Nonparametric Bayesian Methods: Models, Algorithms, and Applications

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

- Bayesian methods that are not parametric
Nonparametric Bayes

• Bayesian methods that are not parametric (wait!)
Nonparametric Bayes

• Bayesian methods that are not parametric
• Bayesian
Nonparametric Bayes

- Bayesian methods that are not parametric
- Bayesian

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

• Bayesian methods that are not parametric
• Bayesian

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

- Bayesian methods that are not parametric
- Bayesian

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

- Bayesian methods that are 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 methods that are 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 methods that are 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 methods that are 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 methods that are 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)

[Ed Bowlby, NOAA]

[wikipedia.org]
Nonparametric Bayes

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

- Bayesian methods that are 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 methods that are 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)

[wiki.org]
[Ed Bowlby, NOAA]
[Fox et al 2014]
[Prabhakaran, Azizi, Carr, Pe’er 2016]
[Lloyd et al 2012; Miller et al 2009]
Nonparametric Bayes

• Bayesian methods that are not parametric

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

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

- Bayesian methods that are 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]

[Prabhakaran, Azizi, Carr, Pe’er 2016]

[Saria et al 2010]

[Lloyd et al 2012; Miller et al 2009]

[Fox et al 2014]

[Ewens 1972; Hartl, Clark 2003; Harris et al 2017]
Nonparametric Bayes

- Bayesian methods that are 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]

[Prabhakaran, Azizi, Carr, Pe’er 2016]

[ESO/L. Calçada/M. Kornmesser 2017]

[Del Pozzo et al 2017, 2018]

[Saria et al 2010]

[Lloyd et al 2012; Miller et al 2009]

[Fox et al 2014]

[Ewens 1972; Hartl, Clark 2003; Harris et al 2017]
Nonparametric Bayes

- Bayesian methods that are 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 methods that are 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

De Finetti's Theorem (roughly): A sequence is infinitely exchangeable if and only if, for all \( N \) and some distribution \( P \):

\[
    \prod_{i=1}^{N} p(X_i) = \prod_{j=1}^{N} p(X_{(j)}),
\]

Motivates:

- Parameters and likelihoods
- Priors

“Nonparametric Bayesian” priors
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 *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 \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:

[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 $P$:
  \[
p(X_1, \ldots, X_N) = \int \prod_{n=1}^{N} p(X_n|\theta) P(d\theta)\]
- 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 \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
  • Priors
  • “Nonparametric Bayesian” priors

[Hewitt, Savage 1955; Aldous 1983]
Roadmap
Roadmap

- Example problem: clustering
Roadmap

- Example problem: clustering
- Example NPBayes model: Dirichlet process
Roadmap

- Example problem: clustering
- Example NPBayes model: Dirichlet process
- Chinese restaurant process
Roadmap

- Example problem: clustering
- Example NPBayes model: Dirichlet process
- Chinese restaurant process
- Inference
Roadmap

- Example problem: clustering
- Example NPBayes model: Dirichlet process
- Chinese restaurant process
- Inference
- Venture further into the wild world of Nonparametric Bayes
Roadmap

- Example problem: clustering
- Example NPBayes model: Dirichlet process
- Chinese restaurant process
- Inference
- Venture further into the wild world of Nonparametric Bayes
- Big questions
Roadmap

• Example problem: clustering
• Example NPBayes model: Dirichlet process
• Chinese restaurant process
• Inference
• Venture further into the wild world of Nonparametric Bayes

• Big questions
  • Why NPBayes?
Roadmap

• Example problem: clustering
• Example NPBayes model: Dirichlet process
• Chinese restaurant process
• Inference
• Venture further into the wild world of Nonparametric Bayes

• Big questions
  • Why NPBayes?
  • What does an infinite/growing number of parameters really mean (in NPBayes)?
Roadmap

- Example problem: clustering
- Example NPBayes model: Dirichlet process
- Chinese restaurant process
- Inference
- Venture further into the wild world of Nonparametric Bayes

- Big questions
  - Why NPBayes?
  - What does an infinite/growing number of parameters really mean (in NPBayes)?
  - Why is NPBayes challenging but practical?
Generative model

- \( \text{Don't know} \mu_1, \mu_2 \)
- \( \text{Don't know} \theta_1, \theta_2 \)
- \( z_n \text{ iid } \sim \text{Categorical}(\theta_1, \theta_2) \)
- \( \mu_k \text{ iid } \sim N(\mu_0, \Sigma_0) \)
- \( \theta_1 \sim \text{Beta}(a_1, a_2) \)
- \( \theta_2 = 1 - \theta_1 \)

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

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

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

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

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

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

- 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 \(\mu_1, \mu_2\)
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 \(\mu_1, \mu_2\)
  \[ \mu_k \overset{iid}{\sim} \mathcal{N}(\mu_0, \Sigma_0) \]
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 \(\rho_1, \rho_2\)
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 \(\rho_1, \rho_2\)
  \[ \rho_1 \sim \text{Beta}(a_1, a_2) \]
  \[ \rho_2 = 1 - \rho_1 \]
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 \((K=2 \text{ clusters})\)
  \[
  z_n \overset{iid}{\sim} \text{Categorical}(\rho_1, \rho_2) \\
  x_n \overset{\text{indep}}{\sim} \mathcal{N}(\mu_{z_n}, \Sigma)
  \]
- 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 (\( K=2 \) 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 \((K=2\ \text{clusters})\)
  \[ z_n \overset{iid}{\sim} \text{Categorical}(\rho_1, \rho_2) \]
  \[ x_n \overset{indep}{\sim} \mathcal{N}(\mu_{z_n}, \Sigma) \]
- 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 \((K=2 \text{ clusters})\)
  \[ z_n \overset{iid}{\sim} \text{Categorical}(\rho_1, \rho_2) \]
  \[ x_n \overset{indep}{\sim} \mathcal{N}(\mu_{z_n}, \Sigma) \]
- 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
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

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

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

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

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

\( a_1, a_2 > 0 \)
\( \rho_1 \in (0, 1) \)
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 \( m \): \( \Gamma(m + 1) = m! \)
- 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 \( x > 0 \): \( \Gamma(x + 1) = x\Gamma(x) \)
- What happens?
  - \( a = a_1 = a_2 \to 0 \)

\( \rho_1 \in (0, 1) \)
\( a_1, a_2 > 0 \)
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 \(\Gamma\)
- integer \(m\): \(\Gamma(m + 1) = m!\)
- for \(x > 0\): \(\Gamma(x + 1) = x\Gamma(x)\)

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

\[\text{[demo]}\]
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 \( m \): \( \Gamma(m + 1) = m! \)
- For \( x > 0 \): \( \Gamma(x + 1) = x \Gamma(x) \)

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

[demo]
Beta distribution review

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

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

$\rho_1 \in (0, 1)$
$a_1, a_2 > 0$
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?
  - \( a = a_1 = a_2 \rightarrow 0 \)
  - \( a = a_1 = a_2 \rightarrow \infty \)
  - \( a_1 > a_2 \) [demo]

- Beta is conjugate to Cat

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

<table>
<thead>
<tr>
<th>density</th>
<th>2.5</th>
<th>2.0</th>
<th>1.5</th>
<th>1.0</th>
<th>0.5</th>
<th>0.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \rho_1 )</td>
<td>0.0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8</td>
<td>1.0</td>
</tr>
</tbody>
</table>
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

$\rho_1 \sim \text{Beta}(a_1, a_2), z \sim \text{Cat}(\rho_1, \rho_2)$
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?
  - \( a = a_1 = a_2 \rightarrow 0 \)
  - \( a = a_1 = a_2 \rightarrow \infty \)
  - \( a_1 > a_2 \)  

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

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

\[ p(\rho_1, z) \propto \rho_1^{1\{z=1\}}(1 - \rho_1)^{1\{z=2\}}. \]
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?
  - $a = a_1 = a_2 \to 0$
  - $a = a_1 = a_2 \to \infty$
  - $a_1 > a_2$
- Beta is conjugate to Cat
  \[ \rho_1 \sim \text{Beta}(a_1, a_2), z \sim \text{Cat}(\rho_1, \rho_2) \]

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

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

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

- 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

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

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

\[ p(\rho_1|z) \propto \]
Beta distribution review

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

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

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

- 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

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

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

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

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

- **Beta is conjugate to Cat**

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

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

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

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

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

\[ \rho_1:K \sim \text{Dirichlet}(a_1:K) \]
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

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

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

\[
\begin{align*}
\rho_1:K & \sim \text{Dirichlet}(a_1:K) \\
\mu_k & \overset{iid}{\sim} \mathcal{N}(\mu_0, \Sigma_0) \\
z_n & \overset{iid}{\sim} \text{Categorical}(\rho_1:K) \\
x_n & \overset{\text{indep}}{\sim} \mathcal{N}(\mu_{z_n}, \Sigma)
\end{align*}
\]
Dirichlet distribution review

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

\(a_k > 0\)
Dirichlet distribution review

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

$a_k > 0$

$\rho_k \in (0, 1)$

$\sum_k \rho_k = 1$
Dirichlet distribution review

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

\[
a_k > 0 \quad \rho_k \in (0, 1) \quad \sum_k \rho_k = 1
\]

• What happens?
Dirichlet distribution review

$$\text{Dirichlet}(\rho_{1:K} | \alpha_{1:K}) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \rho_k^{\alpha_k - 1}$$

- $a = (0.5, 0.5, 0.5)$
- $a = (5, 5, 5)$
- $a = (40, 10, 10)$

- What happens?
Dirichlet distribution review

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

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

\( a = (0.5, 0.5, 0.5) \) \hspace{1cm} \( a = (5, 5, 5) \) \hspace{1cm} \( a = (40, 10, 10) \)
Dirichlet distribution review

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

- What happens?
  - \(a = (0.5, 0.5, 0.5)\)
  - \(a = (5, 5, 5)\)
  - \(a = (40, 10, 10)\)

\(a_k > 0\)
\(\rho_k \in (0, 1)\)
\(\sum_k \rho_k = 1\)

\[\text{demo}\]
Dirichlet distribution review

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

- What happens?  
  \( a = (0.5, 0.5, 0.5) \)  
  \( a = (5, 5, 5) \)  
  \( a = (40, 10, 10) \)

\( a_k > 0 \)  
\( \rho_k \in (0, 1) \)  
\( \sum_{k} \rho_k = 1 \)

**[demo]**
Dirichlet distribution review

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

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

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

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

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

$a = (0.5,0.5,0.5)$  

$a = (5,5,5)$  

$a = (40,10,10)$

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

- Dirichlet is conjugate to Categorical
Dirichlet distribution review

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

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

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

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

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

- 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 $K > N$?
What if $K > N$ ?
What if $K > N$?
What if $K > N$?

$\rho_1 \quad \rho_2 \quad \rho_3 \quad \rho_{1000}$
What if $K > N$?

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

![Diagram showing components and clusters](image)

$\rho_1 \quad \rho_2 \quad \rho_3 \quad \rho_{1000}$
What if $K > N$?

- 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

What if $K > N$?
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

What if $K > N$?
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]
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
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: $\rho_{1:K} \sim \text{Dirichlet}(a_{1:K})$

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

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

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

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

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

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

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

- “Stick breaking”

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

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

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

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

- “Stick breaking”

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

  $$V_2 \sim \text{Beta}(a_2, a_3 + a_4)$$
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”

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

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

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

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

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

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

- “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?

$$
\begin{align*}
V_1 &\sim \text{Beta}(a_1, b_1) & \rho_1 &= V_1 \\
V_2 &\sim \text{Beta}(a_2, b_2) & \rho_2 &= (1 - V_1)V_2
\end{align*}
$$
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?

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

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

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

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

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

[van der Vaart, Ghosal 2017]
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**: $a_k = 1, b_k = \alpha > 0$

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

[van der Vaart, Ghosal 2017]
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; van der Vaart, Ghosal 2017]
Roadmap

• Example problem: clustering
• Example NPBayes model: Dirichlet process
• Chinese restaurant process
• Inference
• Venture further into the wild world of Nonparametric Bayes

• Big questions
  • Why NPBayes?
  • What does an infinite/growing number of parameters really mean (in NPBayes)?
  • Why is NPBayes challenging but practical?
Roadmap

• Example problem: clustering
• Example NPBayes model: Dirichlet process
• Chinese restaurant process
• Inference
• Venture further into the wild world of Nonparametric Bayes

• Big questions
  • Why NPBayes?
  • What does an infinite/growing number of parameters really mean (in NPBayes)?
  • Why is NPBayes challenging but practical?
Roadmap

• Example problem: clustering
• Example NPBayes model: Dirichlet process
• Chinese restaurant process
• Inference
• Venture further into the wild world of Nonparametric Bayes

• Big questions
  • Why NPBayes?
  • What does an infinite/growing number of parameters really mean (in NPBayes)?
  • Why is NPBayes challenging but practical?
Roadmap

- Example problem: clustering
- Example NPBayes model: Dirichlet process
- Chinese restaurant process
- Inference
- Venture further into the wild world of Nonparametric Bayes

- Big questions
  - Why NPBayes? Learn more as acquire more data
  - What does an infinite/growing number of parameters really mean (in NPBayes)?
  - Why is NPBayes challenging but practical?
Roadmap

• Example problem: clustering
• Example NPBayes model: Dirichlet process
• Chinese restaurant process
• Inference
• Venture further into the wild world of Nonparametric Bayes

• Big questions
  • Why NPBayes? Learn more as acquire more data
  • What does an infinite/growing number of parameters really mean (in NPBayes)? Components vs. clusters; latent vs. realized
  • Why is NPBayes challenging but practical?
Roadmap

• Example problem: clustering
• Example NPBayes model: Dirichlet process
• Chinese restaurant process
• Inference
• Venture further into the wild world of Nonparametric Bayes

• Big questions
  • Why NPBayes? Learn more as acquire more data
  • What does an infinite/growing number of parameters really mean (in NPBayes)? Components vs. clusters; latent vs. realized
  • Why is NPBayes challenging but practical? Infinite dimensional parameter; more on this next session!
Exercises

• Prove the beta (Dirichlet) is conjugate to the categorical

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

• How does the number of clusters change as $N$ changes for the Dirichlet model with $K=1000$?

• How does the number of clusters change as the Dirichlet hyperparameter changes for $K=1000$ and $N$ fixed?

• Suppose $\rho_1:K \sim \text{Dirichlet}(a_1:K)$; prove equivalence to

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

A full reference list is provided at the end of the “Part 3” slides.