Nonparametric Bayesian Statistics

Tamara Broderick
ITT Career Development Assistant Professor
Electrical Engineering & Computer Science
MIT
Nonparametric Bayes
Nonparametric Bayes

- Bayesian statistics that is not parametric
Nonparametric Bayes

• Bayesian statistics that is not parametric (wait!)
Nonparametric Bayes

- Bayesian statistics that is not parametric
- Bayesian
Nonparametric Bayes

• Bayesian statistics that is not parametric
• Bayesian

\[ P(\text{parameters}|\text{data}) \propto P(\text{data}|\text{parameters})P(\text{parameters}) \]
Nonparametric Bayes

• Bayesian statistics that is not parametric

• Bayesian

\[ \mathbb{P}(\text{parameters} | \text{data}) \propto \mathbb{P}(\text{data} | \text{parameters}) \mathbb{P}(\text{parameters}) \]

• Not parametric (i.e. not finite parameter, unbounded/growing/infinite number of parameters)
Nonparametric Bayes

• Bayesian statistics that is not parametric
• Bayesian

\[ P(\text{parameters}|\text{data}) \propto P(\text{data}|\text{parameters})P(\text{parameters}) \]
• Not parametric (i.e. not finite parameter, unbounded/growing/infinite number of parameters)

[wikipedia.org]
Nonparametric Bayes

• Bayesian statistics that is not parametric
• Bayesian
  \[ P(\text{parameters}|\text{data}) \propto P(\text{data}|\text{parameters})P(\text{parameters}) \]
• Not parametric (i.e. not finite parameter, unbounded/growing/infinite number of parameters)

“Wikipedia phenomenon”
Nonparametric Bayes

- Bayesian statistics that is not parametric
- Bayesian
  \[ P(\text{parameters}|\text{data}) \propto P(\text{data}|\text{parameters})P(\text{parameters}) \]
- Not parametric (i.e. not finite parameter, unbounded/growing/infinite number of parameters)
Nonparametric Bayes

- Bayesian statistics that is not parametric
- Bayesian
  \[ P(\text{parameters}|\text{data}) \propto P(\text{data}|\text{parameters})P(\text{parameters}) \]
- Not parametric (i.e. not finite parameter, unbounded/growing/infinite number of parameters)

[Ed Bowlby, NOAA]

[wikipedia.org]
Nonparametric Bayes

- Bayesian statistics that is not parametric
- Bayesian

\[ P(\text{parameters}|\text{data}) \propto P(\text{data}|\text{parameters})P(\text{parameters}) \]
- Not parametric (i.e. not finite parameter, unbounded/growing/infinite number of parameters)

[Ed Bowlby, NOAA]

[wikipedia.org]

[Escobar, West 1995; Ghosal et al 1999]
Nonparametric Bayes

• Bayesian statistics that is not parametric
• Bayesian

\[ P(\text{parameters} \mid \text{data}) \propto P(\text{data} \mid \text{parameters})P(\text{parameters}) \]

• Not parametric (i.e. not finite parameter, unbounded/growing/infinite number of parameters)
Nonparametric Bayes

- Bayesian statistics that is not parametric
- Bayesian
  \[ \mathbb{P}(\text{parameters}|\text{data}) \propto \mathbb{P}(\text{data}|\text{parameters}) \mathbb{P}(\text{parameters}) \]
- Not parametric (i.e. not finite parameter, unbounded/growing/infinite number of parameters)

Wikipedia

[Ed Bowlby, NOAA]

[Arjas, Gasbarra 1994]

[Fox et al 2014]

[Escobar, West 1995; Ghosal et al 1999]
Nonparametric Bayes

- Bayesian statistics that is not parametric
- Bayesian
  \[ P(\text{parameters} | \text{data}) \propto P(\text{data} | \text{parameters}) P(\text{parameters}) \]
- Not parametric (i.e. not finite parameter, unbounded/growing/infinite number of parameters)

[Ed Bowlby, NOAA]
[Arjas, Gasbarra 1994]
[Fox et al 2014]

[wikipedia.org]
[Escobar, West 1995; Ghosal et al 1999]
[Arjas, Gasbarra 1994]
[Ewens, 1972; Hartl, Clark 2003]
Nonparametric Bayes

• Bayesian statistics that is not parametric

\[ P(\text{parameters} | \text{data}) \propto P(\text{data} | \text{parameters}) P(\text{parameters}) \]

• Not parametric (i.e. not finite parameter, unbounded/growing/infinite number of parameters)

[Ed Bowlby, NOAA]
[Arjas, Gasbarra 1994]
[Saria et al 2010]

[Escobar, West 1995; Ghosal et al 1999]

[Fox et al 2014]
[Ewens, 1972; Hartl, Clark 2003]
Nonparametric Bayes

- Bayesian statistics that is not parametric
- Bayesian
  \[ P(\text{parameters}|\text{data}) \propto P(\text{data}|\text{parameters})P(\text{parameters}) \]
- Not parametric (i.e. not finite parameter, unbounded/growing/infinite number of parameters)
Nonparametric Bayes

- Bayesian statistics that is not parametric
- Bayesian

\[ P(\text{parameters}|\text{data}) \propto P(\text{data}|\text{parameters})P(\text{parameters}) \]

- Not parametric (i.e. not finite parameter, unbounded/growing/infinite number of parameters)

---

**References**

[Lloyd et al 2012; Miller et al 2010]

[Arias, Gasbarra 1994]

[Escobar, West 1995; Ghosal et al 1999]

[Ed Bowlby, NOAA]

[Saria et al 2010]

[Arjas, Gasbarra 1994]

[Xiaoyu Fox et al 2014]

[Ewens 1972; Hartl, Clark 2003]

[Sudderth, Jordan 2009]
Nonparametric Bayes

• A theoretical motivation: De Finetti’s Theorem
Nonparametric Bayes

• A theoretical motivation: De Finetti’s Theorem
• A data sequence is *infinitely exchangeable* if the distribution of any $N$ data points doesn’t change when permuted: $p(X_1, \ldots, X_N) = p(X_{\sigma(1)}, \ldots, X_{\sigma(N)})$
Nonparametric Bayes

• A theoretical motivation: De Finetti’s Theorem
• A data sequence is *infinitely exchangeable* if the distribution of any $N$ data points doesn’t change when permuted: $p(X_1, \ldots, X_N) = p(X_{\sigma(1)}, \ldots, X_{\sigma(N)})$
• *De Finetti’s Theorem* (roughly): A sequence $X_1, X_2, \ldots$ is infinitely exchangeable if and only if, for all $N$ and some distribution $P$:

[Hewitt, Savage 1955; Aldous 1983]
Nonparametric Bayes

• A theoretical motivation: De Finetti’s Theorem
• A data sequence is \textit{infinitely exchangeable} if the distribution of any $N$ data points doesn’t change when permuted: $p(X_1, \ldots, X_N) = p(X_{\sigma(1)}, \ldots, X_{\sigma(N)})$
• \textit{De Finetti’s Theorem} (roughly): A sequence $X_1, X_2, \ldots$ is infinitely exchangeable if and only if, for all $N$ and some distribution $P$:

\[ p(X_1, \ldots, X_N) = \int_{\theta} \prod_{n=1}^{N} p(X_n|\theta) P(d\theta) \]
Nonparametric Bayes

- A theoretical motivation: De Finetti’s Theorem
- A data sequence is *infinitely exchangeable* if the distribution of any $N$ data points doesn’t change when permuted: $p(X_1, \ldots, X_N) = p(X_{\sigma(1)}, \ldots, X_{\sigma(N)})$
- *De Finetti’s Theorem* (roughly): A sequence $X_1, X_2, \ldots$ is infinitely exchangeable if and only if, for all $N$ and some distribution $P$:
  \[
p(X_1, \ldots, X_N) = \int_{\theta} \prod_{n=1}^{N} p(X_n|\theta) P(d\theta)\]
- Motivates:

[Hewitt, Savage 1955; Aldous 1983]
Nonparametric Bayes

• A theoretical motivation: De Finetti’s Theorem

• A data sequence is \textit{infinitely exchangeable} if the distribution of any \( N \) data points doesn’t change when permuted: \( p(X_1, \ldots, X_N) = p(X_{\sigma(1)}, \ldots, X_{\sigma(N)}) \)

• \textit{De Finetti’s Theorem} (roughly): A sequence \( X_1, X_2, \ldots \) is infinitely exchangeable if and only if, for all \( N \) and some distribution \( P \):

\[
p(X_1, \ldots, X_N) = \int_{\theta} \prod_{n=1}^{N} p(X_n | \theta) P(d\theta)
\]

• Motivates:
  • Parameters and likelihoods

\[\text{[Hewitt, Savage 1955; Aldous 1983]}\]
Nonparametric Bayes

• A theoretical motivation: De Finetti’s Theorem
• A data sequence is \emph{infinitely exchangeable} if the distribution of any \( N \) data points doesn’t change when permuted: \( p(X_1, \ldots, X_N) = p(X_{\sigma(1)}, \ldots, X_{\sigma(N)}) \)

• \emph{De Finetti’s Theorem} (roughly): A sequence \( X_1, X_2, \ldots \) is infinitely exchangeable if and only if, for all \( N \) and some distribution \( P \):

\[
p(X_1, \ldots, X_N) = \int_{\theta} \prod_{n=1}^{N} p(X_n | \theta) P(d\theta)
\]

• Motivates:
  • Parameters and likelihoods
  • Priors

[Hewitt, Savage 1955; Aldous 1983]
Nonparametric Bayes

• A theoretical motivation: De Finetti’s Theorem
• A data sequence is *infinitely exchangeable* if the distribution of any $N$ data points doesn’t change when permuted: $p(X_1, \ldots, X_N) = p(X_{\sigma(1)}, \ldots, X_{\sigma(N)})$

• *De Finetti’s Theorem* (roughly): A sequence $X_1, X_2, \ldots$ is infinitely exchangeable if and only if, for all $N$ and some distribution $P$:

$$p(X_1, \ldots, X_N) = \int_{\theta} \prod_{n=1}^{N} p(X_n | \theta) P(d\theta)$$

• Motivates:
  • Parameters and likelihoods
  • Priors
  • “Nonparametric Bayesian” priors

[Hewitt, Savage 1955; Aldous 1983]
Outline
Outline

• Dirichlet process
Outline

• Dirichlet process
  • Background for intuition
Outline

• Dirichlet process
  • Background for intuition
  • Generative model
Outline

• Dirichlet process
  • Background for intuition
  • Generative model
  • What does a growing/infinite number of parameters really mean (in Nonparametric Bayes)?
Outline

• Dirichlet process
  • Background for intuition
  • Generative model
  • What does a growing/infinite number of parameters really mean (in Nonparametric Bayes)?
• Chinese restaurant process
Outline

• Dirichlet process
  • Background for intuition
  • Generative model
  • What does a growing/infinite number of parameters really mean (in Nonparametric Bayes)?
• Chinese restaurant process
• Inference
Outline

• Dirichlet process
  • Background for intuition
  • Generative model
  • What does a growing/infinite number of parameters really mean (in Nonparametric Bayes)?
• Chinese restaurant process
• Inference
• Venture further into the wild world of Nonparametric Bayesian statistics
Generative model

\[ \begin{align*}
&\text{Don't know } \mu_1, \mu_2 \quad \Rightarrow \quad \\
&\text{Inference goal: assignments of data points to clusters, cluster parameters}.
\end{align*} \]
Generative model

• Finite Gaussian mixture model ($K=2$ clusters)
Generative model

\[ P(\text{parameters}|\text{data}) \propto P(\text{data}|\text{parameters})P(\text{parameters}) \]

- Finite Gaussian mixture model \((K=2\) clusters\)
Generative model

\[ P(\text{parameters} | \text{data}) \propto P(\text{data} | \text{parameters})P(\text{parameters}) \]

- Finite Gaussian mixture model \((K=2\ \text{clusters})\)

\[ z_n \overset{iid}{\sim} \text{Categorical}(\rho_1, \rho_2) \]
Generative model

\[ P(\text{parameters}|\text{data}) \propto P(\text{data}|\text{parameters})P(\text{parameters}) \]

- Finite Gaussian mixture model \((K=2 \text{ clusters})\)
  \[ z_n \xrightarrow{iid} \text{Categorical}(\rho_1, \rho_2) \]
Generative model

\[ \mathbb{P}(\text{parameters} | \text{data}) \propto \mathbb{P}(\text{data} | \text{parameters}) \mathbb{P}(\text{parameters}) \]

- Finite Gaussian mixture model \((K=2 \text{ clusters})\)

\[
\begin{align*}
  z_n & \sim \text{Categorical}(\rho_1, \rho_2) \\
  x_n & \sim \mathcal{N}(\mu_{z_n}, \Sigma)
\end{align*}
\]
Generative model

\[ P(\text{parameters}|\text{data}) \propto P(\text{data}|\text{parameters})P(\text{parameters}) \]

- Finite Gaussian mixture model \((K=2\text{ clusters})\)
  \[ z_n \overset{iid}{\sim} \text{Categorical}(\rho_1, \rho_2) \]
  \[ x_n \overset{indep}{\sim} \mathcal{N}(\mu_{z_n}, \Sigma) \]
- Don’t know \(\mu_1, \mu_2\)
Generative model

\[ P(\text{parameters}|\text{data}) \propto P(\text{data}|\text{parameters})P(\text{parameters}) \]

- Finite Gaussian mixture model \((K=2\text{ clusters})\)
  
  \[ z_n \overset{iid}{\sim} \text{Categorical}(\rho_1, \rho_2) \]
  
  \[ x_n \overset{\text{indep}}{\sim} \mathcal{N}(\mu_{z_n}, \Sigma) \]

- Don’t know \(\mu_1, \mu_2\)

\[ \mu_k \overset{iid}{\sim} \mathcal{N}(\mu_0, \Sigma_0) \]
Generative model

\[ P(\text{parameters}|\text{data}) \propto P(\text{data}|\text{parameters})P(\text{parameters}) \]

- Finite Gaussian mixture model \((K=2\) clusters\)
  \[ z_n^{iid} \sim \text{Categorical}(\rho_1, \rho_2) \]
  \[ x_n^{indep} \sim \mathcal{N}(\mu_{z_n}, \Sigma) \]

- Don’t know \(\mu_1, \mu_2\)
  \[ \mu_k^{iid} \sim \mathcal{N}(\mu_0, \Sigma_0) \]

- Don’t know \(\rho_1, \rho_2\)
Generative model

\[ P(\text{parameters}|\text{data}) \propto P(\text{data}|\text{parameters})P(\text{parameters}) \]

- Finite Gaussian mixture model \((K=2\) clusters\)
  \[ z_n \overset{iid}{\sim} \text{Categorical}(\rho_1, \rho_2) \]
  \[ x_n \overset{\text{indep}}{\sim} \mathcal{N}(\mu_{z_n}, \Sigma) \]

- Don’t know \(\mu_1, \mu_2\)
  \[ \mu_k \overset{iid}{\sim} \mathcal{N}(\mu_0, \Sigma_0) \]

- Don’t know \(\rho_1, \rho_2\)
  \[ \rho_1 \sim \text{Beta}(a_1, a_2) \]
  \[ \rho_2 = 1 - \rho_1 \]
Generative model

\[ P(\text{parameters}|\text{data}) \propto P(\text{data}|\text{parameters})P(\text{parameters}) \]

- Finite Gaussian mixture model \((K=2\) clusters\)
  \[ z_n \overset{iid}{\sim} \text{Categorical}(\rho_1, \rho_2) \]
  \[ x_n \overset{\text{indep}}{\sim} \mathcal{N}(\mu_{z_n}, \Sigma) \]
- Don’t know \(\mu_1, \mu_2\)
  \[ \mu_k \overset{iid}{\sim} \mathcal{N}(\mu_0, \Sigma_0) \]
- Don’t know \(\rho_1, \rho_2\)
  \[ \rho_1 \sim \text{Beta}(a_1, a_2) \]
  \[ \rho_2 = 1 - \rho_1 \]
- Inference goal: assignments of data points to clusters, cluster parameters
Generative model

\[ P(\text{parameters} | \text{data}) \propto P(\text{data} | \text{parameters}) P(\text{parameters}) \]

- Finite Gaussian mixture model \((K=2 \text{ clusters})\)
  \[ z_n \overset{iid}{\sim} \text{Categorical}(\rho_1, \rho_2) \]
  \[ x_n \overset{indep}{\sim} \mathcal{N}(\mu_{z_n}, \Sigma) \]
- Don’t know \(\mu_1, \mu_2\)
  \[ \mu_k \overset{iid}{\sim} \mathcal{N}(\mu_0, \Sigma_0) \]
- Don’t know \(\rho_1, \rho_2\)
  \[ \rho_1 \sim \text{Beta}(a_1, a_2) \]
  \[ \rho_2 = 1 - \rho_1 \]
- Inference goal: assignments of data points to clusters, cluster parameters
Generative model

\[ P(\text{parameters} | \text{data}) \propto P(\text{data} | \text{parameters})P(\text{parameters}) \]

- Finite Gaussian mixture model \((K=2 \text{ clusters})\)
  
  \( z_n \overset{iid}{\sim} \text{Categorical}(\rho_1, \rho_2) \)
  
  \( x_n \overset{indep}{\sim} \mathcal{N}(\mu_{z_n}, \Sigma) \)

- Don’t know \( \mu_1, \mu_2 \)
  
  \( \mu_k \overset{iid}{\sim} \mathcal{N}(\mu_0, \Sigma_0) \)

- Don’t know \( \rho_1, \rho_2 \)
  
  \( \rho_1 \sim \text{Beta}(a_1, a_2) \)
  
  \( \rho_2 = 1 - \rho_1 \)

- Inference goal: assignments of data points to clusters, cluster parameters
Generative model

\[ P(\text{parameters} \mid \text{data}) \propto P(\text{data} \mid \text{parameters})P(\text{parameters}) \]

- Finite Gaussian mixture model \((K=2\) clusters\)
  
  \[ z_n \overset{iid}{\sim} \text{Categorical}(\rho_1, \rho_2) \]
  \[ x_n \overset{\text{indep}}{\sim} \mathcal{N}(\mu_{z_n}, \Sigma) \]

- Don’t know \(\mu_1, \mu_2\)

\[ \mu_k \overset{iid}{\sim} \mathcal{N}(\mu_0, \Sigma_0) \]

- Don’t know \(\rho_1, \rho_2\)

\[ \rho_1 \sim \text{Beta}(a_1, a_2) \]
\[ \rho_2 = 1 - \rho_1 \]

- Inference goal: assignments of data points to clusters, cluster parameters
Generative model

\[ P(\text{parameters}|\text{data}) \propto P(\text{data}|\text{parameters})P(\text{parameters}) \]

- Finite Gaussian mixture model \((K=2 \text{ clusters})\)
  \[ z_n \overset{iid}{\sim} \text{Categorical}(\rho_1, \rho_2) \]
  \[ x_n \overset{\text{indep}}{\sim} \mathcal{N}(\mu_{z_n}, \Sigma) \]
- Don’t know \(\mu_1, \mu_2\)
  \[ \mu_k \overset{iid}{\sim} \mathcal{N}(\mu_0, \Sigma_0) \]
- Don’t know \(\rho_1, \rho_2\)
  \[ \rho_1 \sim \text{Beta}(a_1, a_2) \]
  \[ \rho_2 = 1 - \rho_1 \]
- Inference goal: assignments of data points to clusters, cluster parameters
Beta distribution review

\[ \text{Beta}(\rho_1|a_1, a_2) = \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)} \rho_1^{a_1-1}(1 - \rho_1)^{a_2-1} \]

\[ \rho_1 \in (0, 1) \]

\[ a_1, a_2 > 0 \]
Beta distribution review

$$\text{Beta}(\rho_1|a_1, a_2) = \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)}\rho_1^{a_1-1}(1 - \rho_1)^{a_2-1}$$

- $\rho_1 \in (0, 1)$
- $a_1, a_2 > 0$

- Gamma function $\Gamma$
Beta distribution review

$$\text{Beta}(\rho_1|a_1, a_2) = \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)} \rho_1^{a_1-1}(1 - \rho_1)^{a_2-1}$$

- Gamma function $\Gamma$
- Integer $m$: $\Gamma(m + 1) = m!$

$\rho_1 \in (0, 1)$

$a_1, a_2 > 0$
Beta distribution review

\[
\text{Beta}(\rho_1|a_1, a_2) = \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)} \rho_1^{a_1-1}(1 - \rho_1)^{a_2-1} \quad \rho_1 \in (0, 1) \\
a_1, a_2 > 0
\]

- Gamma function \( \Gamma \)
- integer \( m \): \( \Gamma(m + 1) = m! \)
- for \( x > 0 \): \( \Gamma(x + 1) = x\Gamma(x) \)
Beta distribution review

\[ \text{Beta}(\rho_1|a_1, a_2) = \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)} \rho_1^{a_1-1}(1 - \rho_1)^{a_2-1} \]

\[
\begin{align*}
\rho_1 & \in (0, 1) \\
a_1, a_2 & > 0
\end{align*}
\]

- Gamma function \( \Gamma \)
- Integer \( m \): \( \Gamma(m + 1) = m! \)
- For \( x > 0 \): \( \Gamma(x + 1) = x\Gamma(x) \)

- What happens?
Beta distribution review

\[
\text{Beta}(\rho_1|a_1, a_2) = \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)} \rho_1^{a_1-1}(1 - \rho_1)^{a_2-1}
\]

- Gamma function $\Gamma$
- integer $m$: $\Gamma(m + 1) = m!$
- for $x > 0$: $\Gamma(x + 1) = x\Gamma(x)$

\[
\rho_1 \in (0, 1) \\
a_1, a_2 > 0
\]
Beta distribution review

\[
\text{Beta}(\rho_1|a_1, a_2) = \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)} \rho_1^{a_1-1}(1 - \rho_1)^{a_2-1}
\]

\[
\rho_1 \in (0, 1) \quad a_1, a_2 > 0
\]

- Gamma function \( \Gamma \)
- integer \( m \): \( \Gamma(m + 1) = m! \)
- for \( x > 0 \): \( \Gamma(x + 1) = x\Gamma(x) \)

- What happens?
  - \( a = a_1 = a_2 \to 0 \)
Beta distribution review

\[
\text{Beta}(\rho_1|a_1, a_2) = \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)} \rho_1^{a_1-1}(1 - \rho_1)^{a_2-1}
\]

\(a_1, a_2 > 0\)

- Gamma function \(\Gamma\)
- integer \(m\): \(\Gamma(m + 1) = m!\)
- for \(x > 0\): \(\Gamma(x + 1) = x\Gamma(x)\)

- What happens?
  - \(a = a_1 = a_2 \to 0\)
  - \(a = a_1 = a_2 \to \infty\)
Beta distribution review

\[
\text{Beta}(\rho_1 | a_1, a_2) = \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)} \rho_1^{a_1-1}(1 - \rho_1)^{a_2-1}
\]

\[\rho_1 \in (0, 1) \]
\[a_1, a_2 > 0\]

- Gamma function \( \Gamma \)
- integer \( m \): \( \Gamma(m + 1) = m! \)
- for \( x > 0 \): \( \Gamma(x + 1) = x\Gamma(x) \)

- What happens?
  - \( a = a_1 = a_2 \to 0 \)
  - \( a = a_1 = a_2 \to \infty \)
  - \( a_1 > a_2 \)
Beta distribution review

\[
\text{Beta}(\rho | a_1, a_2) = \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)} \rho^{a_1-1}(1 - \rho)^{a_2-1}
\]

\[\rho \in (0, 1)\]
\[a_1, a_2 > 0\]

- Gamma function \(\Gamma\)
  - integer \(m\): \(\Gamma(m + 1) = m!\)
  - for \(x > 0\): \(\Gamma(x + 1) = x\Gamma(x)\)

- What happens?
  - \(a = a_1 = a_2 \to 0\)
  - \(a = a_1 = a_2 \to \infty\)
  - \(a_1 > a_2\)
Beta distribution review

\[
\text{Beta}(\rho_1|a_1, a_2) = \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)} \rho_1^{a_1-1}(1 - \rho_1)^{a_2-1} \\
\rho_1 \in (0, 1) \quad a_1, a_2 > 0
\]

- Gamma function $\Gamma$
- integer $m$: $\Gamma(m + 1) = m!$
- for $x > 0$: $\Gamma(x + 1) = x\Gamma(x)$

- What happens?
  - $a = a_1 = a_2 \to 0$
  - $a = a_1 = a_2 \to \infty$
  - $a_1 > a_2$ [demo]

- Beta is conjugate to Cat
Beta distribution review

\[
\text{Beta}(\rho_1|a_1, a_2) = \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)} \rho_1^{a_1-1}(1 - \rho_1)^{a_2-1}
\]

\[
\rho_1 \sim \text{Beta}(a_1, a_2),
\]

- Gamma function $\Gamma$
- integer $m$: $\Gamma(m + 1) = m!$
- for $x > 0$: $\Gamma(x + 1) = x\Gamma(x)$

- What happens?
  - $a = a_1 = a_2 \rightarrow 0$
  - $a = a_1 = a_2 \rightarrow \infty$
  - $a_1 > a_2$ [demo]

- Beta is conjugate to Cat

\[
\rho_1 \sim \text{Beta}(a_1, a_2), \ z \sim \text{Cat}(\rho_1, \rho_2)
\]
Beta distribution review

$$\text{Beta}(\rho_1|a_1, a_2) = \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)} \rho_1^{a_1-1}(1 - \rho_1)^{a_2-1}$$

- Gamma function $\Gamma$
- integer $m$: $\Gamma(m + 1) = m!$
- for $x > 0$: $\Gamma(x + 1) = x\Gamma(x)$

What happens?
- $a = a_1 = a_2 \to 0$
- $a = a_1 = a_2 \to \infty$
- $a_1 > a_2$ [demo]

- Beta is conjugate to Cat

$$\rho_1 \sim \text{Beta}(a_1, a_2), z \sim \text{Cat}(\rho_1, \rho_2)$$

$$p(\rho_1, z) \propto$$
Beta distribution review

\[
\text{Beta}(\rho_1|a_1, a_2) = \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)} \rho_1^{a_1-1}(1 - \rho_1)^{a_2-1} \\
\rho_1 \in (0, 1) \\
a_1, a_2 > 0
\]

- **Gamma function** \( \Gamma \)
- **integer** \( m \): \( \Gamma(m + 1) = m! \)
- **for** \( x > 0 \): \( \Gamma(x + 1) = x\Gamma(x) \)

- **What happens?**
  - \( a = a_1 = a_2 \rightarrow 0 \)
  - \( a = a_1 = a_2 \rightarrow \infty \)
  - \( a_1 > a_2 \) [demo]

- **Beta is conjugate to Cat**
  \( \rho_1 \sim \text{Beta}(a_1, a_2), z \sim \text{Cat}(\rho_1, \rho_2) \)

\[
p(\rho_1, z) \propto \rho_1^{1\{z=1\}}(1 - \rho_1)^{1\{z=2\}}.
\]
Beta distribution review

$$\text{Beta}(\rho_1|a_1, a_2) = \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)} \rho_1^{a_1-1} (1 - \rho_1)^{a_2-1}$$

- Gamma function $\Gamma$
- integer $m$: $\Gamma(m + 1) = m!$
- for $x > 0$: $\Gamma(x + 1) = x\Gamma(x)$
- What happens?
  - $a = a_1 = a_2 \rightarrow 0$
  - $a = a_1 = a_2 \rightarrow \infty$
  - $a_1 > a_2$
- Beta is conjugate to Cat

$$\rho_1 \sim \text{Beta}(a_1, a_2), z \sim \text{Cat}(\rho_1, \rho_2)$$

$$p(\rho_1, z) \propto \rho_1^{1\{z=1\}} (1 - \rho_1)^{1\{z=2\}} \rho_1^{a_1-1} (1 - \rho_1)^{a_2-1}$$
Beta distribution review

\[
\text{Beta}(\rho_1|a_1, a_2) = \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)} \rho_1^{a_1-1}(1 - \rho_1)^{a_2-1} \quad \rho_1 \in (0, 1) \\
a_1, a_2 > 0
\]

- Gamma function \( \Gamma \)
- integer \( m \): \( \Gamma(m + 1) = m! \)
- for \( x > 0 \): \( \Gamma(x + 1) = x\Gamma(x) \)
- What happens?
  - \( a = a_1 = a_2 \to 0 \)
  - \( a = a_1 = a_2 \to \infty \)
  - \( a_1 > a_2 \)
- Beta is conjugate to Cat

\[
p(\rho_1, z) \propto \rho_1^{1\{z=1\}}(1 - \rho_1)^{1\{z=2\}} \rho_1^{a_1-1}(1 - \rho_1)^{a_2-1} \\
p(\rho_1|z) \propto \rho_1^{1\{z=1\}}(1 - \rho_1)^{1\{z=2\}} \rho_1^{a_1-1}(1 - \rho_1)^{a_2-1}
\]
Beta distribution review

\[ \text{Beta}(\rho_1|a_1, a_2) = \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)} \rho_1^{a_1-1}(1 - \rho_1)^{a_2-1} \]

\[ \rho_1 \in (0, 1) \quad a_1, a_2 > 0 \]

- Gamma function \( \Gamma \)
- integer \( m \): \( \Gamma(m + 1) = m! \)
- for \( x > 0 \): \( \Gamma(x + 1) = x\Gamma(x) \)

- What happens?
  - \( a = a_1 = a_2 \to 0 \)
  - \( a = a_1 = a_2 \to \infty \)
  - \( a_1 > a_2 \) [demo]

- Beta is conjugate to Cat

\[ \rho_1 \sim \text{Beta}(a_1, a_2), \ z \sim \text{Cat}(\rho_1, \rho_2) \]

\[ p(\rho_1, z) \propto \rho_1^{1\{z=1\}}(1 - \rho_1)^{1\{z=2\}} \rho_1^{a_1-1}(1 - \rho_1)^{a_2-1} \]

\[ p(\rho_1|z) \propto \rho_1^{a_1+1\{z=1\}-1}(1 - \rho_1)^{a_2+1\{z=2\}-1} \]
Beta distribution review

\[
\text{Beta}(\rho_1|a_1, a_2) = \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)} \rho_1^{a_1-1}(1 - \rho_1)^{a_2-1}
\]

\[\rho_1 \in (0, 1), \quad a_1, a_2 > 0\]

- Gamma function \(\Gamma\)
- integer \(m\): \(\Gamma(m + 1) = m!\)
- for \(x > 0\): \(\Gamma(x + 1) = x\Gamma(x)\)
- What happens?
  - \(a = a_1 = a_2 \to 0\)
  - \(a = a_1 = a_2 \to \infty\)
  - \(a_1 > a_2\) [demo]
- Beta is conjugate to Cat
  \(\rho_1 \sim \text{Beta}(a_1, a_2), z \sim \text{Cat}(\rho_1, \rho_2)\)

\[
p(\rho_1, z) \propto \rho_1^{1\{z=1\}}(1 - \rho_1)^{1\{z=2\}} \rho_1^{a_1-1}(1 - \rho_1)^{a_2-1}
\]

\[
p(\rho_1|z) \propto \rho_1^{a_1+1\{z=1\}-1}(1 - \rho_1)^{a_2+1\{z=2\}-1} \propto \text{Beta}(\rho_1|a_1 + 1\{z = 1\}, a_2 + 1\{z = 2\})
\]
Generative model

\[ P(\text{parameters} | \text{data}) \propto P(\text{data} | \text{parameters})P(\text{parameters}) \]

- Finite Gaussian mixture model \((K\text{ clusters})\)
Generative model

\[ P(\text{parameters} | \text{data}) \propto P(\text{data} | \text{parameters}) P(\text{parameters}) \]

- Finite Gaussian mixture model \((K\text{ clusters})\)
Generative model

\[ P(\text{parameters} \mid \text{data}) \propto P(\text{data} \mid \text{parameters})P(\text{parameters}) \]

- Finite Gaussian mixture model \((K\ \text{clusters})\)

\[ \rho_{1:K} \sim \text{Dirichlet}(a_{1:K}) \]
Generative model

\[ P(\text{parameters}|\text{data}) \propto P(\text{data}|\text{parameters})P(\text{parameters}) \]

- Finite Gaussian mixture model \((K\) clusters\)
  
  \[ \rho_{1:K} \sim \text{Dirichlet}(a_{1:K}) \]
  
  \[ \mu_k \overset{iid}{\sim} \mathcal{N}(\mu_0, \Sigma_0) \]
Generative model

\[ \mathbb{P}(\text{parameters}|\text{data}) \propto \mathbb{P}(\text{data}|\text{parameters})\mathbb{P}(\text{parameters}) \]

- Finite Gaussian mixture model \((K\text{ clusters})\)

\[
\begin{align*}
\rho_{1:K} & \sim \text{Dirichlet}(a_{1:K}) \\
\mu_k & \sim \mathcal{N}(\mu_0, \Sigma_0) \\
z_n & \sim \text{Categorical}(\rho_{1:K})
\end{align*}
\]
Generative model

\[ P(\text{parameters}|\text{data}) \propto P(\text{data}|\text{parameters})P(\text{parameters}) \]

- Finite Gaussian mixture model \((K\,\text{clusters})\)

\[ \rho_{1:K} \sim \text{Dirichlet}(a_{1:K}) \]

\[ \mu_k \overset{iid}{\sim} \mathcal{N}(\mu_0, \Sigma_0) \]

\[ z_n \overset{iid}{\sim} \text{Categorical}(\rho_{1:K}) \]

\[ x_n \overset{\text{indep}}{\sim} \mathcal{N}(\mu_{z_n}, \Sigma) \]
Dirichlet distribution review

\[ \text{Dirichlet}(\rho_1:K|a_1:K) = \frac{\Gamma(\sum_{k=1}^{K} a_k)}{\prod_{k=1}^{K} \Gamma(a_k)} \prod_{k=1}^{K} \rho_k^{a_k-1} \quad a_k > 0 \]
Dirichlet distribution review

\[
\text{Dirichlet}(\rho_{1:K} | a_{1:K}) = \frac{\Gamma(\sum_{k=1}^{K} a_k)}{\prod_{k=1}^{K} \Gamma(a_k)} \prod_{k=1}^{K} \rho_k^{a_k-1}
\]

\[a_k > 0\]
\[\rho_k \in (0, 1)\]
\[\sum_k \rho_k = 1\]
Dirichlet distribution review

\[
\text{Dirichlet}(\rho_1:K|a_1:K) = \frac{\Gamma\left(\sum_{k=1}^{K} a_k\right)}{\prod_{k=1}^{K} \Gamma(a_k)} \prod_{k=1}^{K} \rho_k^{a_k-1}
\]

\[a_k > 0\]
Dirichlet distribution review

\[
\text{Dirichlet}(\rho_{1:K} \mid a_{1:K}) = \frac{\Gamma\left(\sum_{k=1}^{K} a_k\right)}{\prod_{k=1}^{K} \Gamma(a_k)} \prod_{k=1}^{K} \rho_k^{a_k-1} \quad a_k > 0
\]

- What happens?
Dirichlet distribution review

\[
\text{Dirichlet}(\rho_{1:K} | a_{1:K}) = \frac{\Gamma\left(\sum_{k=1}^{K} a_k\right)}{\prod_{k=1}^{K} \Gamma(a_k)} \prod_{k=1}^{K} \rho_k^{a_k-1}
\]

\( a > 0 \)

\( a = (0.5,0.5,0.5) \)

\( a = (5,5,5) \)

\( a = (40,10,10) \)

• What happens?
Dirichlet distribution review

Dirichlet\(\rho_{1:K} | a_{1:K}\) = \(\frac{\Gamma\left(\sum_{k=1}^{K} a_k\right)}{\prod_{k=1}^{K} \Gamma(a_k)} \prod_{k=1}^{K} \rho_k^{a_k-1}\)
\[a_k > 0\]

\(a = (0.5, 0.5, 0.5)\)
\(a = (5, 5, 5)\)
\(a = (40, 10, 10)\)

- What happens? \(a = a_k = 1\)
Dirichlet distribution review

$$\text{Dirichlet}(\rho_1:K | a_1:K) = \frac{\Gamma\left(\sum_{k=1}^{K} a_k\right)}{\prod_{k=1}^{K} \Gamma(a_k)} \prod_{k=1}^{K} \rho_k^{a_k-1}$$

$$a_k > 0$$

- What happens?  
  $$a = a_k = 1 \quad a = a_k \to 0$$
Dirichlet distribution review

Dirichlet\((\rho_1:K | a_1:K)\) = \frac{\Gamma\left(\sum_{k=1}^{K} a_k\right)}{\prod_{k=1}^{K} \Gamma(a_k)} \prod_{k=1}^{K} \rho_k^{a_k-1} \quad a_k > 0

a = (0.5, 0.5, 0.5) \quad a = (5, 5, 5) \quad a = (40, 10, 10)

- What happens? \quad a = a_k = 1 \quad a = a_k \to 0 \quad a = a_k \to \infty
Dirichlet distribution review

\[
\text{Dirichlet}(\rho_1:K | a_1:K) = \frac{\Gamma\left(\sum_{k=1}^{K} a_k\right)}{\prod_{k=1}^{K} \Gamma(a_k)} \prod_{k=1}^{K} \rho_k^{a_k-1} \quad a_k > 0
\]

\[a = (0.5, 0.5, 0.5)\]  \[a = (5, 5, 5)\]  \[a = (40, 10, 10)\]

- What happens?
  \[a = a_k = 1\]  \[a = a_k \to 0\]  \[a = a_k \to \infty\]

[demo]
Dirichlet distribution review

\[
\text{Dirichlet}(\rho_{1:K} | a_{1:K}) = \frac{\Gamma\left(\sum_{k=1}^{K} a_k\right)}{\prod_{k=1}^{K} \Gamma(a_k)} \prod_{k=1}^{K} \rho_k^{a_k-1} \quad a_k > 0
\]

\begin{align*}
&\text{a} = (0.5, 0.5, 0.5) \\
&\text{a} = (5, 5, 5) \\
&\text{a} = (40, 10, 10)
\end{align*}

- What happens? \( a = a_k = 1 \) \( a = a_k \to 0 \) \( a = a_k \to \infty \)
- Dirichlet is conjugate to Categorical
Dirichlet distribution review

\[
\text{Dirichlet}(\rho_{1:K} | a_{1:K}) = \frac{\Gamma\left(\sum_{k=1}^{K} a_k\right)}{\prod_{k=1}^{K} \Gamma(a_k)} \prod_{k=1}^{K} \rho_k^{a_k-1} \quad a_k > 0
\]

- What happens?  \( a = a_k = 1 \quad a = a_k \rightarrow 0 \quad a = a_k \rightarrow \infty \)  
- Dirichlet is conjugate to Categorical  
  \( \rho_{1:K} \sim \text{Dirichlet}(a_{1:K}), z \sim \text{Cat}(\rho_{1:K}) \)

\[\begin{align*}
a &= (0.5, 0.5, 0.5) \\
a &= (5, 5, 5) \\
a &= (40, 10, 10)
\end{align*}\]
Dirichlet distribution review

\[
\text{Dirichlet}(\rho_{1:K} | a_{1:K}) = \frac{\Gamma\left(\sum_{k=1}^{K} a_k\right)}{\prod_{k=1}^{K} \Gamma(a_k)} \prod_{k=1}^{K} \rho_k^{a_k-1}, \quad a_k > 0
\]

- What happens? \( a = a_k = 1 \) \( a = a_k \to 0 \) \( a = a_k \to \infty \)
- Dirichlet is conjugate to Categorical
  \( \rho_{1:K} \sim \text{Dirichlet}(a_{1:K}), z \sim \text{Cat}(\rho_{1:K}) \)
  \( \rho_{1:K} | z \stackrel{d}{=} \text{Dirichlet}(a_{1:K}', a_k' = a_k + 1 \{ z = k \}) \)
What if $K > N$?
What if $K > N$?

- Components: number of latent groups
- Clusters: number of components represented in the data
- Number of clusters for $N$ data points is less than $K$ and random
- Number of clusters grows with $N$
What if $K > N$?

- e.g. species sampling, topic modeling, groups on a social network, etc.

![Diagram showing $\rho_1$, $\rho_2$, $\rho_3$, $\rho_{1000}$]
What if $K > N$?

- e.g. species sampling, topic modeling, groups on a social network, etc.

- Components: number of latent groups

- Clusters: number of components represented in the data

- Number of clusters for $N$ data points is $< K$ and random

- Number of clusters grows with $N$

- What if $K > N$?

\[ \rho_1 \quad \rho_2 \quad \rho_3 \quad \rho_{1000} \]

- Components: number of latent groups
What if $K > N$?

- e.g. species sampling, topic modeling, groups on a social network, etc.

<table>
<thead>
<tr>
<th>$\rho_1$</th>
<th>$\rho_2$</th>
<th>$\rho_3$</th>
<th>$\rho_{1000}$</th>
</tr>
</thead>
</table>

- Components: number of latent groups
- Clusters: number of components represented in the data
What if $K > N$?

- e.g. species sampling, topic modeling, groups on a social network, etc.

- Components: number of latent groups

- Clusters: number of components represented in the data

- $[\text{demo 1, demo 2}]$
What if $K > N$?

- e.g. species sampling, topic modeling, groups on a social network, etc.

- Components: number of latent groups

- Clusters: number of components represented in the data

- [demo 1, demo 2]

- Number of clusters for $N$ data points is $< K$ and random
What if $K > N$?

- e.g. species sampling, topic modeling, groups on a social network, etc.

- Components: number of latent groups

- Clusters: number of components represented in the data

- [demo 1, demo 2]

- Number of clusters for $N$ data points is $< K$ and random

- Number of clusters grows with $N$
• Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data

• How to generate $K = \infty$ strictly positive frequencies that sum to one?

Observation: $\sim 1: K \sim \text{Dirichlet}(a_1:K)$
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data
- How to generate $K = \infty$ strictly positive frequencies that sum to one?
Choosing \( K = \infty \)

- Here, difficult to choose finite \( K \) in advance (contrast with small \( K \)): don’t know \( K \), difficult to infer, streaming data

- How to generate \( K = \infty \) strictly positive frequencies that sum to one?
- Observation: \( \rho_{1:K} \sim \text{Dirichlet}(a_{1:K}) \)
Choosing $K = \infty$

• Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data

• How to generate $K = \infty$ strictly positive frequencies that sum to one?
  • Observation: $\rho_{1:K} \sim \text{Dirichlet}(a_{1:K})$
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data
- How to generate $K = \infty$ strictly positive frequencies that sum to one?
- Observation: $\rho_{1:K} \sim \text{Dirichlet}(a_{1:K})$

$\Leftrightarrow \rho_1 \overset{d}{=} \text{Beta}(a_1, \sum_{k=1}^{K} a_k - a_1)$
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data

- How to generate $K = \infty$ strictly positive frequencies that sum to one?
  
  - Observation: $\rho_{1:K} \sim \text{Dirichlet}(a_{1:K})$

\[ \iff \rho_1 \overset{d}{=} \text{Beta}(a_1, \sum_{k=1}^{K} a_k - a_1) \]
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data
- How to generate $K = \infty$ strictly positive frequencies that sum to one?
- Observation: $\rho_{1:K} \sim \text{Dirichlet}(a_{1:K})$

$$\iff \rho_1 \overset{d}{=} \text{Beta}(a_1, \sum_{k=1}^{K} a_k - a_1) \perp \frac{\rho_2, \ldots, \rho_K}{1-\rho_1} \overset{d}{=} \text{Dirichlet}(a_2, \ldots, a_K)$$
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data

- How to generate $K = \infty$ strictly positive frequencies that sum to one?

- Observation: $\rho_{1:K} \sim \text{Dirichlet}(a_{1:K})$

$$\iff \rho_1 \overset{d}{=} \text{Beta}(a_1, \sum_{k=1}^{K} a_k - a_1) \perp \frac{(\rho_2, \ldots, \rho_K)}{1-\rho_1} \overset{d}{=} \text{Dirichlet}(a_2, \ldots, a_K)$$
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data

- How to generate $K = \infty$ strictly positive frequencies that sum to one?
  - Observation: $\rho_{1:K} \sim \text{Dirichlet}(a_{1:K})$

\[
\Leftrightarrow \rho_1 \overset{d}{=} \text{Beta}(a_1, \sum_{k=1}^{K} a_k - a_1) \perp \frac{(\rho_2, \ldots, \rho_K)}{1-\rho_1} \overset{d}{=} \text{Dirichlet}(a_2, \ldots, a_K)
\]
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data

- How to generate $K = \infty$ strictly positive frequencies that sum to one?
  - Observation: $\rho_{1:K} \sim \text{Dirichlet}(a_{1:K})$

\[\Leftrightarrow \rho_1 \overset{d}{=} \text{Beta}(a_1, \sum_{k=1}^{K} a_k - a_1) \perp \frac{(\rho_2, \ldots, \rho_K)}{1-\rho_1} \overset{d}{=} \text{Dirichlet}(a_2, \ldots, a_K)\]

- “Stick breaking”
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data

- How to generate $K = \infty$ strictly positive frequencies that sum to one?
  - Observation: $\rho_{1:K} \sim \text{Dirichlet}(a_{1:K})$

\[
\Leftrightarrow \rho_1 \overset{d}{=} \text{Beta}(a_1, \sum_{k=1}^{K} a_k - a_1) \parallel \frac{(\rho_2, \ldots, \rho_K)}{1 - \rho_1} \overset{d}{=} \text{Dirichlet}(a_2, \ldots, a_K)
\]

- “Stick breaking”

\[V_1 \sim \text{Beta}(a_1, a_2 + a_3 + a_4)\]
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data

- How to generate $K = \infty$ strictly positive frequencies that sum to one?
  - Observation: $\rho_1:K \sim \text{Dirichlet}(a_1:K)$

\[
\Leftrightarrow \rho_1 \overset{d}{=} \text{Beta}(a_1, \sum_{k=1}^{K} a_k - a_1) \perp \frac{\rho_2, \ldots, \rho_K}{1-\rho_1} \overset{d}{=} \text{Dirichlet}(a_2, \ldots, a_K)
\]

- “Stick breaking”

$V_1 \sim \text{Beta}(a_1, a_2 + a_3 + a_4)$ \quad $\rho_1 = V_1$
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data

- How to generate $K = \infty$ strictly positive frequencies that sum to one?
  - Observation: $\rho_{1:K} \sim \text{Dirichlet}(a_{1:K})$

\[
\iff \rho_1 \overset{d}{=} \text{Beta}(a_1, \sum_{k=1}^{K} a_k - a_1) \perp \frac{(\rho_2, \ldots, \rho_K)}{1-\rho_1} \overset{d}{=} \text{Dirichlet}(a_2, \ldots, a_K)
\]

  - “Stick breaking”

\[
V_1 \sim \text{Beta}(a_1, a_2 + a_3 + a_4) \quad \rho_1 = V_1 \\
V_2 \sim \text{Beta}(a_2, a_3 + a_4)
\]
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data
- How to generate $K = \infty$ strictly positive frequencies that sum to one?
  - Observation: $\rho_{1:K} \sim \text{Dirichlet}(a_{1:K})$

\[
\Leftrightarrow \rho_1 \overset{d}{=} \text{Beta}(a_1, \sum_{k=1}^{K} a_k - a_1) \perp \frac{(\rho_2, \ldots, \rho_K)}{1-\rho_1} \overset{d}{=} \text{Dirichlet}(a_2, \ldots, a_K)
\]

- “Stick breaking”

\[
V_1 \sim \text{Beta}(a_1, a_2 + a_3 + a_4) \quad \rho_1 = V_1
\]
\[
V_2 \sim \text{Beta}(a_2, a_3 + a_4) \quad \rho_2 = (1 - V_1)V_2
\]
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data

- How to generate $K = \infty$ strictly positive frequencies that sum to one?
  - Observation: $\rho_{1:K} \sim \text{Dirichlet}(a_{1:K})$

\[ \Leftrightarrow \rho_1 \overset{d}{=} \text{Beta}(a_1, \sum_{k=1}^{K} a_k - a_1) \perp (\rho_2, \ldots, \rho_K) \overset{d}{=} \text{Dirichlet}(a_2, \ldots, a_K) \]

- “Stick breaking”

\[ V_1 \sim \text{Beta}(a_1, a_2 + a_3 + a_4) \quad \rho_1 = V_1 \]
\[ V_2 \sim \text{Beta}(a_2, a_3 + a_4) \quad \rho_2 = (1 - V_1)V_2 \]
\[ V_3 \sim \text{Beta}(a_3, a_4) \]
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data
- How to generate $K = \infty$ strictly positive frequencies that sum to one?
- Observation: $\rho_{1:K} \sim \text{Dirichlet}(a_{1:K})$

$$\Leftrightarrow \rho_1 \overset{d}{=} \text{Beta}(a_1, \sum_{k=1}^{K} a_k - a_1) \perp \frac{\rho_2, \ldots, \rho_K}{1-\rho_1} \overset{d}{=} \text{Dirichlet}(a_2, \ldots, a_K)$$

• “Stick breaking”

$$V_1 \sim \text{Beta}(a_1, a_2 + a_3 + a_4) \quad \rho_1 = V_1$$
$$V_2 \sim \text{Beta}(a_2, a_3 + a_4) \quad \rho_2 = (1 - V_1)V_2$$
$$V_3 \sim \text{Beta}(a_3, a_4) \quad \rho_3 = (1 - V_1)(1 - V_2)V_3$$
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data
- How to generate $K = \infty$ strictly positive frequencies that sum to one?
- Observation: $\rho_{1:K} \sim \text{Dirichlet}(a_{1:K})$

$$\Leftrightarrow \rho_1 \overset{d}{=} \text{Beta}(a_1, \sum_{k=1}^{K} a_k - a_1) \perp \frac{(\rho_2, \ldots, \rho_K)}{1-\rho_1} \overset{d}{=} \text{Dirichlet}(a_2, \ldots, a_K)$$

- “Stick breaking”
  
  $$V_1 \sim \text{Beta}(a_1, a_2 + a_3 + a_4) \quad \rho_1 = V_1$$
  $$V_2 \sim \text{Beta}(a_2, a_3 + a_4) \quad \rho_2 = (1 - V_1)V_2$$
  $$V_3 \sim \text{Beta}(a_3, a_4) \quad \rho_3 = (1 - V_1)(1 - V_2)V_3$$
  $$\rho_4 = 1 - \sum_{k=1}^{3} \rho_k$$
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data

- How to generate $K = \infty$ strictly positive frequencies that sum to one?
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data
- How to generate $K = \infty$ strictly positive frequencies that sum to one?
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data
- How to generate $K = \infty$ strictly positive frequencies that sum to one?
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data
- How to generate $K = \infty$ strictly positive frequencies that sum to one?

\[ V_1 \sim \text{Beta}(a_1, b_1) \]
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data
- How to generate $K = \infty$ strictly positive frequencies that sum to one?

$$V_1 \sim \text{Beta}(a_1, b_1) \quad \rho_1 = V_1$$
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data

- How to generate $K = \infty$ strictly positive frequencies that sum to one?

\[
V_1 \sim \text{Beta}(a_1, b_1) \quad \rho_1 = V_1 \\
V_2 \sim \text{Beta}(a_2, b_2)
\]
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data

- How to generate $K = \infty$ strictly positive frequencies that sum to one?

\[
V_1 \sim \text{Beta}(a_1, b_1) \quad \rho_1 = V_1 \\
V_2 \sim \text{Beta}(a_2, b_2) \quad \rho_2 = (1 - V_1)V_2
\]
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data

- How to generate $K = \infty$ strictly positive frequencies that sum to one?

\[
V_1 \sim \text{Beta}(a_1, b_1) \quad \rho_1 = V_1 \\
V_2 \sim \text{Beta}(a_2, b_2) \quad \rho_2 = (1 - V_1)V_2
\]
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data

- How to generate $K = \infty$ strictly positive frequencies that sum to one?

\[ V_1 \sim \text{Beta}(a_1, b_1) \quad \rho_1 = V_1 \]
\[ V_2 \sim \text{Beta}(a_2, b_2) \quad \rho_2 = (1 - V_1)V_2 \]
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data
- How to generate $K = \infty$ strictly positive frequencies that sum to one?

\[ V_1 \sim \text{Beta}(a_1, b_1) \quad \rho_1 = V_1 \]
\[ V_2 \sim \text{Beta}(a_2, b_2) \quad \rho_2 = (1 - V_1)V_2 \]
\[ \ldots \quad V_k \sim \text{Beta}(a_k, b_k) \]
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data
- How to generate $K = \infty$ strictly positive frequencies that sum to one?

\[
V_1 \sim \text{Beta}(a_1, b_1) \quad \rho_1 = V_1 \\
V_2 \sim \text{Beta}(a_2, b_2) \quad \rho_2 = (1 - V_1)V_2 \\
\vdots \\
V_k \sim \text{Beta}(a_k, b_k) \quad \rho_k = \left[ \prod_{j=1}^{k-1} (1 - V_j) \right] V_k
\]
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data

- How to generate $K = \infty$ strictly positive frequencies that sum to one?

\[
\begin{align*}
V_1 & \sim \text{Beta}(a_1, b_1) & \rho_1 &= V_1 \\
V_2 & \sim \text{Beta}(a_2, b_2) & \rho_2 &= (1 - V_1)V_2 \\
& \vdots \\
V_k & \sim \text{Beta}(a_k, b_k) & \rho_k &= \left[ \prod_{j=1}^{k-1} (1 - V_j) \right] V_k
\end{align*}
\]

[Ishwaran, James 2001]
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data

- How to generate $K = \infty$ strictly positive frequencies that sum to one?

- **Dirichlet process stick-breaking:** $a_k = 1, b_k = \alpha > 0$

\[
\begin{align*}
V_1 &\sim \text{Beta}(a_1, b_1) & \rho_1 = V_1 \\
V_2 &\sim \text{Beta}(a_2, b_2) & \rho_2 = (1 - V_1)V_2 \\
\vdots & & \\
V_k &\sim \text{Beta}(a_k, b_k) & \rho_k = \left[ \prod_{j=1}^{k-1} (1 - V_j) \right] V_k
\end{align*}
\]

[Ishwaran, James 2001]
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data
- How to generate $K = \infty$ strictly positive frequencies that sum to one?
  - **Dirichlet process stick-breaking**: $a_k = 1, b_k = \alpha > 0$
  - Griffiths-Engen-McCloskey (GEM) distribution:
    \[
    \rho = (\rho_1, \rho_2, \ldots) \sim \text{GEM}(\alpha) \]
    \[
    V_1 \sim \text{Beta}(a_1, b_1) \quad \rho_1 = V_1 \\
    V_2 \sim \text{Beta}(a_2, b_2) \quad \rho_2 = (1 - V_1)V_2 \\
    \cdots \\
    V_k \sim \text{Beta}(a_k, b_k) \quad \rho_k = \left[ \prod_{j=1}^{k-1} (1 - V_j) \right] V_k
    \]

[McCloskey 1965; Engen 1975; Patil and Taillie 1977; Ewens 1987; Sethuraman 1994; Ishwaran, James 2001]
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data
- How to generate $K = \infty$ strictly positive frequencies that sum to one?
  - **Dirichlet process stick-breaking**: $a_k = 1$, $b_k = \alpha > 0$
  - Griffiths-Engen-McCloskey (GEM) distribution:
    \[
    \theta = (\theta_1, \theta_2, \ldots) \sim \text{GEM}(\alpha)
    \]
    \[
    V_k \sim \text{Beta}(1, \alpha) \quad \rho_k = \left[ \prod_{j=1}^{k-1} (1 - V_j) \right] V_k
    \]

[McCloskey 1965; Engen 1975; Patil and Taillie 1977; Ewens 1987; Sethuraman 1994; Ishwaran, James 2001]
Choosing $K = \infty$

- Here, difficult to choose finite $K$ in advance (contrast with small $K$): don’t know $K$, difficult to infer, streaming data

- How to generate $K = \infty$ strictly positive frequencies that sum to one?

  - **Dirichlet process stick-breaking**: $a_k = 1, b_k = \alpha > 0$
  
  - Griffiths-Engen-McCloskey (GEM) distribution:
  
  $$
  \rho = (\rho_1, \rho_2, \ldots) \sim \text{GEM}(\alpha)
  $$

\[
V_k \sim \text{Beta}(1, \alpha) \quad \rho_k = \left[ \prod_{j=1}^{k-1} (1 - V_j) \right] V_k
\]

[McCloskey 1965; Engen 1975; Patil and Taillie 1977; Ewens 1987; Sethuraman 1994; Ishwaran, James 2001]
Distributions
Distributions

• Beta → random distribution over 1, 2
Distributions

• Beta $\rightarrow$ random distribution over $1, 2$

• Dirichlet $\rightarrow$ random distribution over $1, 2, \ldots, K$
Distributions

- Beta → random distribution over 1, 2
- Dirichlet → random distribution over 1, 2, ..., $K$
- GEM / Dirichlet process stick-breaking → random distribution over 1, 2, ...
Distributions

• Beta $\rightarrow$ random distribution over 1, 2

• Dirichlet $\rightarrow$ random distribution over 1, 2, \ldots, $K$

• GEM / Dirichlet process stick-breaking $\rightarrow$ random distribution over 1, 2, \ldots

• Infinity of parameters: components

• Growing number of parameters: clusters
Distributions

• Beta → random distribution over 1, 2

• Dirichlet → random distribution over 1, 2, ..., $K$

• GEM / Dirichlet process
  stick-breaking → random distribution over 1, 2, ...
Distributions

• Beta $\rightarrow$ random distribution over 1, 2

• Dirichlet $\rightarrow$ random distribution over 1, 2, $\ldots$, $K$

• GEM / Dirichlet process stick-breaking $\rightarrow$ random distribution over 1, 2, $\ldots$

\[ \rho = (\rho_1, \rho_2, \ldots) \sim \text{GEM}(\alpha) \]
Distributions

• Beta $\rightarrow$ random distribution over 1, 2

• Dirichlet $\rightarrow$ random distribution over 1, 2, ..., $K$

• GEM / Dirichlet process stick-breaking $\rightarrow$ random distribution over 1, 2, ...

$$\rho = (\rho_1, \rho_2, \ldots) \sim \text{GEM}(\alpha)$$

$$\phi_k \overset{iid}{\sim} G_0$$
Distributions

• Beta \rightarrow \text{random distribution over 1, 2}

• Dirichlet \rightarrow \text{random distribution over 1, 2, \ldots, } K

• GEM / Dirichlet process stick-breaking \rightarrow \text{random distribution over 1, 2, \ldots}

\[ \rho = (\rho_1, \rho_2, \ldots) \sim \text{GEM(} \alpha \text{)} \]

\[ \phi_k \overset{iid}{\sim} G_0 \]

\[ G = \sum_{k=1}^{\infty} \rho_k \delta_{\phi_k} \]
Distributions

• Beta $\rightarrow$ random distribution over $1, 2$

• Dirichlet $\rightarrow$ random distribution over $1, 2, \ldots, K$

• GEM / Dirichlet process stick-breaking $\rightarrow$ random distribution over $1, 2, \ldots$

\[
\rho = (\rho_1, \rho_2, \ldots) \sim \text{GEM}(\alpha)
\]
\[
\phi_k \overset{iid}{\sim} G_0
\]
\[
G = \sum_{k=1}^{\infty} \rho_k \delta_{\phi_k}
\]
Distributions

• Beta → random distribution over 1, 2

• Dirichlet → random distribution over 1, 2, ..., K

• GEM / Dirichlet process stick-breaking → random distribution over 1, 2, ...

• **Dirichlet process** → random distribution over Φ:

  \[ \rho = (\rho_1, \rho_2, \ldots) \sim \text{GEM}(\alpha) \]

  \[ \phi_k \overset{iid}{\sim} G_0 \]

  \[ G = \sum_{k=1}^{\infty} \rho_k \delta_{\phi_k} \]

  \[ \Phi \]

[Ferguson 1973]
Dirichlet process mixture model
Dirichlet process mixture model

• Gaussian mixture model
Dirichlet process mixture model

• Gaussian mixture model

\[ \rho = (\rho_1, \rho_2, \ldots) \sim \text{GEM}(\alpha) \]
Dirichlet process mixture model

• Gaussian mixture model

\[
\rho = (\rho_1, \rho_2, \ldots) \sim \text{GEM}(\alpha)
\]
Dirichlet process mixture model

- Gaussian mixture model

\[ \rho = (\rho_1, \rho_2, \ldots) \sim \text{GEM}(\alpha) \]
\[ \mu_k \sim \mathcal{N}(\mu_0, \Sigma_0), \ k = 1, 2, \ldots \]
Dirichlet process mixture model

- Gaussian mixture model

\[ \rho = (\rho_1, \rho_2, \ldots) \sim \text{GEM}(\alpha) \]

\[ \mu_k \overset{iid}{\sim} \mathcal{N}(\mu_0, \Sigma_0), \; k = 1, 2, \ldots \]
Dirichlet process mixture model

• Gaussian mixture model

\[ \rho = (\rho_1, \rho_2, \ldots) \sim \text{GEM}(\alpha) \]

\[ \mu_k \overset{iid}{\sim} \mathcal{N}(\mu_0, \Sigma_0), \; k = 1, 2, \ldots \]

• i.e. \( G = \sum_{k=1}^{\infty} \rho_k \delta_{\mu_k} \)
Dirichlet process mixture model

• Gaussian mixture model

\[ \rho = (\rho_1, \rho_2, \ldots) \sim \text{GEM}(\alpha) \]

\[ \mu_k \overset{iid}{\sim} \mathcal{N}(\mu_0, \Sigma_0), \, k = 1, 2, \ldots \]

• i.e. \[ G = \sum_{k=1}^{\infty} \rho_k \delta_{\mu_k} \overset{d}{=} \text{DP}(\alpha, \mathcal{N}(\mu_0, \Sigma_0)) \]
Dirichlet process mixture model

- Gaussian mixture model

\[ \rho = (\rho_1, \rho_2, \ldots) \sim \text{GEM}(\alpha) \]

\[ \mu_k \overset{iid}{\sim} \mathcal{N}(\mu_0, \Sigma_0), \ k = 1, 2, \ldots \]

- i.e. \[ G = \sum_{k=1}^{\infty} \rho_k \delta_{\mu_k} \overset{d}{=} \text{DP}(\alpha, \mathcal{N}(\mu_0, \Sigma_0)) \]

\[ z_n \overset{iid}{\sim} \text{Categorical}(\rho) \]
Dirichlet process mixture model

• Gaussian mixture model

\[ \rho = (\rho_1, \rho_2, \ldots) \sim \text{GEM}(\alpha) \]
\[ \mu_k \overset{iid}{\sim} \mathcal{N}(\mu_0, \Sigma_0), k = 1, 2, \ldots \]

i.e. \[ G = \sum_{k=1}^{\infty} \rho_k \delta_{\mu_k} \overset{d}{=} \text{DP}(\alpha, \mathcal{N}(\mu_0, \Sigma_0)) \]

\[ z_n \overset{iid}{\sim} \text{Categorical}(\rho) \]
\[ \mu^*_n = \mu_{z_n} \]
Dirichlet process mixture model

- Gaussian mixture model
  \[ \rho = (\rho_1, \rho_2, \ldots) \sim \text{GEM}(\alpha) \]
  \[ \mu_k \overset{iid}{\sim} \mathcal{N}(\mu_0, \Sigma_0), \; k = 1, 2, \ldots \]
- i.e. \[ G = \sum_{k=1}^{\infty} \rho_k \delta_{\mu_k} \overset{d}{=} \text{DP}(\alpha, \mathcal{N}(\mu_0, \Sigma_0)) \]

- \[ z_n \overset{iid}{\sim} \text{Categorical}(\rho) \]
- \[ \mu^*_n = \mu_{z_n} \]
Dirichlet process mixture model

- Gaussian mixture model
  \[
  \rho = (\rho_1, \rho_2, \ldots) \sim \text{GEM}(\alpha)
  \]
  \[
  \mu_k \overset{iid}{\sim} \mathcal{N}(\mu_0, \Sigma_0), \ k = 1, 2, \ldots
  \]
- i.e. \( G = \sum_{k=1}^{\infty} \rho_k \delta_{\mu_k} \overset{d}{=} \text{DP}(\alpha, \mathcal{N}(\mu_0, \Sigma_0)) \)

- \( z_n \overset{iid}{\sim} \text{Categorical}(\rho) \)
- \( \mu^*_n = \mu_{z_n} \)
- i.e. \( \mu^*_n \overset{iid}{\sim} G \)
**Dirichlet process mixture model**

- **Gaussian mixture model**

\[
\rho = (\rho_1, \rho_2, \ldots) \sim \text{GEM}(\alpha) \\
\mu_k \overset{iid}{\sim} \mathcal{N}(\mu_0, \Sigma_0), \ k = 1, 2, \ldots \\
\text{i.e. } G = \sum_{k=1}^{\infty} \rho_k \delta_{\mu_k} \overset{d}{=} \text{DP}(\alpha, \mathcal{N}(\mu_0, \Sigma_0))
\]

- \(z_n \overset{iid}{\sim} \text{Categorical}(\rho)\)

\(\mu_n^* = \mu_{z_n}\)

- \(\text{i.e. } \mu_n^* \overset{iid}{\sim} G\)

\(x_n \overset{indep}{\sim} \mathcal{N}(\mu_n^*, \Sigma)\)
Dirichlet process mixture model

- Gaussian mixture model
  \[
  \rho = (\rho_1, \rho_2, \ldots) \sim \text{GEM}(\alpha)
  \]
  \[
  \mu_k \overset{iid}{\sim} \mathcal{N}(\mu_0, \Sigma_0), \quad k = 1, 2, \ldots
  \]
- i.e. \( G = \sum_{k=1}^{\infty} \rho_k \delta_{\mu_k} \overset{d}{=} \text{DP}(\alpha, \mathcal{N}(\mu_0, \Sigma_0)) \)

\[
\begin{align*}
  z_n & \overset{iid}{\sim} \text{Categorical}(\rho) \\
  \mu_n^* & = \mu_{z_n} \\
  \mu_n^* & \overset{iid}{\sim} G
\end{align*}
\]

- i.e. \( \mu_n^* \sim G \)

\[
\begin{align*}
x_n & \overset{\text{indep}}{\sim} \mathcal{N}(\mu_n^*, \Sigma)
\end{align*}
\]
Exercises

- Prove the Dirichlet is conjugate to the categorical
- What is the posterior after $N$ data points?
- Suppose $\rho \sim \text{Beta}(a_1, K)$; prove that $\rho \sim \text{Dirichlet}(a_1:K)$
- Code your own GEM simulator for $\rho$; why is this hard?
- Simulate drawing cluster indicators ($z$) from your $ho \sim \text{Dirichlet}(a_1:K)$

$\rho \overset{\sim}{\leftarrow} \text{Beta}(a_1, K) \quad z_k \overset{\sim}{\leftarrow} \text{Dirichlet}(a_1, ..., a_K)$

- Compare the number of clusters as $N$ changes in the GEM case with the growth in the $K=1000$ case
- How do the two compare when you change $\alpha$ ?

1, 2, 3, 4, ...
Exercises

- Prove the Dirichlet is conjugate to the categorical

- What is the posterior after $N$ data points?

- Suppose $\rho \sim \text{Beta}(a, K)$; prove that $\lambda = 1$, $K \sim \text{Dirichlet}(a_1, \ldots, a_K)$.

- Code your own GEM simulator for $\rho$; why is this hard?

- Simulate drawing cluster indicators ($z$) from your $\rho$; compare the number of clusters as $N$ changes in the GEM case with the growth in the $K=1000$ case when you change $a$.

[slides, code: www.tamarabroderick.com/tutorials.html]
Exercises

- Prove the beta (Dirichlet) is conjugate to the categorical
- What is the posterior after \( N \) data points?
- Suppose \( \rho \mapsto 1: K \mapsto \text{Dirichlet}(\alpha_1: K) \mapsto \text{Beta}(\alpha_1, \sum_{k=1}^X a_k) \)?\(^2\)
- Compare the number of clusters as \( N \) changes in the GEM case with the growth in the \( K = 1000 \) case
- How do the two compare when you change \( \alpha \)?

[slides, code: www.tamarabroderick.com/tutorials.html]
• Prove the beta (Dirichlet) is conjugate to the categorical
• What is the posterior after $N$ data points?
Exercises

- Prove the beta (Dirichlet) is conjugate to the categorical
- What is the posterior after $N$ data points?
- Suppose $\rho_{1:K} \sim \text{Dirichlet}(a_{1:K})$; prove equivalence to
  
  $\rho_1 \overset{d}{=} \text{Beta}(a_1, \sum_{k=1}^{K} a_k - a_1) \parallel \frac{(\rho_2, \ldots, \rho_K)}{1-\rho_1} \overset{d}{=} \text{Dirichlet}(a_2, \ldots, a_K)$

[slides, code: www.tamarabroderick.com/tutorials.html]
Exercises

• Prove the beta (Dirichlet) is conjugate to the categorical
• What is the posterior after $N$ data points?
• Suppose $\rho_{1:K} \sim \text{Dirichlet}(a_{1:K})$; prove equivalence to

$$\rho_1 \overset{d}{=} \text{Beta}(a_1, \sum_{k=1}^{K} a_k - a_1) \parallel \frac{(\rho_2, \ldots, \rho_K)}{1-\rho_1} \overset{d}{=} \text{Dirichlet}(a_2, \ldots, a_K)$$

• Code your own GEM simulator for $\rho$; why is this hard?
Exercises

• Prove the beta (Dirichlet) is conjugate to the categorical
  • What is the posterior after $N$ data points?
• Suppose $\rho_1:K \sim \text{Dirichlet}(a_1:K)$; prove equivalence to
  $$\rho_1 \overset{d}{=} \text{Beta}(a_1, \sum_{k=1}^{K} a_k - a_1) \perp \frac{(\rho_2, \ldots, \rho_K)}{1-\rho_1} \overset{d}{=} \text{Dirichlet}(a_2, \ldots, a_K)$$
• Code your own GEM simulator for $\rho$; why is this hard?
• Simulate drawing cluster indicators ($z$) from your $\rho$
Exercises

- Prove the beta (Dirichlet) is conjugate to the categorical
- What is the posterior after $N$ data points?
- Suppose $\rho_{1:K} \sim \text{Dirichlet}(a_{1:K})$; prove equivalence to
  \[\rho_1 \overset{d}{=} \text{Beta}(a_1, \sum_{k=1}^{K} a_k - a_1) \parallel \frac{(\rho_2, \ldots, \rho_K)}{1-\rho_1} \overset{d}{=} \text{Dirichlet}(a_2, \ldots, a_K)\]
- Code your own GEM simulator for $\rho$; why is this hard?
- Simulate drawing cluster indicators ($z$) from your $\rho$

- Compare the number of clusters as $N$ changes in the GEM case with the growth in the $K=1000$ case
Exercises

• Prove the beta (Dirichlet) is conjugate to the categorical
  • What is the posterior after $N$ data points?
• Suppose $\rho_1:K \sim \text{Dirichlet}(a_1:K)$; prove equivalence to
  \[ \rho_1 \overset{d}{=} \text{Beta}(a_1, \sum_{k=1}^{K} a_k - a_1) \perp \frac{(\rho_2, \ldots, \rho_K)}{1-\rho_1} \overset{d}{=} \text{Dirichlet}(a_2, \ldots, a_K) \]
• Code your own GEM simulator for $\rho$; why is this hard?
• Simulate drawing cluster indicators ($z$) from your $\rho$
  • Compare the number of clusters as $N$ changes in the GEM case with the growth in the $K=1000$ case
• How does the growth in $N$ change when you change $\alpha$?
Exercises

• Prove the beta (Dirichlet) is conjugate to the categorical
  • What is the posterior after $N$ data points?
• Suppose $\rho_{1:K} \sim \text{Dirichlet}(a_{1:K})$; prove equivalence to
  $\rho_1 \overset{d}{=} \text{Beta}(a_1, \sum_{k=1}^{K} a_k - a_1) \perp \frac{(\rho_2, \ldots, \rho_K)}{1-\rho_1} \overset{d}{=} \text{Dirichlet}(a_2, \ldots, a_K)$
• Code your own GEM simulator for $\rho$; why is this hard?
• Simulate drawing cluster indicators ($z$) from your $\rho$
  • Compare the number of clusters as $N$ changes in the GEM case with the growth in the $K=1000$ case
• How does the growth in $N$ change when you change $\alpha$?
• How does the distribution of # clusters at $N$ change with $\alpha$?
References
A full reference list is provided at the end of the “Part III” slides.