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”
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]
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)

[Wikipedia]

[Ed Bowlby, NOAA]

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

[Arjas, Gasbarra 1994]
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]

[Ed Bowlby, NOAA]

[Arjas, Gasbarra 1994]

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

[Fox, et al 2014]

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

• Bayesian statistics that is not parametric

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

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

[Ed Bowlby, NOAA]

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

[Arjas, Gasbarra 1994]

[Saria et al 2010]

[Fox, et al 2014]

[Arjas, Gasbarra 1994]

[Ewens, 1972; Hartl, Clark 2003]

[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

[Ed Bowlby, NOAA] [Fox, et al 2014]

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

[Arjas, Gasbarra 1994]

[Saria et al 2010]

[Lloyd et al 2012; Miller et al, 2010]

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

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

[Ed Bowlby, NOAA]

[Wikipedia.org]

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

[Saria et al 2010]

[Arjas, Gasbarra 1994]

[Fox, et al 2014]

[Lloyd et al 2012; Miller et al, 2010]

[Ewens, 1972; Hartl, Clark 2003]

[Sudderth, Jordan 2009]
Nonparametric Bayes

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

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

\(\text{De Finetti’s Theorem}\) (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

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

\[ \mu_1, \mu_2 \]
\[ \zeta_n \text{iid} \sim \text{Categorical}(\zeta_1, \zeta_2) \]
\[ \mu_k \text{iid} \sim \mathcal{N}(\mu_0, \Sigma_0) \]
\[ \zeta_1 \sim \text{Beta}(a_1, a_2) \]
\[ \zeta_2 = 1 - \zeta_1 \]

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

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

\[ z_n \overset{iid}{\sim} \text{Categorical}(\neu_1, \neu_2) \]

\[ \mu_k \overset{iid}{\sim} \mathcal{N}(\mu_0, \S_0) \]

\[ \neu_1 \overset{}{\sim} \text{Beta}(a_1, a_2) \]

\[ \neu_2 = 1 - \neu_1 \]

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 \((K=2 \text{ clusters})\)
Generative model

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

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

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

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

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

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

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

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

- Finite Gaussian mixture model (\(K=2\) 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 ($K=2$ 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 $\mu_1, \mu_2$
  \[ \mu_k \overset{iid}{\sim} \mathcal{N}(\mu_0, \Sigma_0) \]
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 \( \rho_1, \rho_2 \)
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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$a_1, a_2 > 0$$
Beta distribution review

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

- 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 $m$: $\Gamma(m) = (m - 1)!$

\[ \rho_1 \in (0, 1) \quad a_1, a_2 > 0 \]
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 \( m \): \( \Gamma(m) = (m - 1)! \)
- for \( x > 0 \): \( \Gamma(x) = x\Gamma(x - 1) \)

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

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

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

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

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

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

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

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

- What happens?

\[
\begin{align*}
\rho_1 & \in (0, 1) \\
a_1, a_2 & > 0
\end{align*}
\]
### 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?
- \(a = a_1 = a_2 \to 0\)
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?
  - $a = a_1 = a_2 \to 0$
  - $a = a_1 = a_2 \to \infty$
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?
  - \(a = a_1 = a_2 \rightarrow 0\)
  - \(a = a_1 = a_2 \rightarrow \infty\)
  - \(a_1 > a_2\)
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

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

\[a_1, a_2 > 0\]

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

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

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

- Beta is conjugate to Cat
Beta distribution review

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

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

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

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

- Beta is conjugate to Cat

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

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

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

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

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

- 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

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

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

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

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

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

- Gamma function \( \Gamma \)
- integer \( m \): \( \Gamma(m) = (m - 1)! \)
- 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
  \( \rho_1 \sim \text{Beta}(a_1, a_2), z \sim \text{Cat}(\rho_1, \rho_2) \)

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

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

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

- 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$a_k > 0$

$\rho_k \in (0, 1)$

$$\sum_{k} \rho_k = 1$$
Dirichlet distribution review

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

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

• What happens?
Dirichlet distribution review

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

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

- What happens?
Dirichlet distribution review

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

$a_k > 0$

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

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

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

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

$$a = a_k = 1$$

$$a = a_k \to 0$$
Dirichlet distribution review

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

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

- What happens?
  \[a = a_k = 1\] \[a = a_k \rightarrow 0\] \[a = a_k \rightarrow \infty\]
Dirichlet distribution review

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

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

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

[demo]
Dirichlet distribution review

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

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

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

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

- 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

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

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

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

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

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

• 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 $\ll K$ and random

Number of clusters grows with $N$

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

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

> \[ \begin{array}{cccccc}
> \rho_1 & \rho_2 & \rho_3 & \ldots & \rho_{1000} \\
> \end{array} \]
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

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

Number of clusters grows with $N$

• Components: number of latent groups

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

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

\[ \cdots \]

- 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.

- 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.

$\rho_1 \rho_2 \rho_3 \ldots \rho_{1000}$

- 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.

\[ \begin{align*}
\rho_1 & \quad \rho_2 & \quad \rho_3 & \quad \rho_{1000} \\
\end{align*} \]

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

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

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

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

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

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

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

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

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

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

- “Stick breaking”

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

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

- “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”

\[ V_1 \sim \text{Beta}(a_1, a_2 + a_3 + a_4) \quad \rho_1 = V_1 \]
\[ V_2 \sim \text{Beta}(a_2, a_3 + a_4) \quad \rho_2 = (1 - V_1)V_2 \]
\[ V_3 \sim \text{Beta}(a_3, a_4) \quad \rho_3 = (1 - V_1)(1 - V_2)V_3 \]
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: $\rho_{1:K} \sim \text{Dirichlet}(a_{1:K})$

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

- “Stick breaking”

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[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**: $a_k = 1, b_k = \alpha > 0$

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

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

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

• Beta → random distribution
• Dirichlet → random distribution over
• GEM / Dirichlet stick-breaking → random distribution over
• Dirichlet process → random distribution over
Distributions

- Beta → random distribution over 1, 2
Distributions

- Beta $\rightarrow$ random distribution over 1, 2
Distributions

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

- Beta $\rightarrow$ random distribution over 1, 2
- Dirichlet $\rightarrow$ random distribution over 1, 2, ..., $K$
Distributions

- Beta $\rightarrow$ random distribution over 1, 2
- Dirichlet $\rightarrow$ random distribution over 1, 2, $\ldots$, $K$
- GEM / Dirichlet stick-breaking $\rightarrow$ random distribution over 1, 2, $\ldots$
Distributions

- Beta → random distribution over $1, 2$

- Dirichlet → random distribution over $1, 2, \ldots, K$

- GEM / Dirichlet stick-breaking → random distribution over $1, 2, \ldots$
Distributions

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

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

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

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

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

- Dirichlet $\rightarrow$ random distribution over 1, 2, …, $K$

- GEM / Dirichlet stick-breaking $\rightarrow$ random distribution over 1, 2, …

\[ \rho = (\rho_1, \rho_2, \ldots) \sim \text{GEM}(\alpha) \]
\[ \phi_k \overset{iid}{\sim} G_0 \]
Distributions

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

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

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

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

• Beta → random distribution over 1, 2

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

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

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

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

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

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

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

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

• **Dirichlet process** $\rightarrow$
  random distribution over $\Phi$:
  $\rho = (\rho_1, \rho_2, \ldots) \sim \text{GEM}(\alpha)$
  $\phi_k \overset{iid}{\sim} G_0$
  $G = \sum_{k=1}^{\infty} \rho_k \delta_{\phi_k}$

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

• Gaussian mixture model
Dirichlet process mixture model

• Gaussian mixture model

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

- Gaussian mixture model

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

- Gaussian mixture model

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

- Gaussian mixture model

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

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

\[ G = \mathbb{D} \mathcal{P} \left( \alpha, \mathcal{N}(\mu_0, \Sigma_0) \right) \]
Dirichlet process mixture model

- Gaussian mixture model

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

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

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

- Gaussian mixture model

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

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

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

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

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

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

\[
\begin{align*}
  z_n &\sim \text{Categorical}(\rho) \\
  \mu^*_n &= \mu_{z_n}
\end{align*}
\]
Dirichlet process mixture model

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

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

- Gaussian mixture model

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

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

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

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

\[ \mu_n^* = \mu_{z_n} \]

- i.e. \( \mu_n^* \overset{iid}{\sim} G \)
Dirichlet process mixture model

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

- \( z_n \overset{iid}{\sim} \text{Categorical}(\rho) \)
- \( \mu_n^* = \mu_{z_n} \)
- i.e. \( \mu_n^* \overset{iid}{\sim} G \)

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

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

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

- \[ z_n \overset{iid}{\sim} \text{Categorical}(\rho) \]
- \[ \mu_n^* = \mu_{z_n} \]

- i.e. \[ \mu_n^* \overset{iid}{\sim} G \]

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

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

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

- \[ z_n \overset{iid}{\sim} \text{Categorical}(\rho) \]
  \[ \mu_n^* = \mu_{z_n} \]

- i.e. \[ \mu_n^* \overset{iid}{\sim} G \]

- \[ x_n \overset{indep}{\sim} \mathcal{N}(\mu_n^*, \Sigma) \]

[demo]
Dirichlet process mixture model

- More generally

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

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

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

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

\[ \mu_n^* = \mu_{z_n} \]

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

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

- More generally

\[ \rho = (\rho_1, \rho_2, \ldots) \sim \text{GEM}(\alpha) \]
\[ \phi_k \overset{iid}{\sim} G_0 \quad k = 1, 2, \ldots \]

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

\[ z_n \overset{iid}{\sim} \text{Categorical}(\rho) \]
\[ \mu_n^* = \mu_{z_n} \]

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

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

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

  \[ \phi_k \overset{iid}{\sim} G_0 \quad k = 1, 2, \ldots \]
  i.e. \[ G = \sum_{k=1}^{\infty} \rho_k \delta_{\mu_k} \overset{d}{=} \text{DP}(\alpha, \mathcal{N}(\mu_0, \Sigma_0)) \]

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

  \[ \mu^*_n = \mu_{z_n} \]
  i.e. \[ \mu^*_n \overset{iid}{\sim} G \]

- \[ x_n \overset{indep}{\sim} \mathcal{N}(\mu^*_n, \Sigma) \]
Dirichlet process mixture model

- More generally
  \[ \rho = (\rho_1, \rho_2, \ldots) \sim \text{GEM}(\alpha) \]
  \[ \phi_k \overset{iid}{\sim} G_0 \]
  \[ k = 1, 2, \ldots \]
- i.e. \[ G = \sum_{k=1}^{\infty} \rho_k \delta_{\mu_k} \overset{d}{=} \text{DP}(\alpha, \mathcal{N}(\mu_0, \Sigma_0)) \]

- \[ z_n \overset{iid}{\sim} \text{Categorical}(\rho) \]
- \[ \mu_n^* = \mu_{z_n} \]
- i.e. \[ \mu_n^* \overset{iid}{\sim} G \]

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

• More generally

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

\[ \phi_k \sim G_0 \quad \text{iid} \quad k = 1, 2, \ldots \]

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

\[ z_n \sim \text{Categorical}(\rho) \]

\[ \mu^*_n = \mu_{z_n} \]

i.e. \[ \mu^*_n \sim G \]

\[ x_n \overset{\text{indep}}{\sim} \mathcal{N}(\mu^*_n, \Sigma) \]
Dirichlet process mixture model

- More generally
  \[ \rho = (\rho_1, \rho_2, \ldots) \sim \text{GEM}(\alpha) \]
  \[ \phi_k \overset{iid}{\sim} G_0 \quad k = 1, 2, \ldots \]
- i.e. \( G = \sum_{k=1}^{\infty} \rho_k \delta_{\phi_k} \overset{d}{=} \text{DP}(\alpha, G_0) \)

- \( z_n \overset{iid}{\sim} \text{Categorical}(\rho) \)
- \( \mu_n^* = \mu_{z_n} \)
- i.e. \( \mu_n^* \overset{iid}{\sim} G \)

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

- More generally

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

$$\phi_k \overset{iid}{\sim} G_0 \quad k = 1, 2, \ldots$$

- i.e. $G = \sum_{k=1}^{\infty} \rho_k \delta_{\phi_k} \overset{d}{=} \text{DP}(\alpha, G_0)$

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

$$\theta_n = \phi_{z_n}$$

- i.e. $\mu_n^* \overset{iid}{\sim} G$

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

• More generally

\[ \rho = (\rho_1, \rho_2, \ldots) \sim \text{GEM}(\alpha) \]
\[ \phi_k \overset{iid}{\sim} G_0 \quad k = 1, 2, \ldots \]

• i.e. \( G = \sum_{k=1}^{\infty} \rho_k \delta_{\phi_k} \overset{d}{=} \text{DP}(\alpha, G_0) \)

\[ z_n \overset{iid}{\sim} \text{Categorical}(\rho) \]
\[ \theta_n = \phi_{z_n} \]

• i.e. \( \theta_n \overset{iid}{\sim} G \)

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

- More generally
  \[ \rho = (\rho_1, \rho_2, \ldots) \sim \text{GEM}(\alpha) \]
  \[ \phi_k \overset{iid}{\sim} G_0 \quad k = 1, 2, \ldots \]
  i.e. \[ G = \sum_{k=1}^{\infty} \rho_k \delta_{\phi_k} \overset{d}{=} \text{DP}(\alpha, G_0) \]

- \[ z_n \overset{iid}{\sim} \text{Categorical}(\rho) \]
  \[ \theta_n = \phi_{z_n} \]
  i.e. \[ \theta_n \overset{iid}{\sim} G \]

- \[ x_n \overset{\text{indep}}{\sim} F(\theta_n) \]
Dirichlet process mixture model

- More generally
  \[ \rho = (\rho_1, \rho_2, \ldots) \sim \text{GEM}(\alpha) \]
  \[ \phi_k \overset{iid}{\sim} G_0 \]
  \[ k = 1, 2, \ldots \]

- i.e. \[ G = \sum_{k=1}^{\infty} \rho_k \delta_{\phi_k} \overset{d}{=} \text{DP}(\alpha, G_0) \]

- \[ z_n \overset{iid}{\sim} \text{Categorical}(\rho) \]
- \[ \theta_n = \phi_{z_n} \]

- i.e. \[ \theta_n \overset{iid}{\sim} G \]

- \[ x_n \overset{indep}{\sim} F(\theta_n) \]

[Antoniak 1974; Ferguson 1983; West, Müller, Escobar 1994; Escobar, West 1995; MacEachern, Müller 1998]
DPMM Exercises

- Code your own DPMM simulator
- How does the number of clusters vary with $N$? (theory/simulations)
- How does the number of clusters vary with $\alpha$? (theory/simulations)
- For fixed $N$, what is the distribution over # clusters?
- What is the marginal distribution of the GEM on any finite collection of positive integers?
- Let $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_J$ be disjoint subsets of $\Phi$; when $G \sim \text{DP}(\alpha, G_0)$, what is the distribution of $\text{DPMM}(\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_J)$?

\[ \rho_2 \]
\[ \phi_2 \]
DPMM Exercises

• Code your own DPMM simulator
DPMM Exercises

- Code your own DPMM simulator
- How does the number of clusters vary with $N$? (theory/simulations)
- How does the number of clusters vary with $\alpha$? (theory/simulations)
- For fixed $N$, what is the distribution over # clusters?
- What is the marginal distribution of the GEM on any finite collection of positive integers?
- Let $A_1, A_2, \ldots, A_J$ be disjoint subsets of $\Phi$; when $G \sim\text{DP}(\alpha, G_0)$, what is the distribution of $(G(A_1), G(A_2), \ldots, G(A_J))$?
DPMM Exercises

• Code your own DPMM simulator
• How does the number of clusters vary with $N$? (theory/simulations)
• How does the number of clusters vary with $\alpha$? (theory/simulations)
DPMM Exercises

- Code your own DPMM simulator
- How does the number of clusters vary with $N$? (theory/simulations)
- How does the number of clusters vary with $\alpha$? (theory/simulations)
- For fixed $N$, what is the distribution over # clusters?


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

