Machine Learning Methods for Economists
Probability Models for Discrete Data

Stephen Hansen
University of Oxford

Text and Unstructured Data

Most empirical work in economics relies on inherently quantitative data: prices, demand, votes, etc.

But a large amount of unstructured text is also generated in economic environments: company reports, policy committee deliberations, court decisions, media articles, political speeches, etc.

One can use such data qualitatively, but increasing interest in treating text quantitatively.

We shall also see that the empirical analysis of text is part of a more general problem of treating high-dimensional count data, and discuss other applications.

Terminology

A single observation in a textual database is called a *document*.

The set of documents that make up the dataset is called a *corpus*.

We often have covariates associated with each document that are sometimes called *metadata*.

FOMC Example

Running example is corpus of verbatim FOMC transcripts from the era of Alan Greenspan:

- 149 meetings from August 1987 through January 2006.
- A document is a single statement by a speaker in a meeting (46,502).
- Baseline data has 6,249,776 total words and 26,030 unique words.
- Metadata include macro conditions, speaker characteristics, etc.

Notation

The corpus is composed of D documents indexed by d.

After pre-processing, each document is a finite, length-N_d list of terms $\mathbf{w}_d = (w_{d,1}, \ldots, w_{d,N_d})$ with generic element $w_{d,n}$.

Let $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_D)$ be a list of all terms in the corpus, and let $N \equiv \sum_d N_d$ be the total number of terms in the corpus.

Suppose there are V **unique** terms in $\mathbf{w}$, where $1 \leq V \leq N$, each indexed by v.

We can then map each term in the corpus into this index, so that $w_{d,n} \in \{1, \ldots, V\}$.

Let $x_{d,v} \equiv \sum_n \mathbb{1}(w_{d,n} = v)$ be the count of term v in document d.

Example

Consider three documents:

1. 'stephen is nice'
2. 'john is also nice'
3. 'george is mean'

We can consider the set of unique terms as $\{\text{stephen}, \text{is}, \text{nice}, \text{john}, \text{also}, \text{george}, \text{mean}\}$ so that $V = 7$.

Construct the following index:

stephen	is	nice	john	also	george	mean
1	2	3	4	5	6	7

We then have $\mathbf{w}_1 = (1, 2, 3)$; $\mathbf{w}_2 = (4, 2, 5, 3)$; $\mathbf{w}_3 = (6, 2, 7)$.

Moreover $x_{1,1} = 1$, $x_{2,1} = 0$, $x_{3,1} = 0$, etc.

Document-Term Matrix

A popular quantitative representation of text is the *document-term matrix* $\mathbf{X}$, which collects the counts $x_{d,v}$ into a $D \times V$ matrix.

In the previous example, we have

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

The key characteristics of the document-term matrix are its:

1. High dimensionality
2. Sparsity

Ngram Models

In the above example, we made an explicit choice to count individual terms, which destroys all information on word order.

In some contexts, this may be sufficient for our information needs, but in others we might lose valuable information.

We could alternatively have counted all adjacent two-term phrases, called bigrams or, more generally, all adjacent N-term phrases, called Ngrams.

This is perfectly consistent with the model described above, where v now indexes unique bigrams rather than unique unigrams:

stephen.is	is.nice	john.is	is.also	also.nice	george.is	is.mean
1	2	3	4	5	6	7

We then have $\mathbf{w}_1 = (1, 2)$; $\mathbf{w}_2 = (3, 4, 5)$; $\mathbf{w}_3 = (6, 7)$.

Alternative Application I: Survey Data

Ongoing project to document CEO time use (with O. Bandiera, A. Prat, and R. Sadun), and its effect on firm performance.

Data on each 15-minute block of time for one week of 1,114 CEOs' time classified according to

1. type (e.g. meeting, public event, etc.)
2. duration (15m, 30m, etc.)
3. planning (planned or unplanned)
4. number of participants (one, more than one)
5. functions of participants, divided between employees of the firms or "insiders" (finance, marketing, etc.) and "outsiders" (clients, banks, etc.).

There are 4,253 unique combinations of these five features in the data.

One can summarize the data with a 1114×4253 matrix where the (i, j)th element is the number of 15-minute time blocks that CEO i spends in activities with a particular combination of features j.

Alternative Application II: Demand

Increasingly common to get detailed data on consumers' shopping behavior.

Imagine a dataset in which consumer $i \in \{1, \ldots, N\}$ makes multiple visits to a store that sells M possible goods bundles.

Then we can define an $N \times M$ matrix that counts the number of times each consumer buys each bundle.

Dimensionality Reduction

A large body of literature in text mining seeks to reduce the dimensionality of the document-term matrix.

Common approach in economics is to use dictionary methods, or counts of key words (see for example Baker, Bloom and Davis 2016).

However these have some limitations:

1. Focus on very narrow subset of dimensions of variation.
2. Unclear which are the relevant words in many cases.

Without any strong theoretical guidance on which are the important dimensions of variation, ideally we would let the data speak.

Probability Models

As a stepping stone for building dimensionality reduction algorithms, we will introduce basic models for the probabilistic modeling of discrete data in the rest of the lecture.

Probability models are sometimes called *generative* models in the machine learning literature.

In supervised learning, generative models allow us to model the full joint distribution $p(y_d, \mathbf{x}_d)$, which we revisit in the final lecture.

In unsupervised learning, generative models allow us to given a statistical interpretation to the hidden structure in a corpus.

For now, we ignore document heterogeneity, and instead introduce models that will form the building blocks for probabilistic unsupervised learning.

Simple Probability Model

Consider the list of terms $\mathbf{w} = (w_1, \ldots, w_N)$ where $w_n \in \{1, \ldots, V\}$.

Suppose that each term is iid, and that $\Pr[\, w_n = v \,] = \beta_v \in [0, 1]$.

Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_V) \in \Delta^{V-1}$ be the parameter vector we want to estimate.

The probability of the data given the parameters is

$$\Pr[\, \mathbf{w} \mid \boldsymbol{\beta} \,] = \prod_n \sum_v \mathbb{1}(w_n = v)\beta_v = \prod_v \beta_v^{x_v}$$

where x_v is the count of term v in $\mathbf{w}$.

Note that term counts are a sufficient statistic for $\mathbf{w}$ in the estimation of $\boldsymbol{\beta}$. The independence assumption provides statistical foundations for the bag-of-words model.

Maximum Likelihood Inference

We can estimate $\boldsymbol{\beta}$ with maximum likelihood. The Lagrangian is

$$\mathfrak{L}(\boldsymbol{\beta}, \lambda) = \underbrace{\sum_v x_v \log(\beta_v)}_{\text{log-likelihood}} + \lambda \underbrace{\left(1 - \sum_v \beta_v\right)}_{\text{Constraint on } \boldsymbol{\beta}}.$$

First order condition is $\frac{x_v}{\beta_v} - \lambda = 0 \Rightarrow \beta_v = \frac{x_v}{\lambda}$.

Constraint gives $\frac{\sum_v x_v}{\lambda} = 1 \Rightarrow \lambda = \sum_v x_v = N$.

So MLE estimate is $\widehat{\beta}_v = \frac{x_v}{N}$, the frequency of term v in list of terms.

Implications of MLE

Suppose you do not speak Portuguese, but someone lists for you 10,000 possible words the spoken language might contain.

You are then shown a single snippet of text 'eles bebem'. The parameters that best explain this data put $1/2$ probability each on 'eles' and on 'bebem' and 0 on every other possible word.

Is this a reasonable model? We 'know' that working languages contain hundreds of regularly spoken words; we 'know' that the distribution of word frequencies is highly skewed; we 'know' that the language is similar to Spanish, and should inherit a similar frequency distribution; and so on.

The MLE estimates relies solely on the data we observe.

More subtle problem is to take V to be the number of unique observations, which may be misleading even with large samples (*black swan paradox*).

Bayesian Inference

Bayesian inference treats β as a random variable drawn from a *prior distribution*, which can encode any knowledge we might have.

On the other hand, we treat the data as a fixed quantity that provides information about β.

The *likelihood principle* states that all relevant information about an unknown quantity θ is contained in the likelihood function of θ for the given data (Berger and Wolpert 1988).

Bayesian inference is consistent with the likelihood principle, frequentist reasoning need not be.

"Many Bayesians became Bayesians only because the LP left them little choice" (Berger and Wolpert 1988).

Bayes' Rule

Bayesian inference is operationalized via the application of Bayes' rule:

$$p\left(\boldsymbol{\beta} \mid \mathbf{w}\right) = \frac{p\left(\mathbf{w} \mid \boldsymbol{\beta}\right) p(\boldsymbol{\beta})}{p(\mathbf{w})}$$

where:

- $p\left(\boldsymbol{\beta} \mid \mathbf{w}\right)$ is the *posterior distribution*.
- $p\left(\mathbf{w} \mid \boldsymbol{\beta}\right)$ is the *likelihood function*.
- $p(\boldsymbol{\beta})$ is the *prior distribution* on the parameter vector.
- $p(\mathbf{w})$ is a normalizing constant sometimes called the *evidence*.

The prior is often parametrized by *hyperparameters*.

What is a Bayesian Estimate?

There are several ways of reporting the Bayesian estimate of $\boldsymbol{\beta}$:

1. MAP estimate is the value at which the posterior distribution is highest, i.e. its mode.
2. Expected value of $\boldsymbol{\beta}$ under the posterior.
3. Choose point estimate to minimize some expected loss function.
4. Compute credible interval $\Pr[\boldsymbol{\beta} \in A \mid \mathbf{w}]$ for some set A.

We are also sometimes interested in $\Pr[w_{N+1} \mid \mathbf{w}]$ for some unseen data w_{N+1}. This is called the *predictive distribution*.

All of these depend fundamentally on the posterior distribution.

If we can compute the posterior, we can do Bayesian inference.

Penalized Regression Revisited

Consider the parametric linear regression model in which $y_i \sim \mathcal{N}(\mathbf{x}_i^T \boldsymbol{\beta}, I_N \sigma^2)$ in which σ is known.

Suppose we draw each regression coefficient β_i from a normal prior $\mathcal{N}(0, \tau^2)$.

The posterior distribution over $\boldsymbol{\beta}$ is then proportional to $p(\, y \mid \beta, \mathbf{x}\,)\, p(\, \beta\,)$.

We know that $p(\, \mathbf{y} \mid \boldsymbol{\beta}, \mathbf{X}\,)\, p(\, \beta\,) \propto \prod_i \exp\left(-\frac{(y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2}{2\sigma^2}\right) \prod_j \exp\left(-\frac{\beta_j^2}{2\tau^2}\right)$.

MAP estimate can be obtained by minimizing

$$\sum_i \left(y_i - \mathbf{x}^T \boldsymbol{\beta}\right)^2 + \frac{\sigma^2}{\tau^2} \sum_j \beta_j^2$$

which is exactly the ridge regression model.

LASSO

Now suppose we draw each regression coefficient β_j from a Laplace prior so that $\Pr[\beta_j \mid \lambda] \propto \exp(-\lambda|\beta_j|)$.

The Laplace distribution has a spike at 0 which promotes sparsity.

The objection function for MAP estimation can be written

$$\mathrm{RSS}(\boldsymbol{\beta}) + \lambda \sum_j |\beta_j|.$$

Choosing Priors

A popular choice for the prior distribution is that it be *conjugate*, i.e. the posterior distribution belongs to the same parametric family as the prior. This facilitates analytic computation of the posterior.

All distributions in the exponential family have conjugate prior distributions...but are they meaningful?

The Dirichlet distribution is conjugate to the categorical/multinomial distributions (as we shall see).

Choosing Priors

A popular choice for the prior distribution is that it be *conjugate*, i.e. the posterior distribution belongs to the same parametric family as the prior. This facilitates analytic computation of the posterior.

All distributions in the exponential family have conjugate prior distributions...but are they meaningful?

The Dirichlet distribution is conjugate to the categorical/multinomial distributions (as we shall see).

When the conjugate prior is not sufficiently expressive, we can adopt another prior and simulate the posterior distribution.

For example, the log-normal distribution more naturally embeds dependence on covariates and correlation in text.

Choosing Priors

A popular choice for the prior distribution is that it be *conjugate*, i.e. the posterior distribution belongs to the same parametric family as the prior. This facilitates analytic computation of the posterior.

All distributions in the exponential family have conjugate prior distributions...but are they meaningful?

The Dirichlet distribution is conjugate to the categorical/multinomial distributions (as we shall see).

When the conjugate prior is not sufficiently expressive, we can adopt another prior and simulate the posterior distribution.

For example, the log-normal distribution more naturally embeds dependence on covariates and correlation in text.

Once we choose a prior, we still need to choose hyperparameters: can set at some value consistent with our domain-specific knowledge, or select them to maximize the evidence (empirical Bayes).

Dirichlet Prior

The Dirichlet distribution is parametrized by $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_V)$; is defined on the $V - 1$ simplex; and has probability density function

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

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

Marginal distribution is

$$\beta_v \sim \text{Beta}(\eta_v, \sum_v \eta_v - \eta_v)$$

Mean and variance are

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

Interpreting the Dirichlet

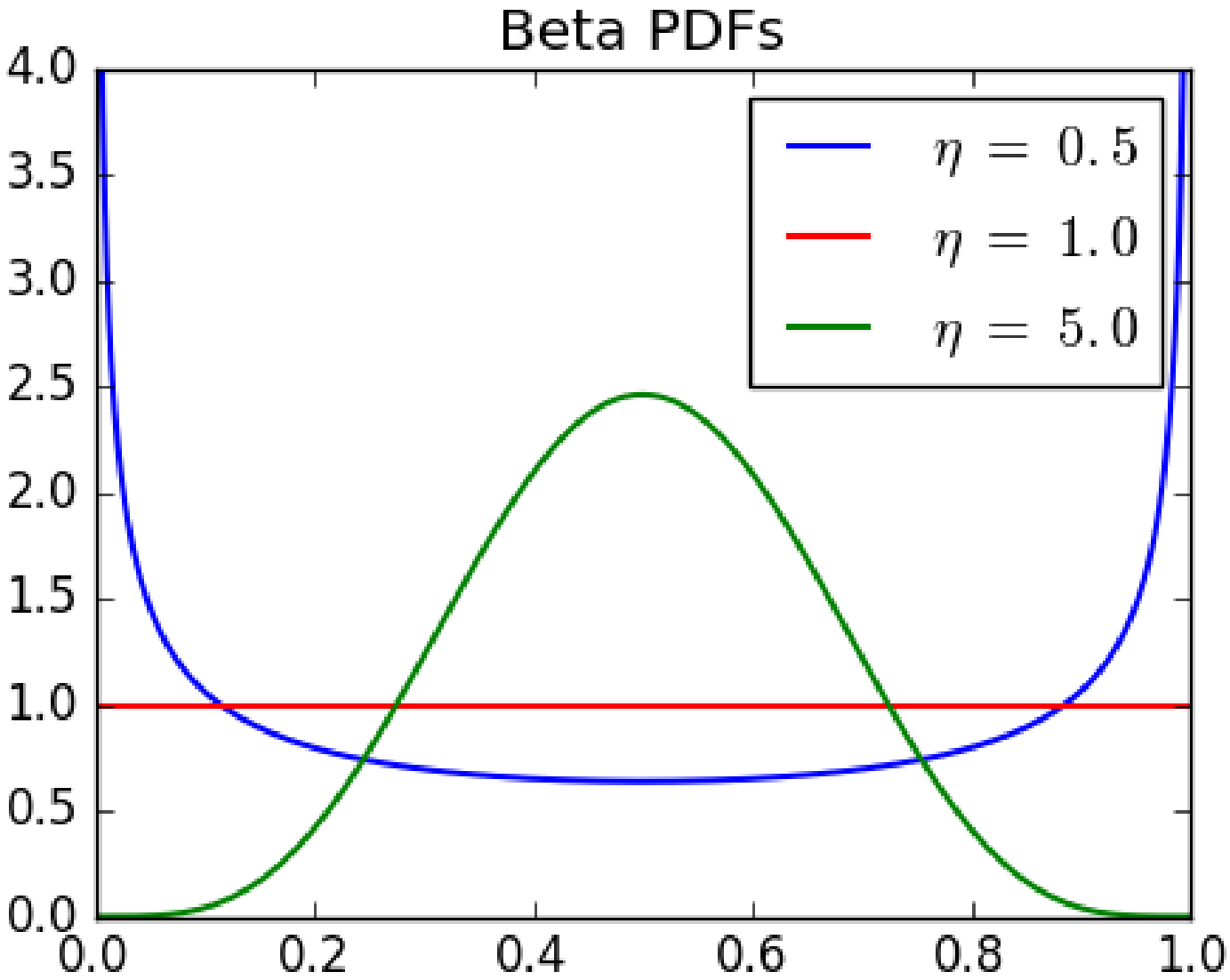
Consider a symmetric Dirichlet in which $\eta_v = \eta$ for all v. Agnostic about favoring one component over another.

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

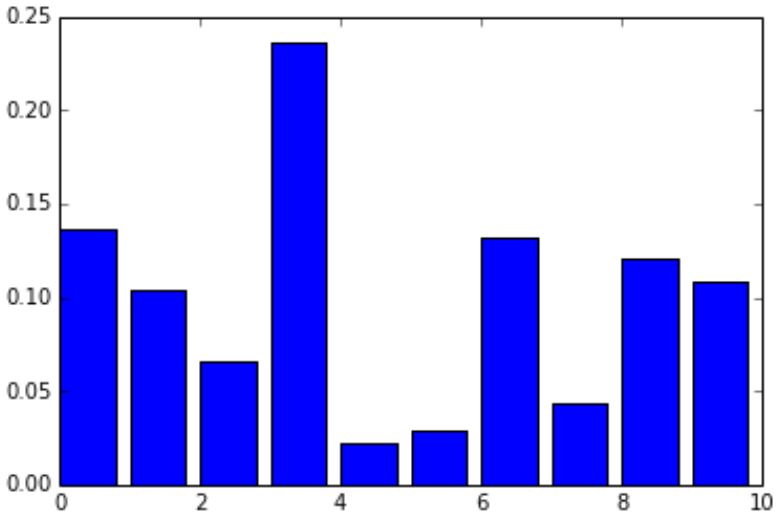
1. $\eta = 1$ is a uniform distribution.
2. $\eta > 1$ puts relatively more weight in center of simplex.
3. $\eta < 1$ puts relatively more weight on corners of simplex.

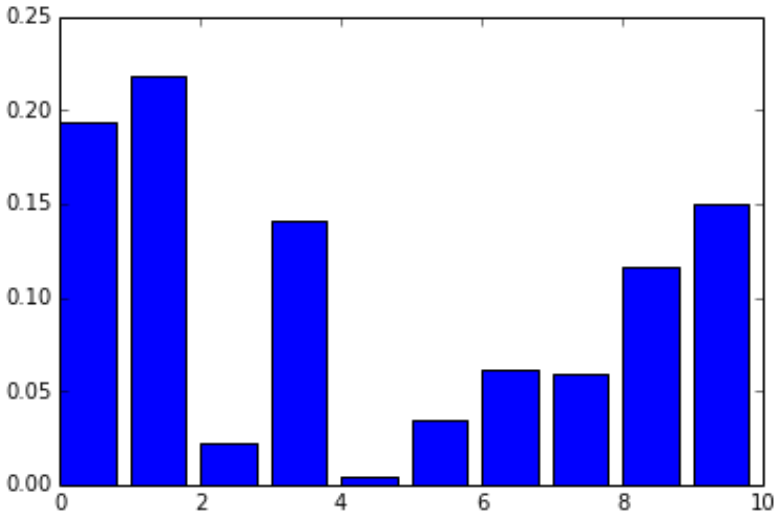
When $V = 2$, the Dirichlet becomes the beta distribution.

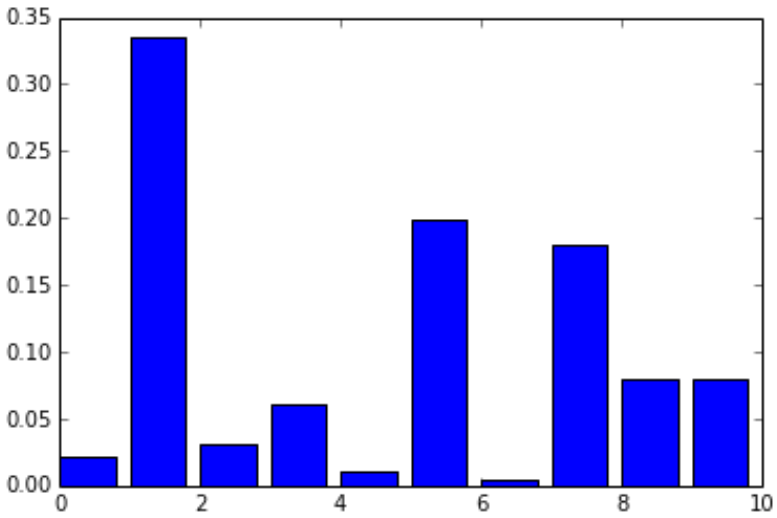
Beta with Different Parameters

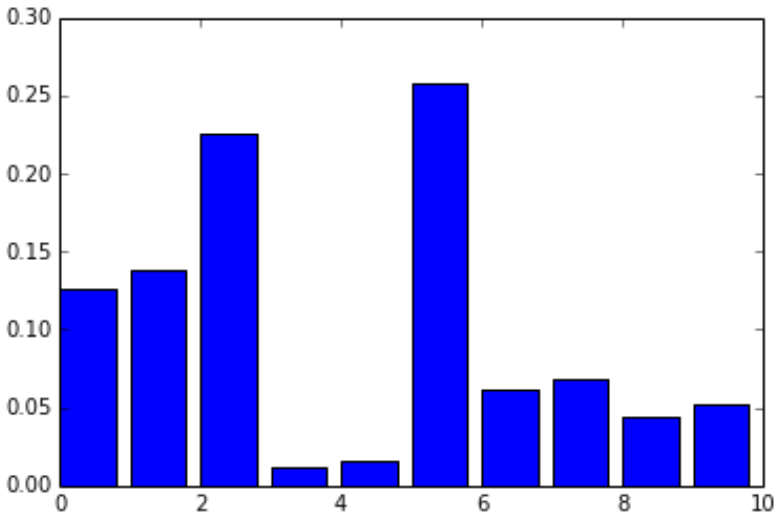


Draws from Dirichlet with $\eta = 1$

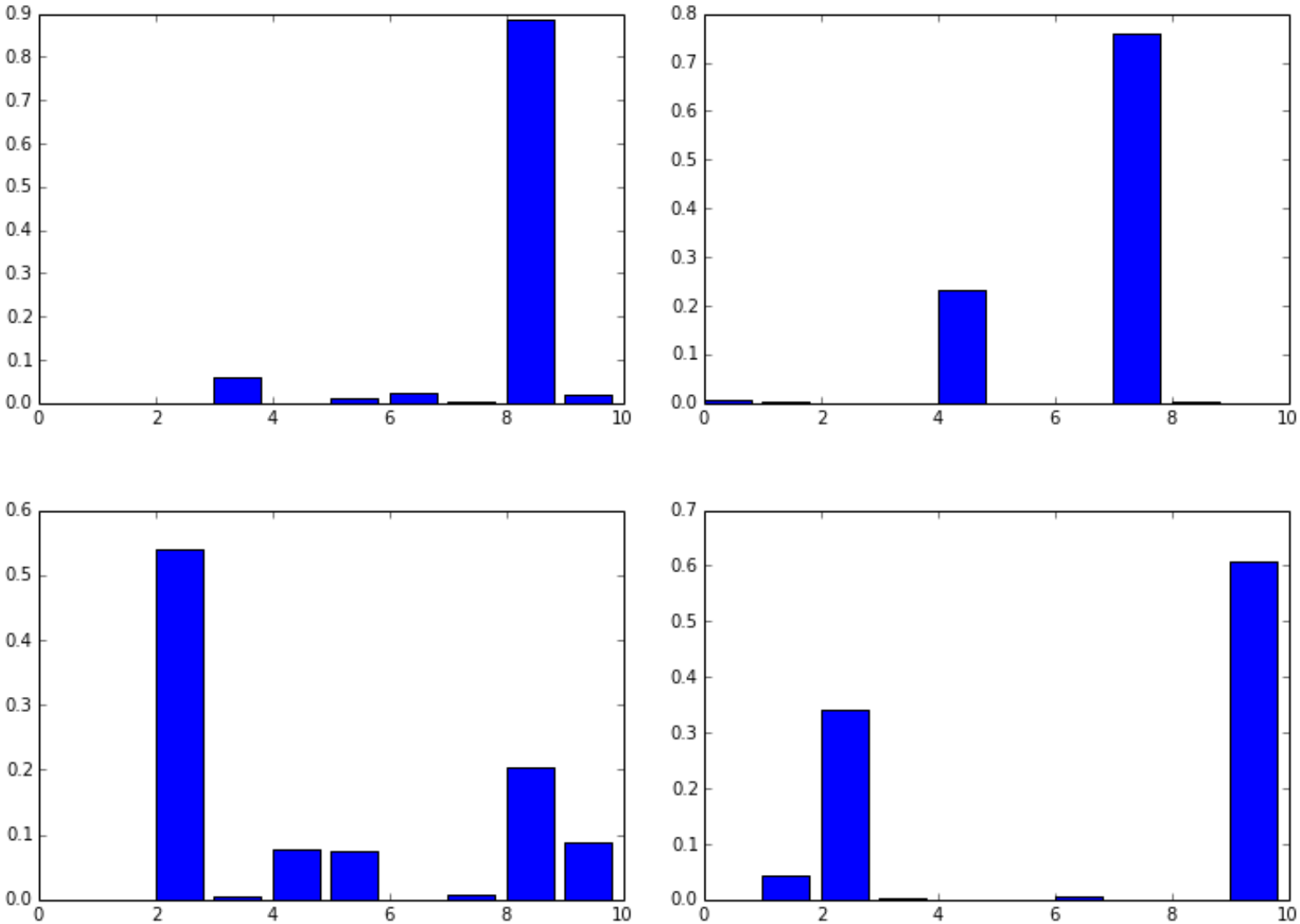




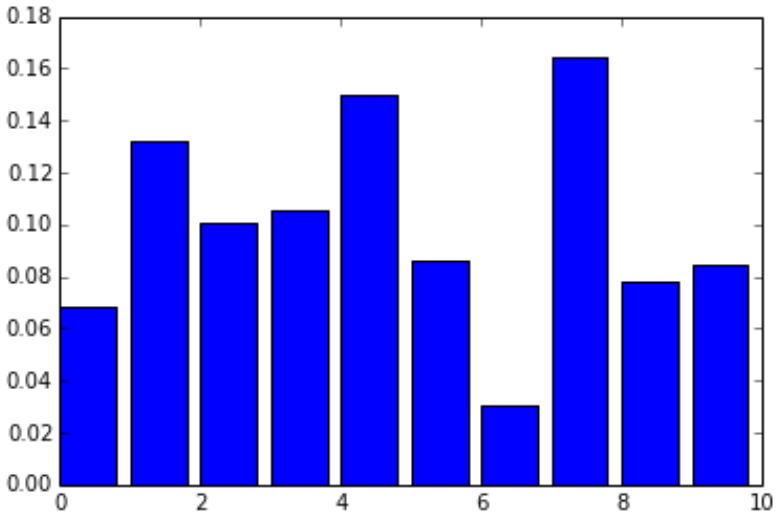


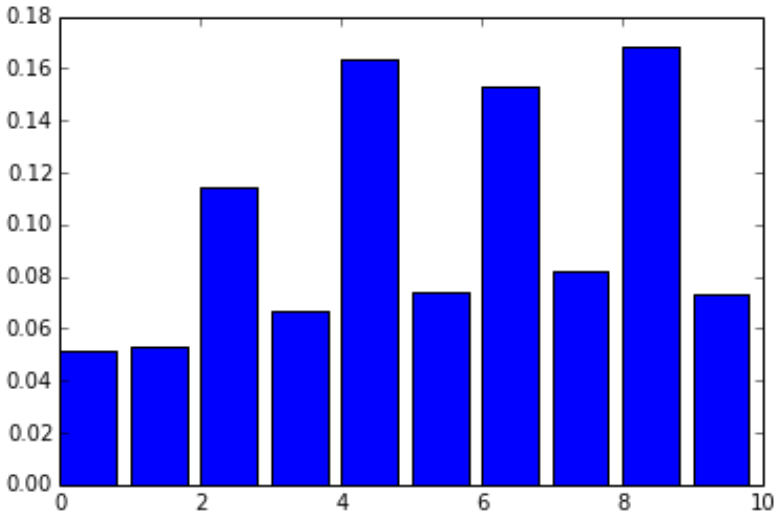


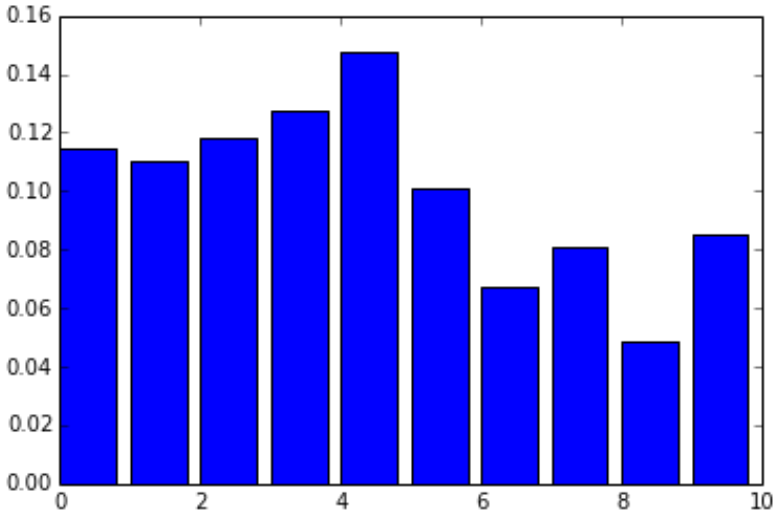
Draws from Dirichlet with $\eta = 0.1$

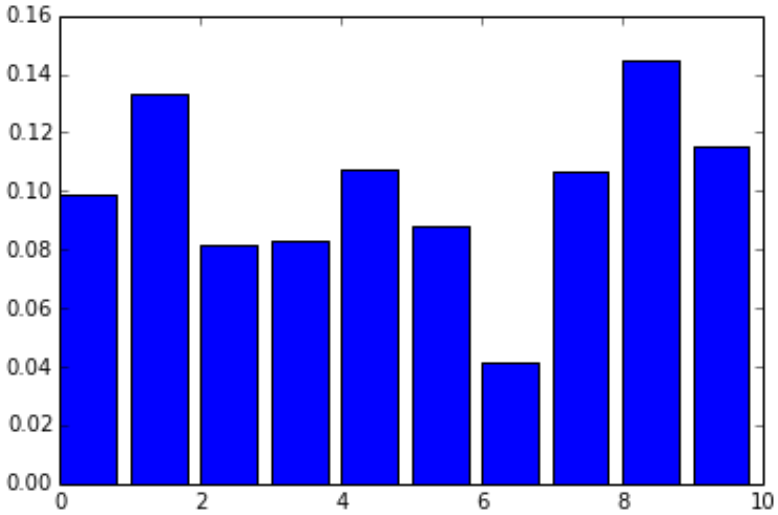


Draws from Dirichlet with $\eta = 10$









Posterior Distribution

$$\Pr[\boldsymbol{\beta} \mid \mathbf{w}] \propto \Pr[\mathbf{w} \mid \boldsymbol{\beta}] \Pr[\boldsymbol{\beta}] \propto \prod_{v=1}^{V} \beta_v^{x_v} \prod_{v=1}^{V} \beta_v^{\eta_v - 1} = \prod_{v=1}^{V} \beta_v^{x_v + \eta_v - 1}.$$

Posterior is a Dirichlet with parameters $(\eta'_1, \ldots, \eta'_V)$ where $\eta'_v \equiv \eta_v + x_v$.

Add term counts to the prior distribution's parameters to form posterior distribution. The Dirichlet hyperparameters can be viewed as *pseudo-counts*, i.e. observations made before observing $\mathbf{w}$.

Therefore we obtain

$$\mathbb{E}[\beta_v \mid \mathbf{w}] = \frac{\eta_v + x_v}{\sum_v \eta_v + N},$$

which also corresponds to the predictive distribution $\Pr[w_{N+1} = v \mid \mathbf{w}]$.

MAP estimator of β_v is

$$\frac{\eta_v + x_v - 1}{\sum_v (\eta_v + x_v) - 2}.$$

Example

Suppose we begin with a possible vocabulary of size $V = 25$.

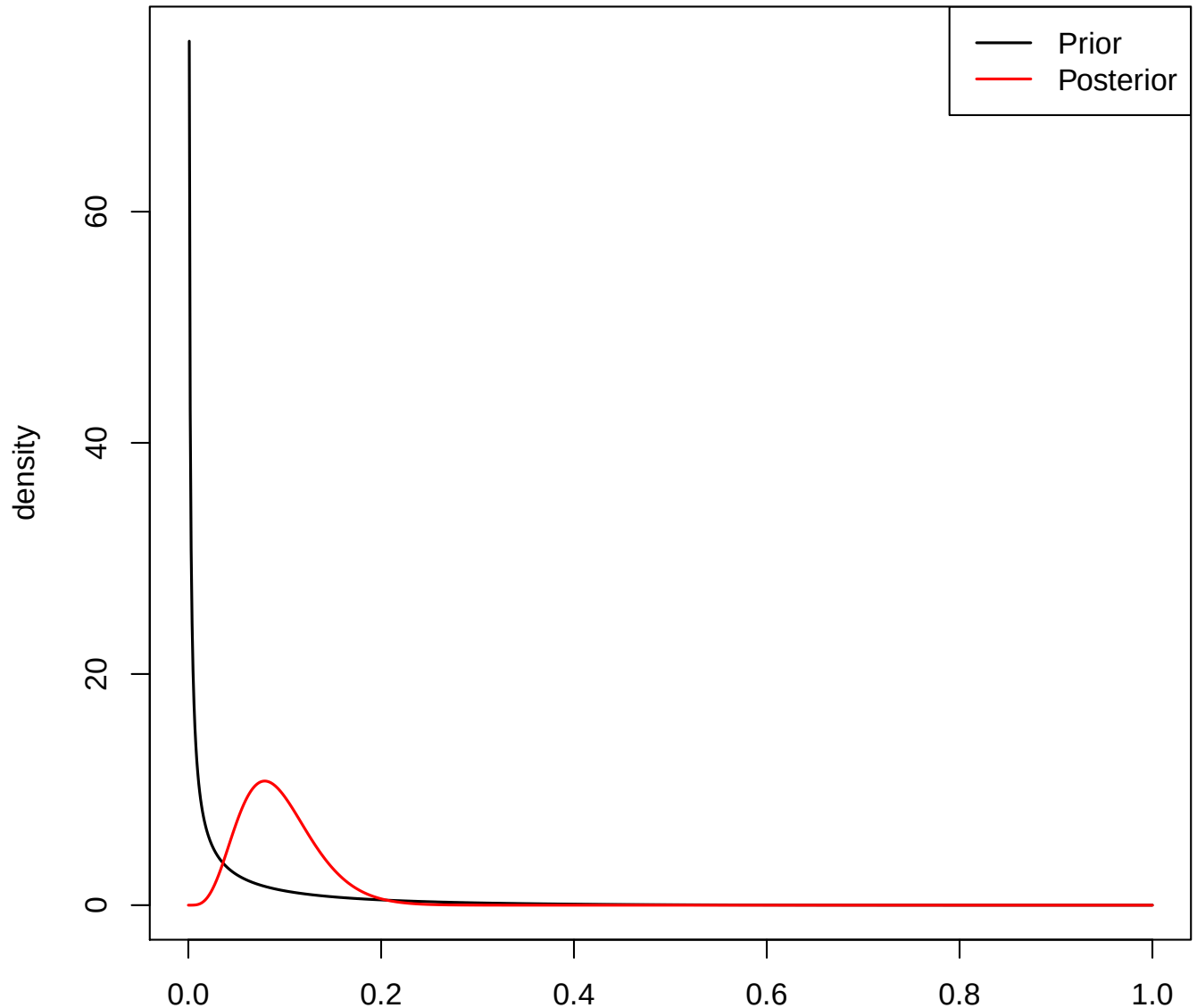
We observe $N = 50$ total words and terms v appears 5 times.

The MLE point estimate of β_v is 0.1.

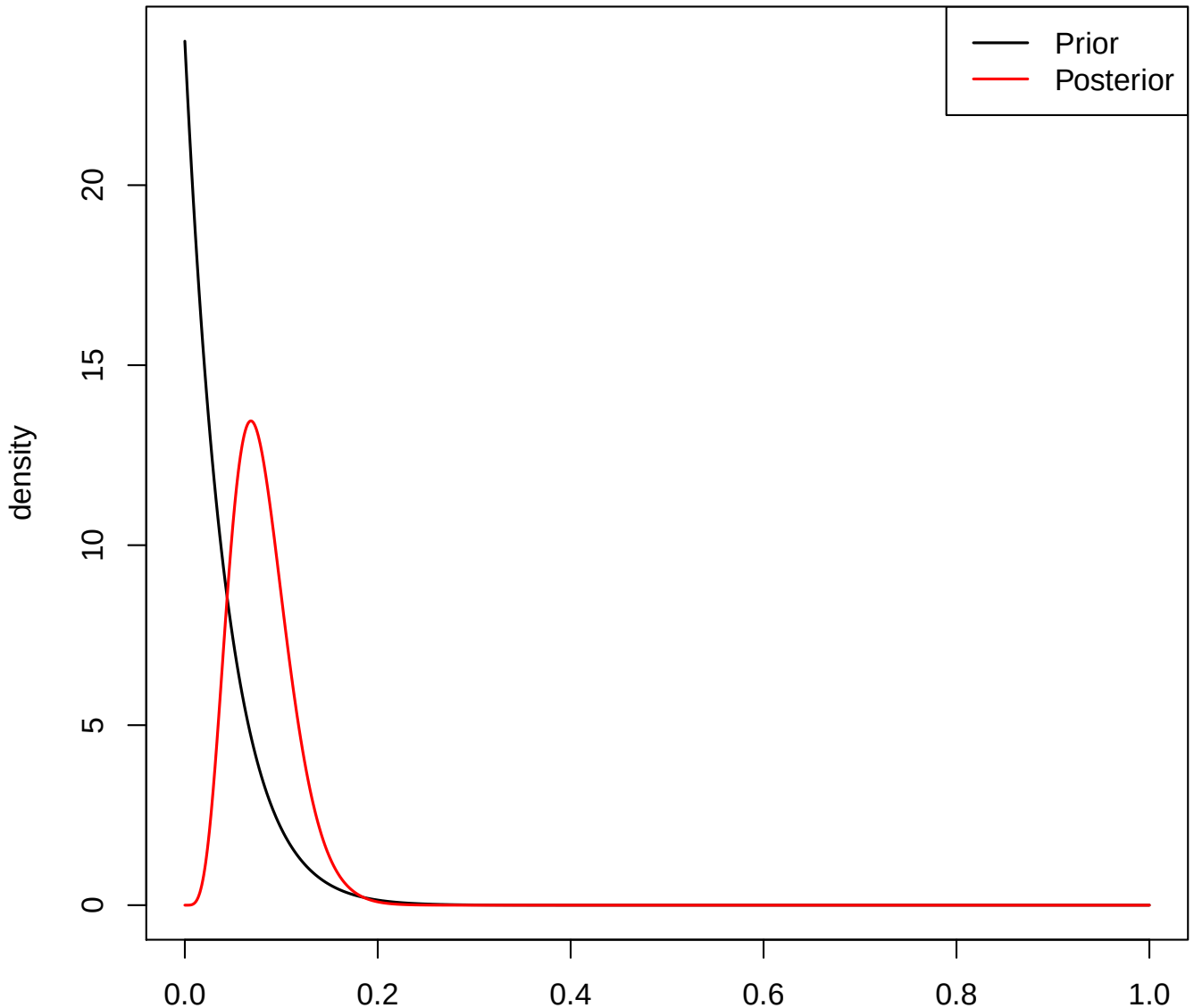
The Bayesian estimate of β_v depends on the prior.

Consider symmetric Dirichlet with hyperparameter η.

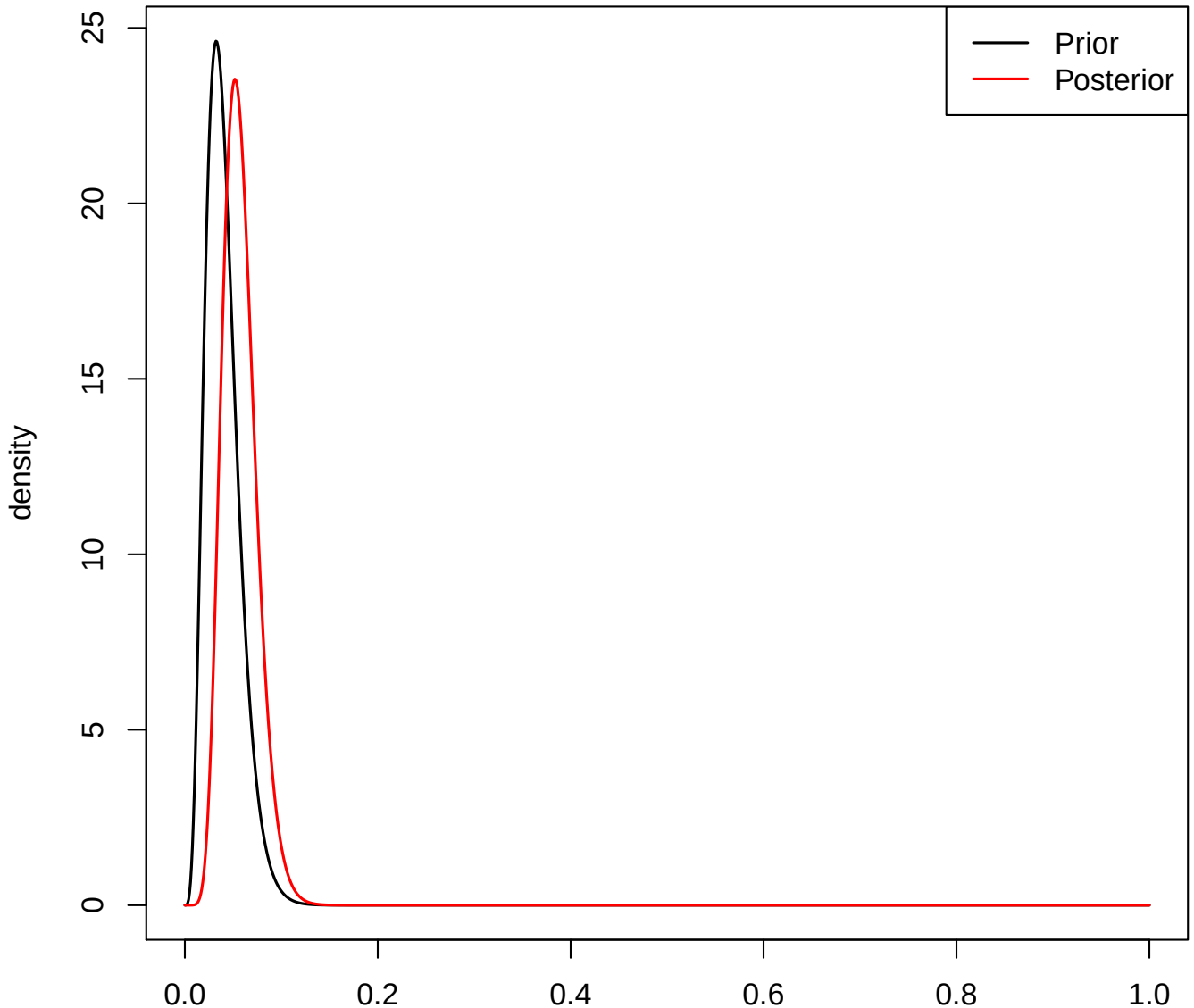
Sparsity-inducing prior ($\eta = 0.2$)



Uniform prior ($\eta = 1$)



Density-inducing prior ($\eta = 5$)



Data Overwhelming the Prior

Recall the MLE estimates for $\widehat{\beta}_v$ satisfies $N\widehat{\beta}_v = x_v$. We then have

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

If we take the limit as $N \to \infty$, 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.

More general result: Bernstein-von Mises theorem.

Conclusion

In MLE we treat parameters as constants, and choose them to maximize the 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.

Outside of special cases, obtaining closed-form solutions for the posterior is impossible; this held back Bayesian methods for decades.

Computation is a large component of Bayesian machine learning (we will see a simple example later).