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

[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"

[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]
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]

[www.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

$$P(\text{parameters} | \text{data}) \propto P(\text{data} | \text{parameters}) P(\text{parameters})$$

- 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]

[Arjas, Gasbarra 1994]

[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)

[Wikipedia](https://en.wikipedia.org)

[Ed Bowlby, NOAA]

[Escobar, West 1995; Ghosal et al 1999]

[Arjas, Gasbarra 1994]

[Fox et al 2014]

[Ewens, 1972; Hartl, Clark 2003]
Nonparametric Bayes

• Bayesian statistics that is not parametric

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

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

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

[Ed Bowlby, NOAA]

[Arjas, Gasbarra 1994]

[Saria et al 2010]

[Fox et al 2014]

[Lloyd et al 2012; Miller et al 2010]

[Ewens, 1972; Hartl, Clark 2003]

[Escobar, West 1995; Ghosal et al 1999]

[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)
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: $p(X_1, \ldots, X_N) = p(X_{\sigma(1)}, \ldots, X_{\sigma(N)})$

\[ p(X_1, \ldots, X_N) = p(X_{\sigma(1)}, \ldots, X_{\sigma(N)}) \]
Nonparametric Bayes

• A theoretical motivation: De Finetti’s Theorem
• A data sequence is \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 $P$:

[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 $P$:

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

[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:

[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

[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

[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

- Don't know $\mu_1, \mu_2$
- Don't know $\gamma_1, \gamma_2$

$z_n \overset{iid}{\sim} \text{Categorical} (\gamma_1, \gamma_2)$

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

$\gamma_1 \overset{\sim}{\sim} \text{Beta} (a_1, a_2)$

$\gamma_2 = 1 - \gamma_1$
Generative model

- Finite Gaussian mixture model ($K=2$ clusters)

\[ z_n \text{iid} \sim \text{Categorical}(\pi_1, \pi_2) \]

\[ \pi_1 \sim \text{Beta}(a_1, a_2) \]

\[ \pi_2 = 1 - \pi_1 \]

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

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 \((K=2\) clusters\)

\[ z_n \overset{iid}{\sim} \text{Categorical}(\propto_1, \propto_2) \]

\[ \mu_k \overset{iid}{\sim} N(\mu_0, \Theta_0) \]

\[ \propto_1 \overset{}{\sim} \text{Beta}(a_1, a_2) \]

\[ \propto_2 = 1 - \propto_1 \]
Generative model

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

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

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

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

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

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

\[ x_n \overset{indep}{\sim} \mathcal{N}(\mu_{z_n}, \Sigma) \]
Generative model

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

- Finite Gaussian mixture model \((K=2\) 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 \((K=2 \text{ clusters})\)
  \[ z_n \overset{iid}{\sim} \text{Categorical}(\rho_1, \rho_2) \]
  \[ x_n \overset{indep}{\sim} \mathcal{N}(\mu_{z_n}, \Sigma) \]
- Don’t know \(\mu_1, \mu_2\)
  \[ \mu_k \overset{iid}{\sim} \mathcal{N}(\mu_0, \Sigma_0) \]
- Don’t know \(\rho_1, \rho_2\)
Generative model

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

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

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

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

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

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

- 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

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

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

- 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
Generative model

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

- Finite Gaussian mixture model \((K=2\ \text{clusters})\)
  \[ z_n \overset{\text{iid}}{\sim} \text{Categorical}(\rho_1, \rho_2) \]
  \[ x_n \overset{\text{indep}}{\sim} \mathcal{N}(\mu_{z_n}, \Sigma) \]
- Don’t know \(\mu_1, \mu_2\)
  \[ \mu_k \overset{\text{iid}}{\sim} \mathcal{N}(\mu_0, \Sigma_0) \]
- 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

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

- Gamma function \( \Gamma \)

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

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

- Gamma function \( \Gamma \)
- integer \( m \): \( \Gamma(m) = (m - 1)! \) 

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

- What happens?
  - \( a = a_1 = a_2 \rightarrow 0 \)
  - \( a = a_1 = a_2 \rightarrow \infty \)
  - \( a_1 > a_2 \)

[demo]
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 \( m \): \( \Gamma(m) = (m - 1)! \)
- for \( x > 0 \): \( \Gamma(x) = x\Gamma(x - 1) \)
- 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

[demo]
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 $m$: $\Gamma(m) = (m-1)!$
- for $x > 0$: $\Gamma(x) = x\Gamma(x-1)$

- 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

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

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

where \( \rho_1 \in (0, 1) \)

\[ a_1, a_2 > 0 \]

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

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

\[ p(\rho_1, z) \propto \]

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

\( a_1, a_2 > 0 \)
Beta distribution review

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

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

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

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

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

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

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

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

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

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

\[ p(\rho_1|z) \propto \]
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 \( m \): \( \Gamma(m) = (m-1)! \)
- for \( x > 0 \): \( \Gamma(x) = x\Gamma(x-1) \)
- 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

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

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

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

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

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

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

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

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

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

- Finite Gaussian mixture model (\( K \) clusters)

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

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

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

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

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

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

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

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

\[
\begin{align*}
\rho_{1:K} & \sim \text{Dirichlet}(a_{1:K}) \\
\mu_k & \sim \mathcal{N}(\mu_0, \Sigma_0) \\
z_n & \sim \text{Categorical}(\rho_{1:K}) \\
x_n & \sim \text{indep} \mathcal{N}(\mu_{z_n}, \Sigma)
\end{align*}
\]
Dirichlet distribution review

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

\[a_k > 0\]
Dirichlet distribution review

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

\[
\rho_k \in (0, 1) \quad \sum_{k} \rho_k = 1
\]
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

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

- What happens?
Dirichlet distribution review

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

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} \quad a_k > 0

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

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

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

\(a_k > 0\)

- What happens?
  \(a = a_k = 1\)
  \(a = a_k \to 0\)

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

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

- \(a = a_k = 1\)  \(a = a_k \to 0\)  \(a = a_k \to \infty\)
Dirichlet distribution review

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

\[
a_k > 0
\]

- What happens?

\[
a = (0.5,0.5,0.5)
\]

\[
a = (5,5,5)
\]

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

\[
a = a_k = 1 \quad a = a_k \to 0 \quad a = a_k \to \infty
\]

[demo]
Dirichlet distribution review

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

\[a_k > 0\]

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

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

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

\[a_k > 0\]

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

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

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

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

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

\[
\rho_{1:K} | z \overset{d}{=} \text{Dirichlet}(a'_{1:K}), a'_k = a_k + 1 \{z = k\}
\]
What if $K \gg N$?

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

$\implies 1 \implies 2 \implies 3 \ldots \implies 1000$

- 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 $K \gg N$?

- 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 $K \gg N$?

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

![Diagram showing components and clusters with labels $\rho_1$, $\rho_2$, $\rho_3$, ..., $\rho_{1000}$]
What if $K \gg N$?

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

$\implies 1 \implies 2 \implies 3 \ldots \implies 1000$

- Components: number of latent groups

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

- 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
What if $K \gg N$?

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

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

- Components: number of latent groups
- Clusters: number of components represented in the data
- [demo 1, demo 2]
What if $K \gg N$?

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

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

- 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 $K \gg N$?

- 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: $\Gamma_1: K \sim \text{Dirichlet}(a_1:K)$
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: $\rho_{1:K} \sim \text{Dirichlet}(a_{1:K})$

\[
\Rightarrow \rho_1 \overset{d}{=} \text{Beta}(a_1, \sum_{k=1}^{K} a_k - a_1)
\]
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: $\rho_{1:K} \sim \text{Dirichlet}(a_{1:K})$

\[ \Rightarrow \rho_1 \overset{d}{=} \text{Beta}(a_1, \sum_{k=1}^{K} a_k - a_1) \]
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: $\rho_{1:K} \sim \text{Dirichlet}(a_{1:K})$

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

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

\[ \Rightarrow \rho_1 \overset{d}{=} \text{Beta}(a_1, \sum_{k=1}^{K} a_k - a_1) \perp \frac{\rho_2, \ldots, \rho_{K}}{1-\rho_1} \overset{d}{=} \text{Dirichlet}(a_2, \ldots, a_K) \]
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: $\rho_{1:K} \sim \text{Dirichlet}(a_{1:K})$

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

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

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

  - “Stick breaking”

\[
V_1 \sim \text{Beta}(a_1, a_2 + a_3 + a_4)
\]
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: $\rho_{1:K} \sim \text{Dirichlet}(a_{1:K})$

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

• “Stick breaking”

\[
V_1 \sim \text{Beta}(a_1, a_2 + a_3 + a_4) \quad \rho_1 = V_1
\]
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: $\rho_{1:K} \sim \text{Dirichlet}(a_{1:K})$

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

- “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: $\rho_1:K \sim \text{Dirichlet}(a_1:K)$

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

- “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: $\rho_1:K \sim \text{Dirichlet}(a_1:K)$

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

- “Stick breaking”

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

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

- “Stick breaking”

\[
\begin{align*}
V_1 & \sim \text{Beta}(a_1, a_2 + a_3 + a_4) & \rho_1 = V_1 \\
V_2 & \sim \text{Beta}(a_2, a_3 + a_4) & \rho_2 = (1-V_1)V_2 \\
V_3 & \sim \text{Beta}(a_3, a_4) & \rho_3 = (1-V_1)(1-V_2)V_3
\end{align*}
\]
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: $\rho_1:K \sim \text{Dirichlet}(a_1:K)$

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

- “Stick breaking”

  $V_1 \sim \text{Beta}(a_1, a_2 + a_3 + a_4)$ \hspace{1cm} $\rho_1 = V_1$

  $V_2 \sim \text{Beta}(a_2, a_3 + a_4)$ \hspace{1cm} $\rho_2 = (1 - V_1)V_2$

  $V_3 \sim \text{Beta}(a_3, a_4)$ \hspace{1cm} $\rho_3 = (1 - V_1)(1 - V_2)V_3$

  $\rho_4 = 1 - \sum_{k=1}^{3} \rho_k$
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?

\[
\begin{align*}
V_1 & \sim \text{Beta}(a_1, b_1) \quad \rho_1 = V_1 \\
V_2 & \sim \text{Beta}(a_2, b_2) \quad \rho_2 = (1 - V_1)V_2 \\
\vdots & \\
V_k & \sim \text{Beta}(a_k, b_k)
\end{align*}
\]
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?

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

\[
\begin{align*}
V_1 & \sim \text{Beta}(a_1, b_1) & \rho_1 &= V_1 \\
V_2 & \sim \text{Beta}(a_2, b_2) & \rho_2 &= (1 - V_1)V_2 \\
\vdots & \cdots & \rho_k &= \left[\prod_{j=1}^{k-1}(1 - V_j)\right] V_k \\
V_k & \sim \text{Beta}(a_k, b_k) & & \text{[Ishwaran, James 2001]}
\end{align*}
\]
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**: $a_k = 1, b_k = \alpha > 0$

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

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

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

- Code your own GEM simulator to draw $\rho$
- Simulate drawing cluster indicators ($z$) from the distribution you generated in the first exercise
- Compare the growth in the number of clusters as $N$ changes in the GEM case with the growth in the $K=1000$ case
- How does the expected number of clusters in the GEM case change with $N$ and with the GEM parameter $\alpha$?
References for Part 1, page 1


References for Part 1, page 2


S Saria, D Koller, and A Penn. Learning individual and population traits from clinical temporal data. NIPS, 2010.
