approaches become equivalent.