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

$$\mathbb{P}(\text{parameters})$$

Nonparametric Bayes

- Bayesian methods that are not parametric
- Bayesian

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

Nonparametric Bayes

- Bayesian methods that are not parametric
- Bayesian

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

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)

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)

[wikipedia.org]

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)

[wikipedia.org]

"Wikipedia phenomenon"

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)

[wikipedia.org]

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)

[wikipedia.org]


[Ed Bowlby, NOAA]

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)

[wikipedia.org]

[Ed Bowlby, NOAA]

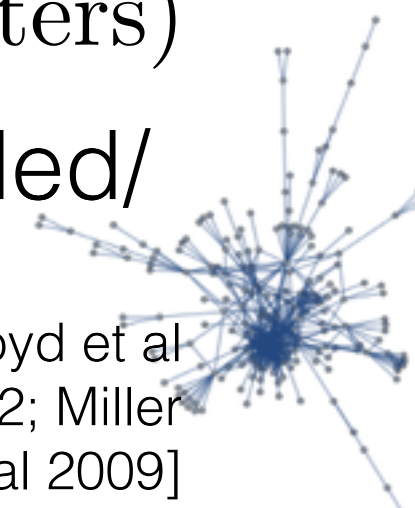
[Fox et al 2014]

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)



[Lloyd et al 2012; Miller et al 2009]

[wikipedia.org]

[Ed Bowlby, NOAA]

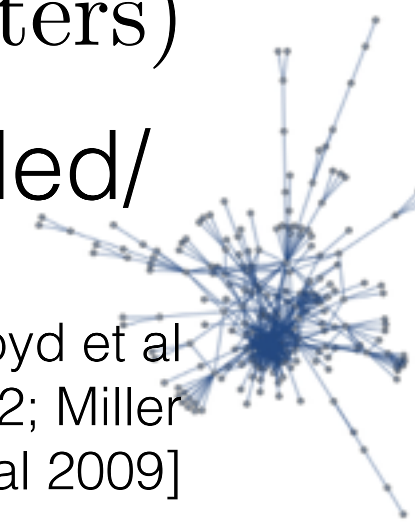
[Fox et al 2014]

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)



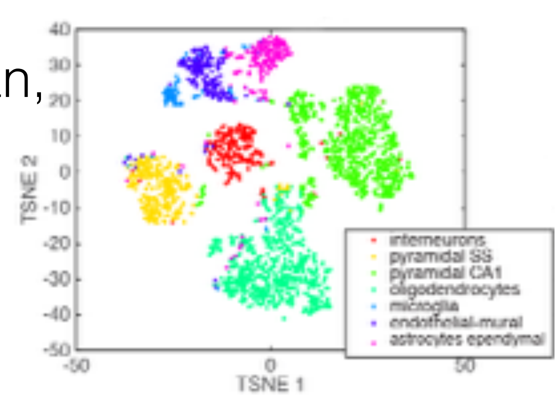
[Lloyd et al 2012; Miller et al 2009]

[wikipedia.org]

[Ed Bowlby, NOAA]

[Fox et al 2014]

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

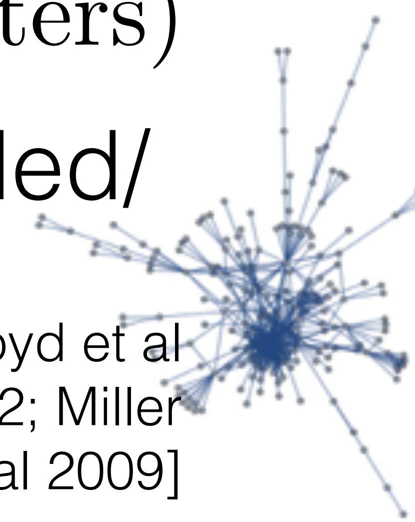


Nonparametric Bayes

- 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)



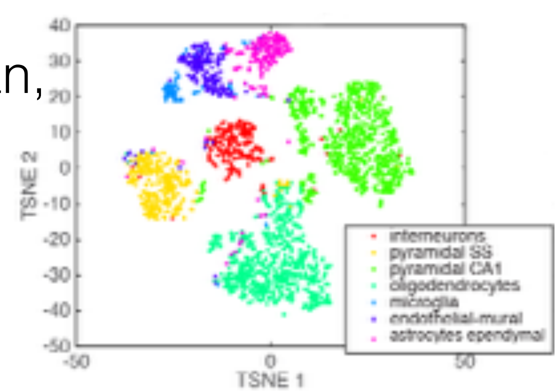
[Lloyd et al 2012; Miller et al 2009]

[wikipedia.org]

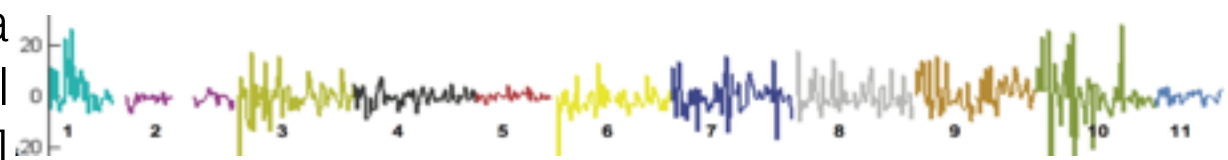
[Ed Bowlby, NOAA]

[Fox et al 2014]

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



[Saria et al 2010]

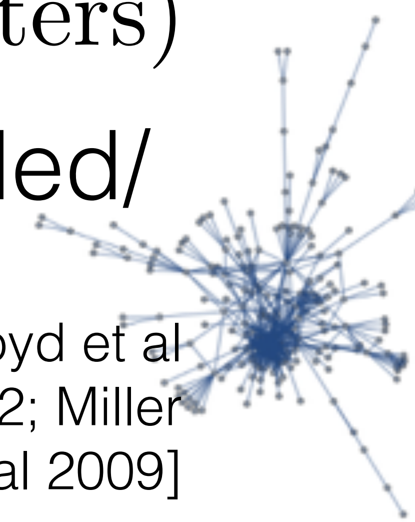


Nonparametric Bayes

- 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)



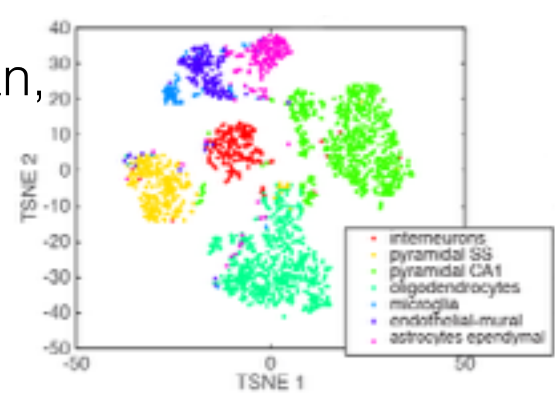
[Lloyd et al 2012; Miller et al 2009]

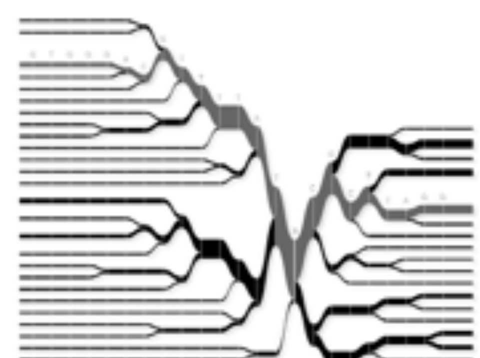
[wikipedia.org]

[Ed Bowlby, NOAA]

[Fox et al 2014]

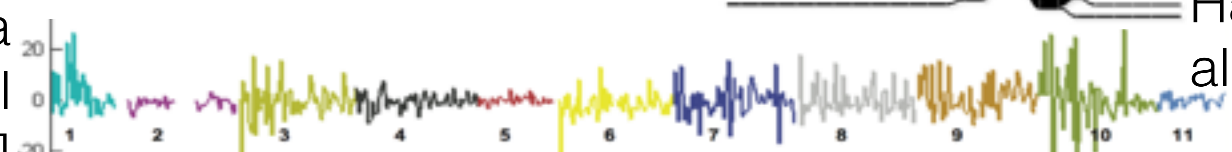
[Prabhakaran, Azizi, Carr, Pe'er 2016]





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

[Saria et al 2010]

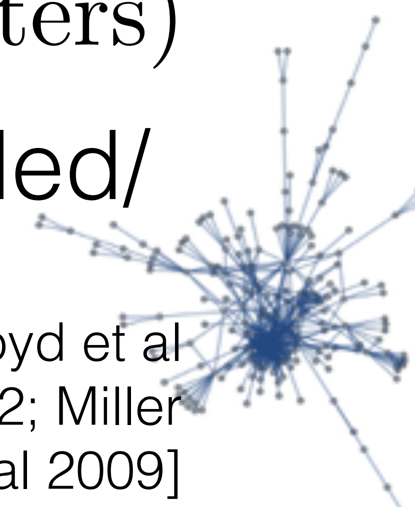


Nonparametric Bayes

- Bayesian methods that are not parametric
- Bayesian

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

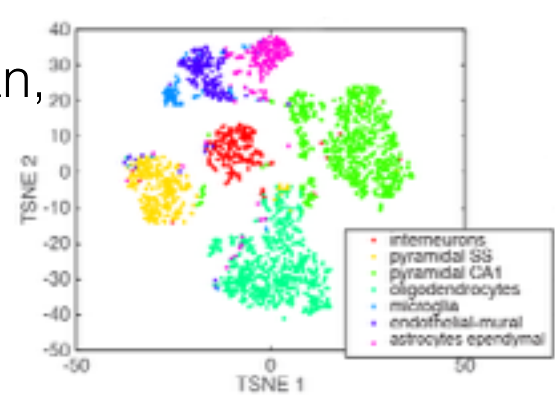
- Not parametric (i.e. not finite parameter, unbounded/ growing/infinite number of parameters)

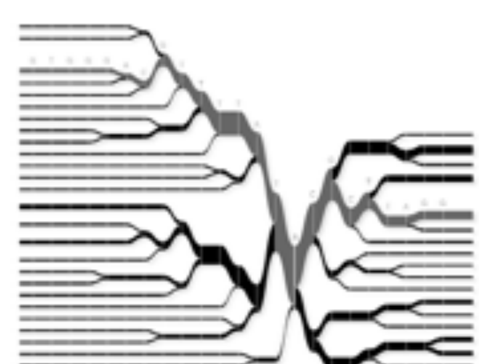

[Lloyd et al 2012; Miller et al 2009]


[Ed Bowlby, NOAA]


[Fox et al 2014]

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

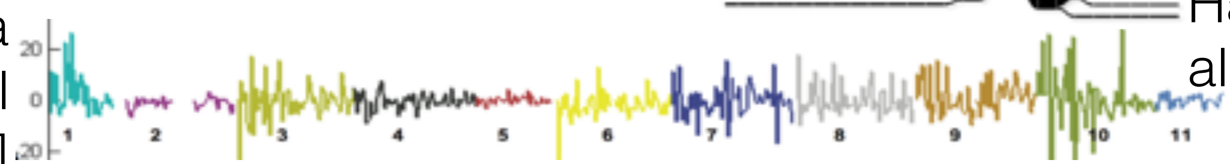



[Ewens 1972; Hartl, Clark 2003; Harris et al 2017]


[ESO/ L. Calçada/ M. Kornmesser 2017]

[Del Pozzo et al 2017, 2018]

[Saria et al 2010]

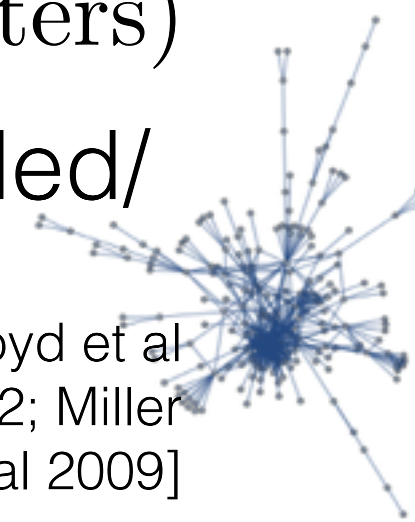


Nonparametric Bayes

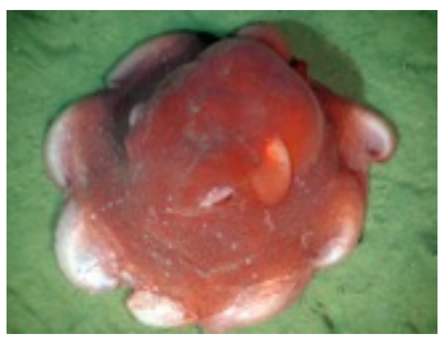
- Bayesian methods that are not parametric
- Bayesian

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

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



[Lloyd et al 2012; Miller et al 2009]



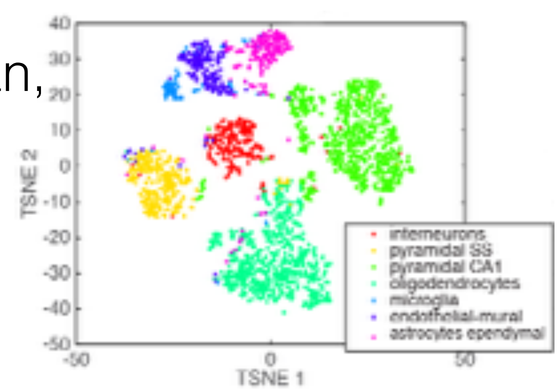
[Ed Bowlby, NOAA]

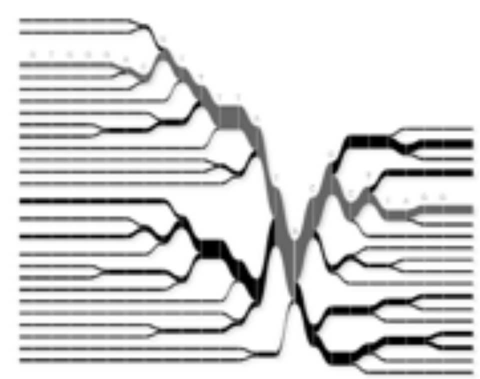
[Fox et al 2014]

[MIT xPRO]

[Lan et al 2015]

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



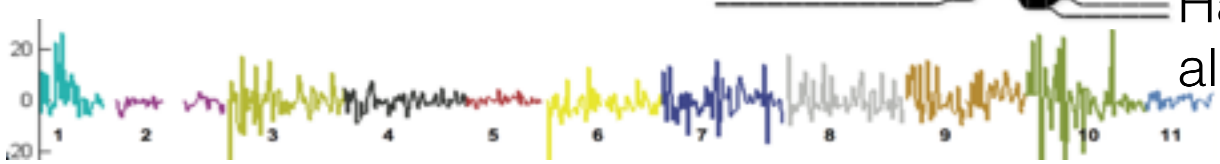


[Ewens 1972; Hartl, Clark 2003; Harris et al 2017]

[ESO/ L. Calçada/ M. Kornmesser 2017]

[Del Pozzo et al 2017, 2018]

[Saria et al 2010]

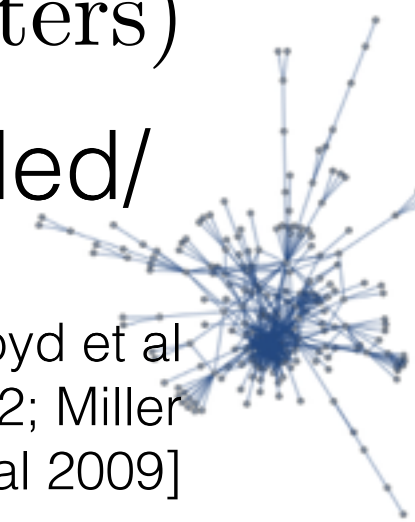


Nonparametric Bayes

- 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)



[Lloyd et al 2012; Miller et al 2009]

[Ed Bowlby, NOAA]

[Fox et al 2014]

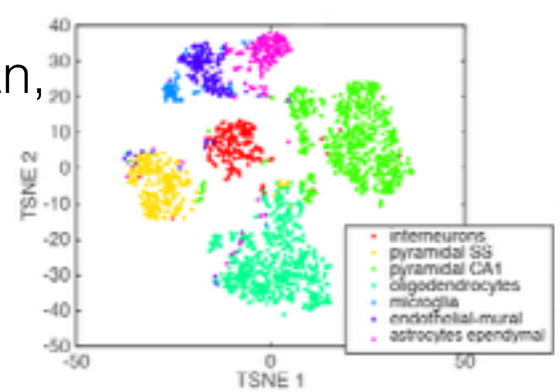
[MIT xPRO]

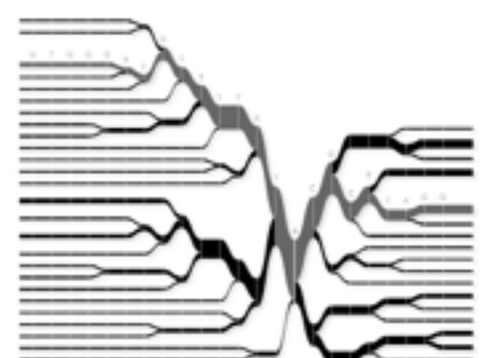
[Lan et al 2015]

[ESO/ L. Calçada/ M. Kornmesser 2017]

[Del Pozzo et al 2017, 2018]

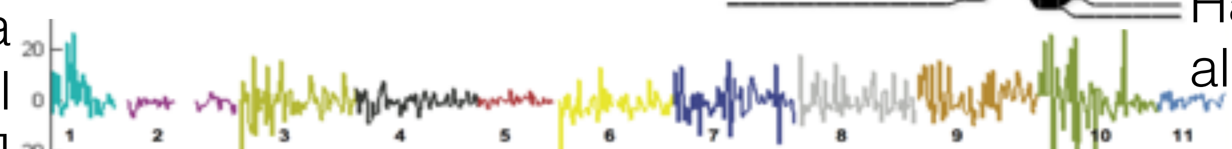
[Prabhakaran, Azizi, Carr, Pe'er 2016]





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

[Saria et al 2010]



[Xu et al 2015]

[Cassidy et al 2015]

Nonparametric Bayes

- A theoretical motivation: De Finetti's Theorem

Nonparametric Bayes

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

Nonparametric Bayes

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

[Hewitt, Savage 1955; Aldous 1983]

Nonparametric Bayes

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

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

[Hewitt, Savage 1955; Aldous 1983]

Nonparametric Bayes

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

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

- Motivates:

[Hewitt, Savage 1955; Aldous 1983]

Nonparametric Bayes

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

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

- Motivates:
  - Parameters and likelihoods

[Hewitt, Savage 1955; Aldous 1983]

Nonparametric Bayes

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

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

- Motivates:
  - Parameters and likelihoods
  - Priors

[Hewitt, Savage 1955; Aldous 1983]

Nonparametric Bayes

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

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

- Motivates:
  - Parameters and likelihoods
  - Priors
  - "Nonparametric Bayesian" priors

[Hewitt, Savage 1955; Aldous 1983]

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

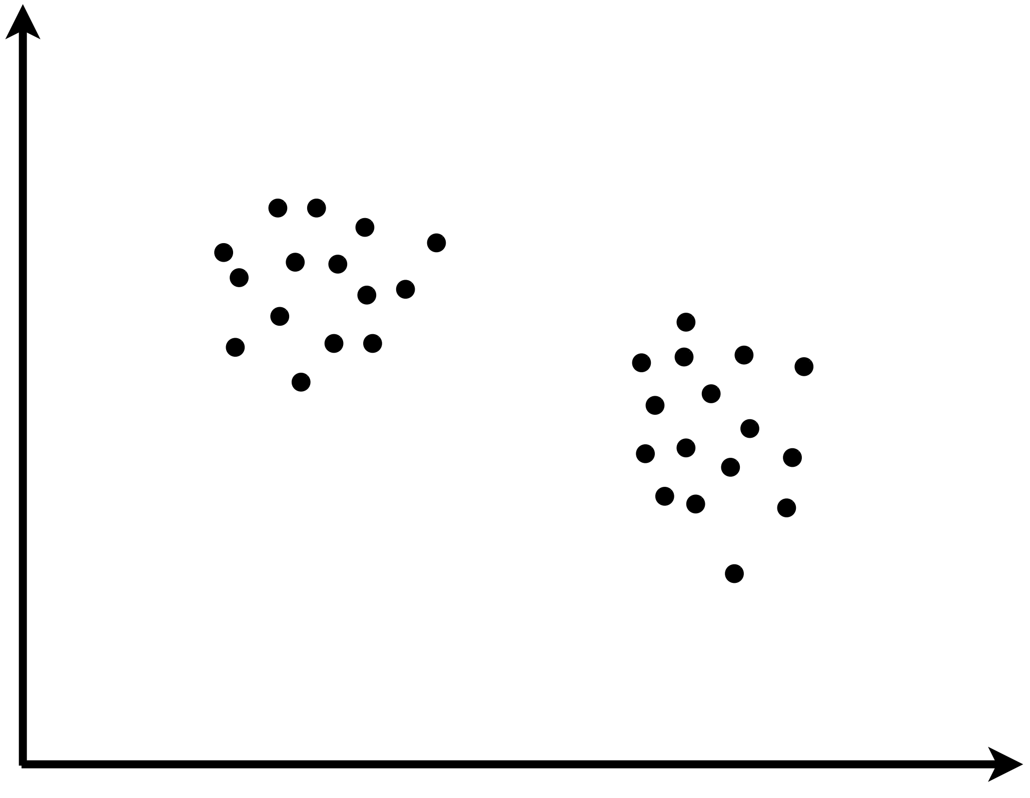
- Big questions
  - Why NPBayes?
  - What does an infinite/growing number of parameters really mean (in NPBayes)?

Roadmap

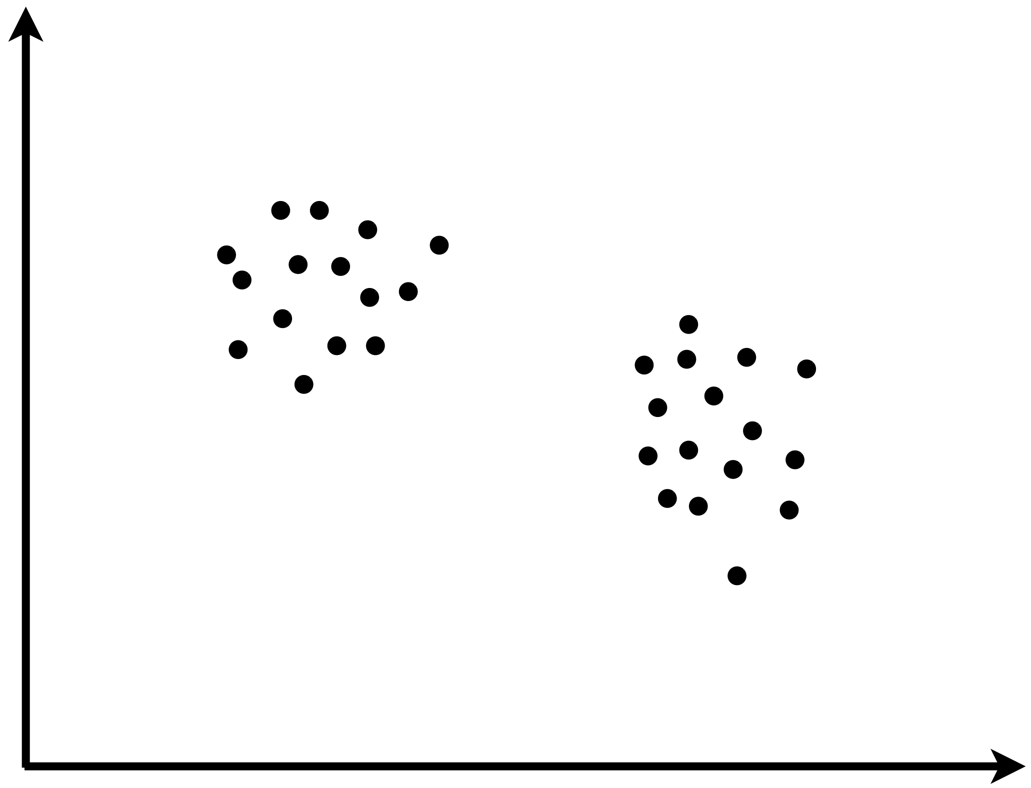
- Example problem: clustering
- Example NPBayes model: Dirichlet process
- Chinese restaurant process
- Inference
- Venture further into the wild world of Nonparametric Bayes

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

Generative model



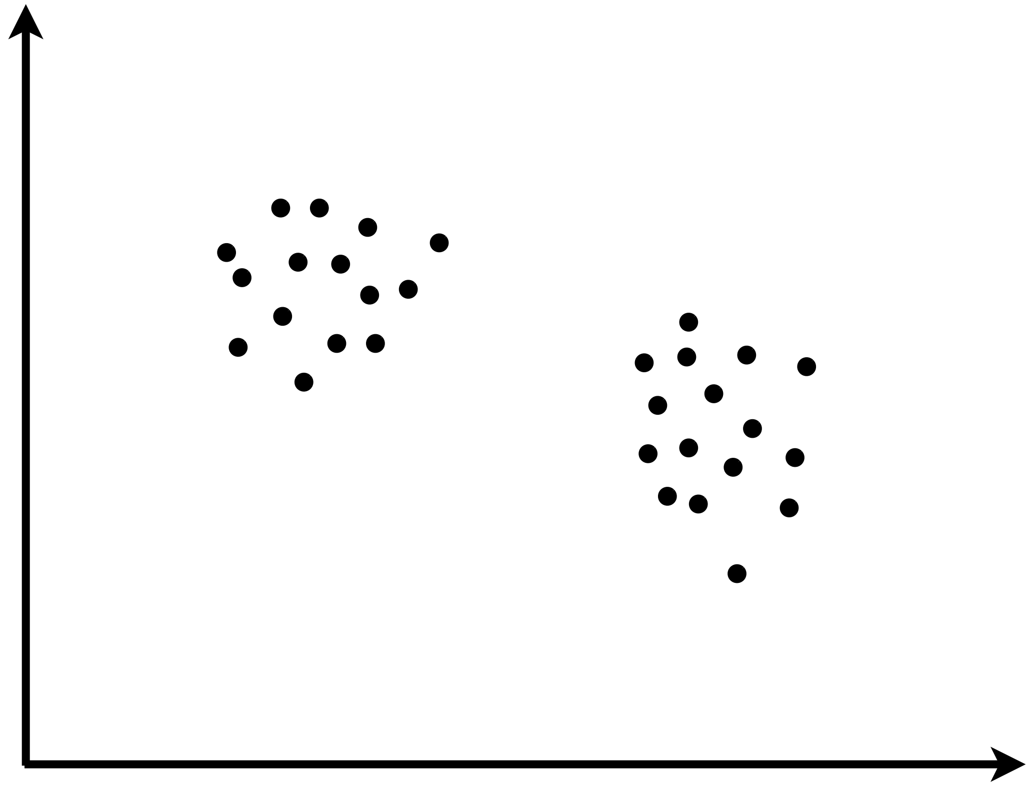
Generative model



- Finite Gaussian mixture model (K=2 clusters)

Generative model

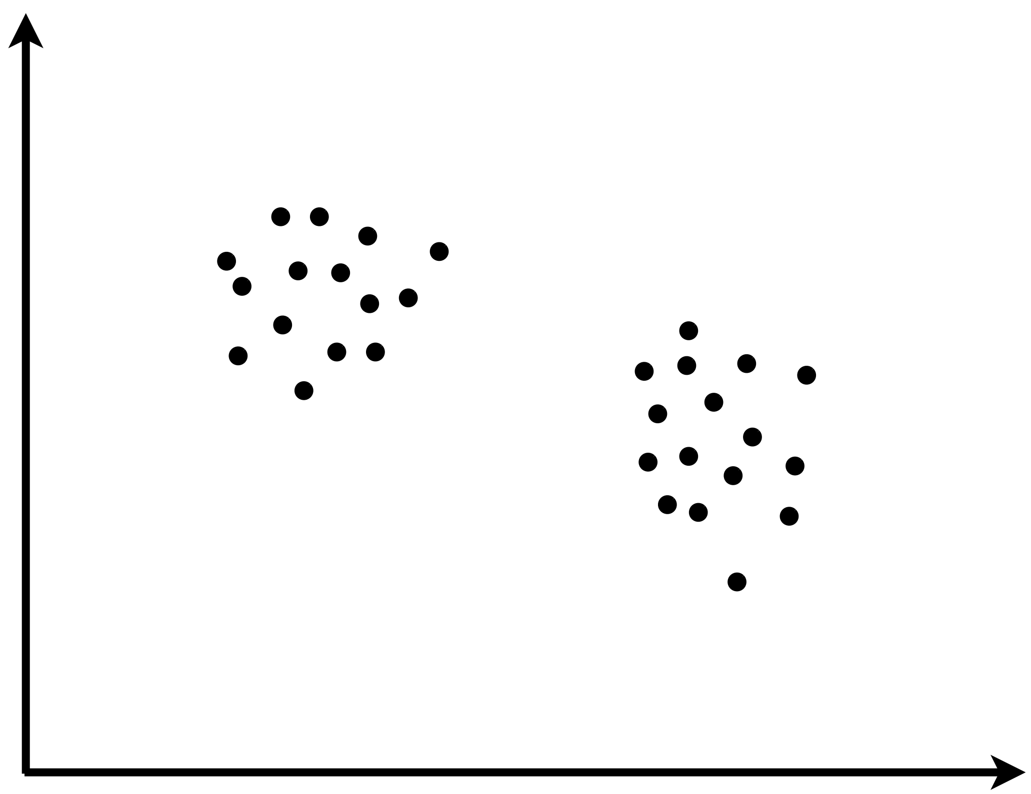
$$\mathbb{P}(\text{parameters}|\text{data}) \propto \mathbb{P}(\text{data}|\text{parameters})\mathbb{P}(\text{parameters})$$



- Finite Gaussian mixture model (K=2 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 clusters)

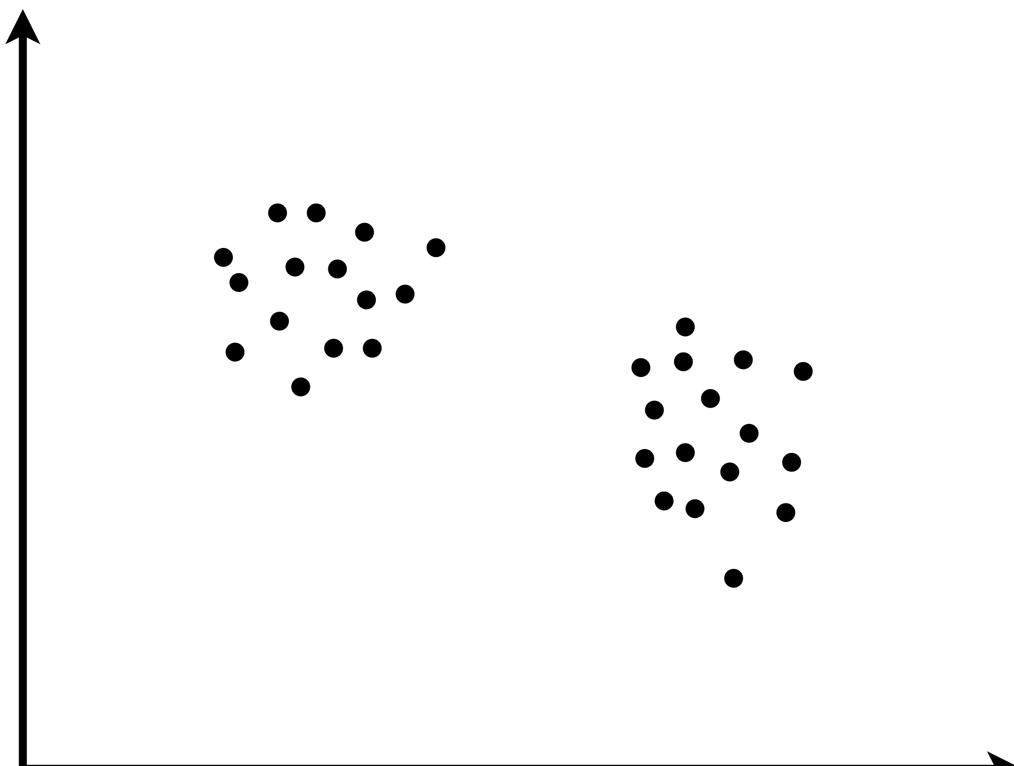
$$z_n \overset{iid}{\sim} \text{Categorical}(\rho_1, \rho_2)$$

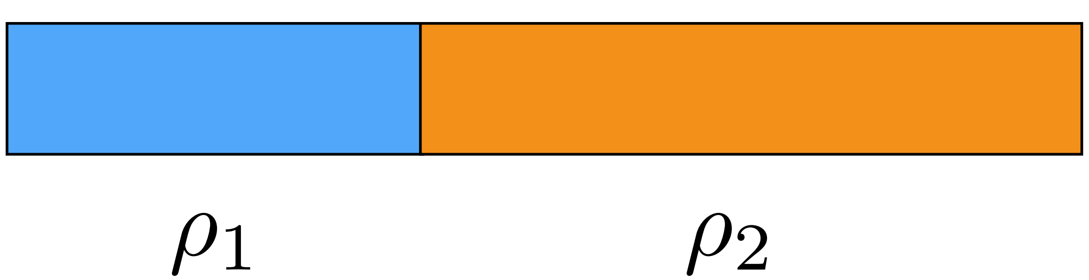
Generative model

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

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

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



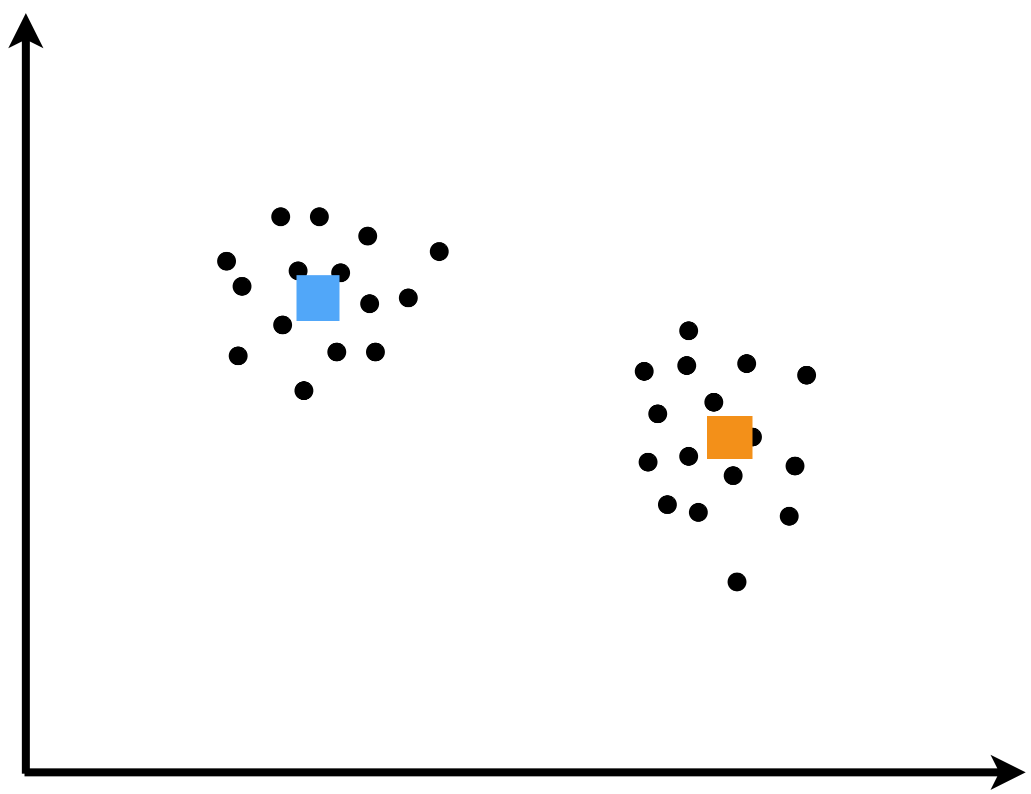


Generative model

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

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

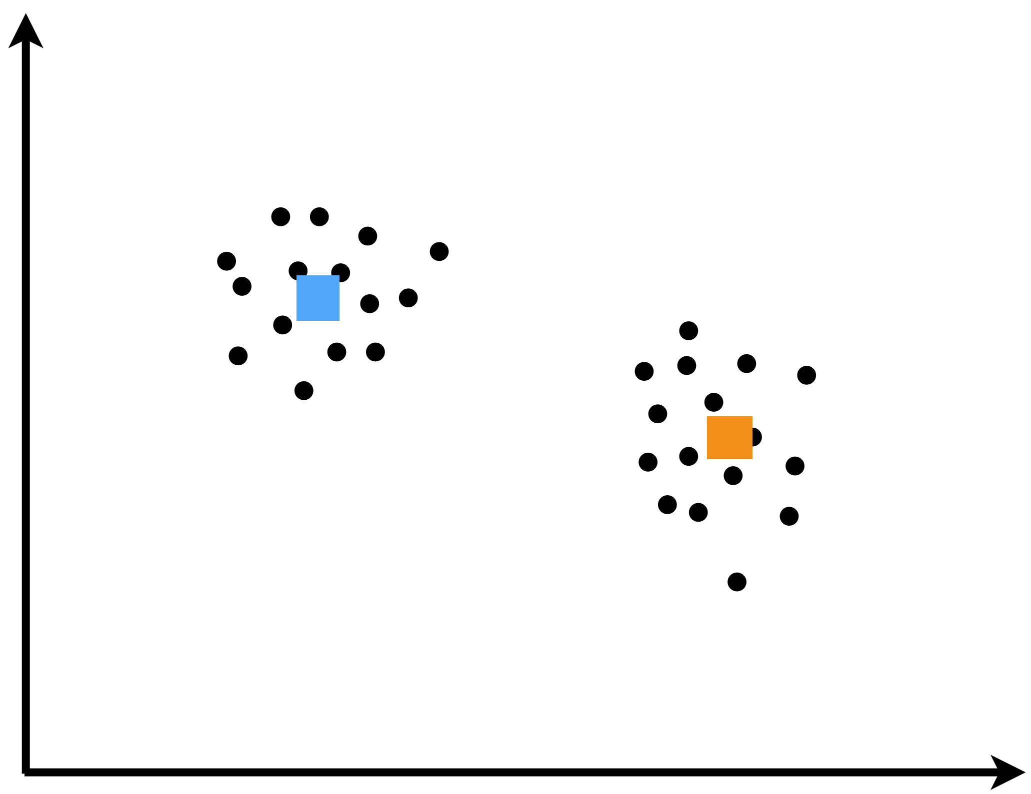
$$z_n \overset{iid}{\sim} \text{Categorical}(\rho_1, \rho_2)$$
$$x_n \overset{indep}{\sim} \mathcal{N}(\mu_{z_n}, \Sigma)$$

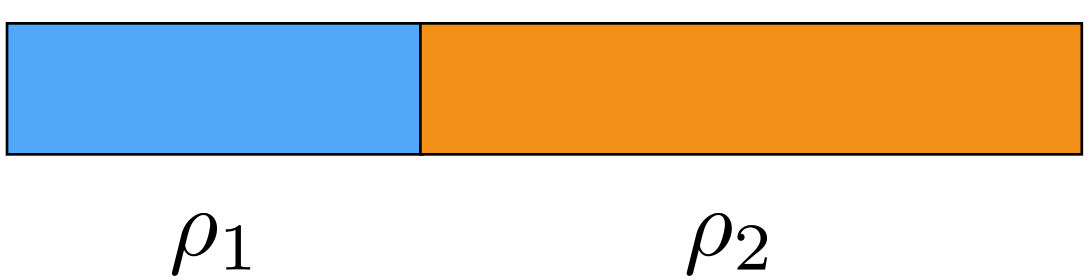


ρ_1 ρ_2

Generative model

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





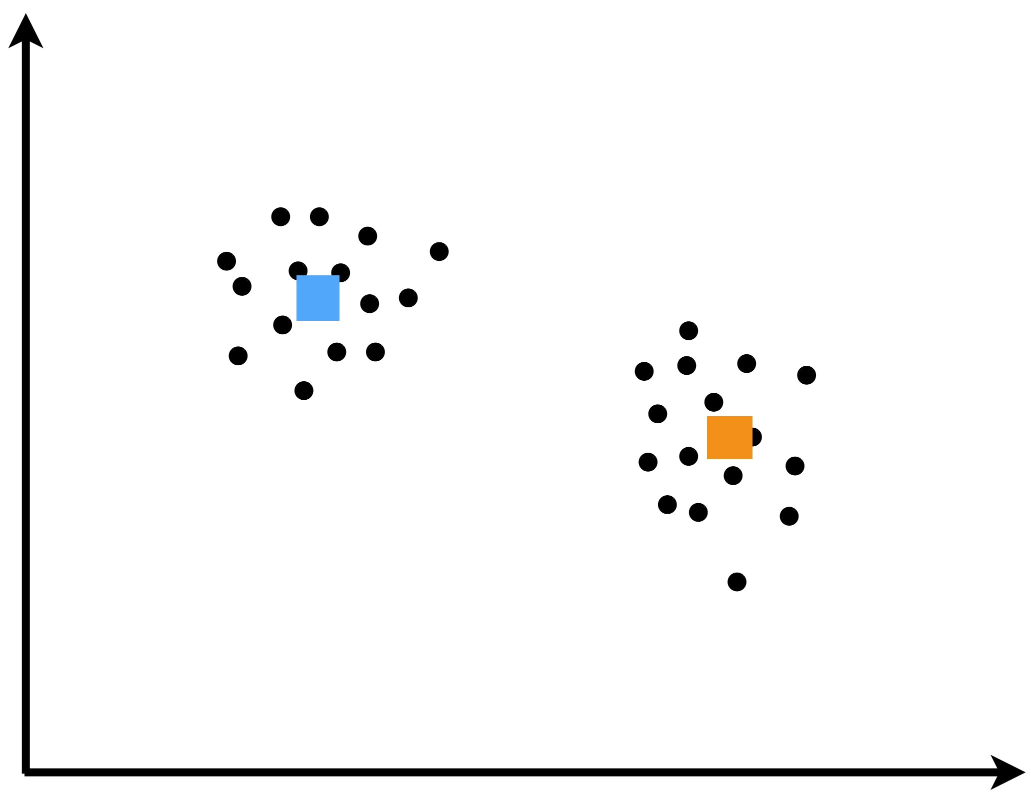
- Finite Gaussian mixture model (*K*=2 clusters)

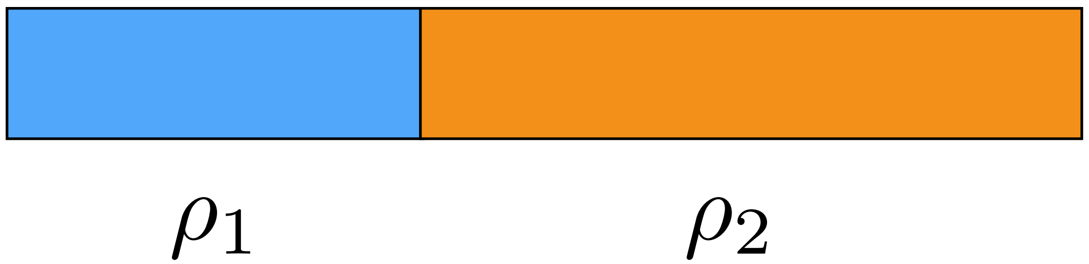
$$z_n \overset{iid}{\sim} \text{Categorical}(\rho_1, \rho_2)$$
$$x_n \overset{indep}{\sim} \mathcal{N}(\mu_{z_n}, \Sigma)$$

- Don't know μ_1, μ_2

Generative model

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





- Finite Gaussian mixture model (K=2 clusters)

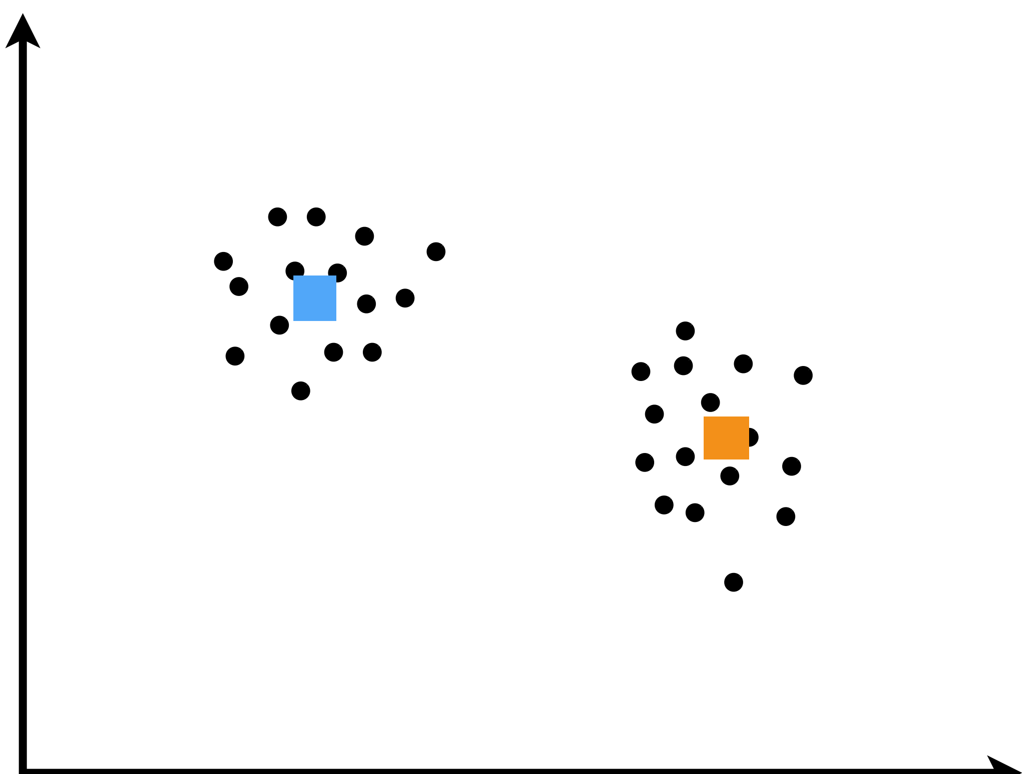
$$z_n \overset{iid}{\sim} \text{Categorical}(\rho_1, \rho_2)$$
$$x_n \overset{indep}{\sim} \mathcal{N}(\mu_{z_n}, \Sigma)$$

- Don't know μ_1, μ_2

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

Generative model

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



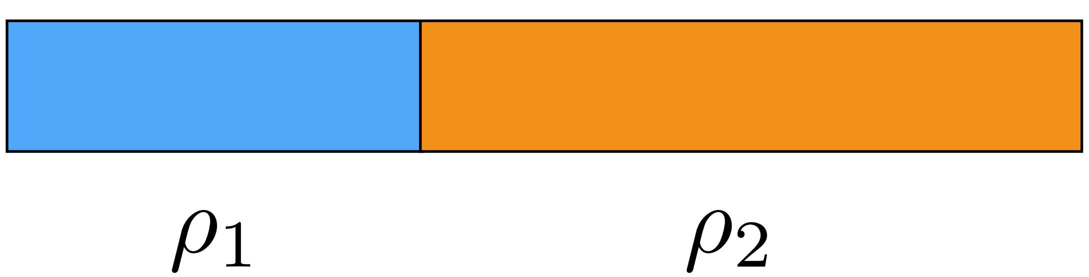
- Finite Gaussian mixture model (K=2 clusters)

$$z_n \overset{iid}{\sim} \text{Categorical}(\rho_1, \rho_2)$$
$$x_n \overset{indep}{\sim} \mathcal{N}(\mu_{z_n}, \Sigma)$$

- Don't know μ_1, μ_2

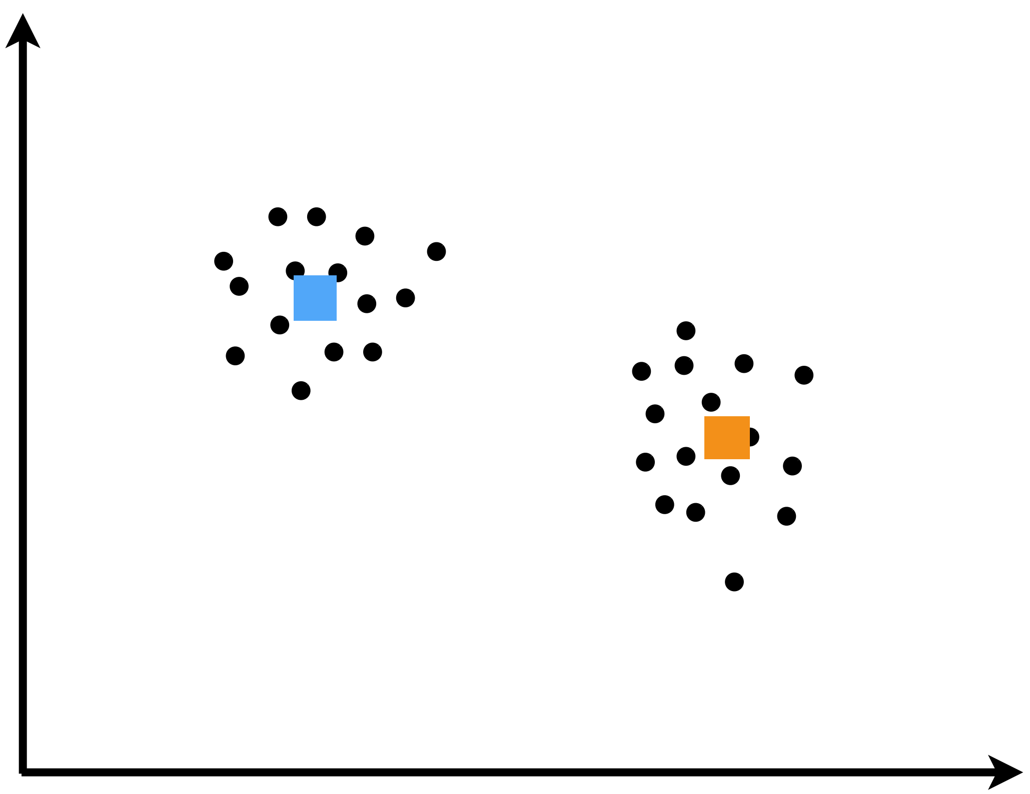
$$\mu_k \overset{iid}{\sim} \mathcal{N}(\mu_0, \Sigma_0)$$

- Don't know ρ_1, ρ_2



Generative model

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



- Finite Gaussian mixture model (K=2 clusters)

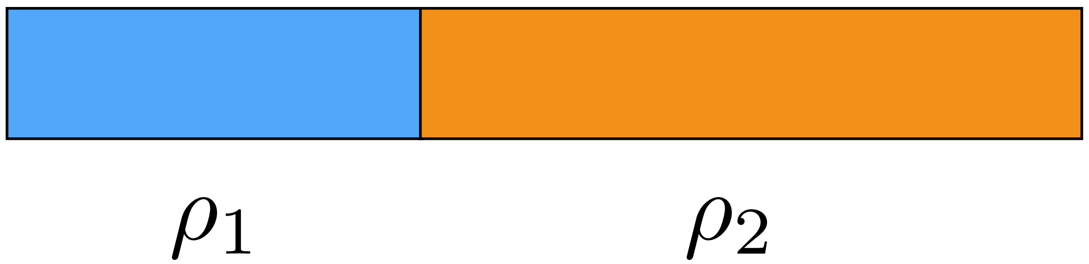
$$z_n \overset{iid}{\sim} \text{Categorical}(\rho_1, \rho_2)$$
$$x_n \overset{indep}{\sim} \mathcal{N}(\mu_{z_n}, \Sigma)$$

- Don't know μ_1, μ_2

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

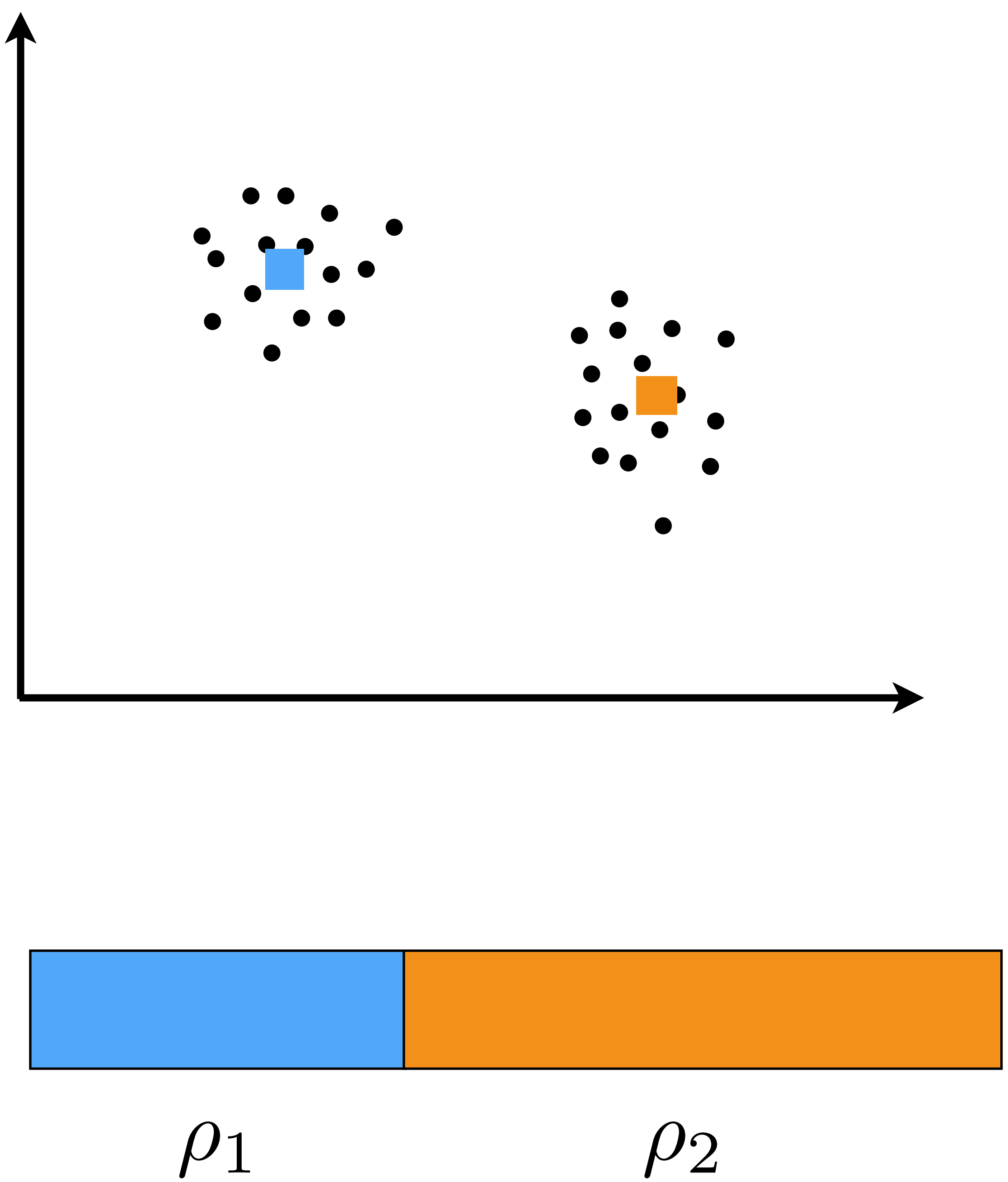
- Don't know ρ_1, ρ_2

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



Generative model

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



- Finite Gaussian mixture model (K=2 clusters)

$$z_n \overset{iid}{\sim} \text{Categorical}(\rho_1, \rho_2)$$
$$x_n \overset{indep}{\sim} \mathcal{N}(\mu_{z_n}, \Sigma)$$

- Don't know μ_1, μ_2

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

- Don't know ρ_1, ρ_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 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 μ_1, μ_2

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

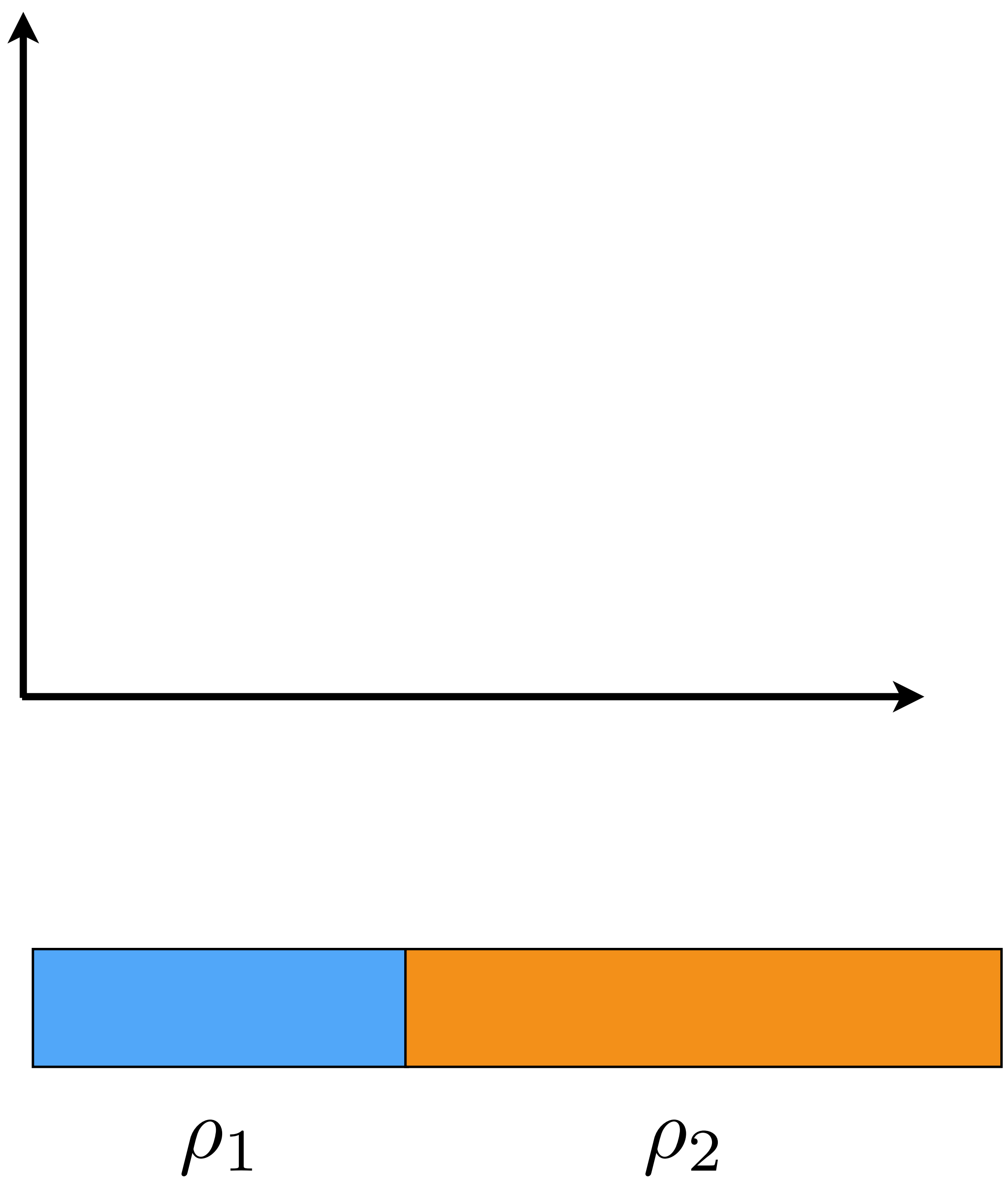
- Don't know ρ_1, ρ_2

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

- 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 clusters)

$$z_n \stackrel{iid}{\sim} \text{Categorical}(\rho_1, \rho_2)$$
$$x_n \stackrel{indep}{\sim} \mathcal{N}(\mu_{z_n}, \Sigma)$$

- Don't know μ_1, μ_2

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

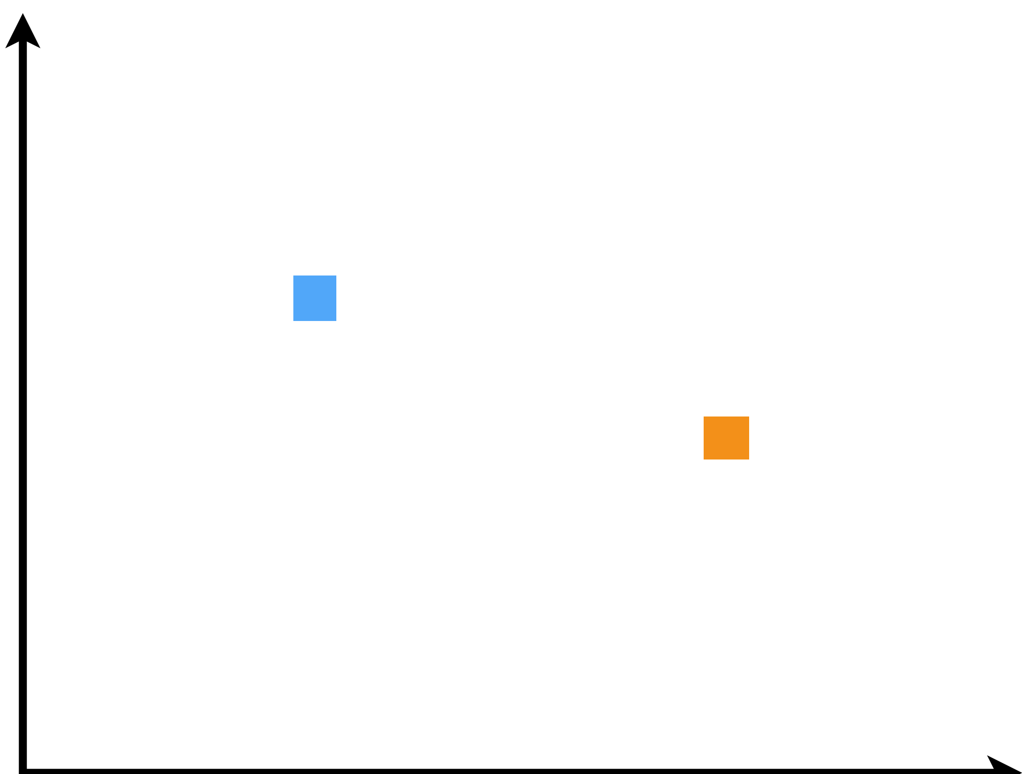
- Don't know ρ_1, ρ_2

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

- 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 clusters)

$$z_n \overset{iid}{\sim} \text{Categorical}(\rho_1, \rho_2)$$
$$x_n \overset{indep}{\sim} \mathcal{N}(\mu_{z_n}, \Sigma)$$

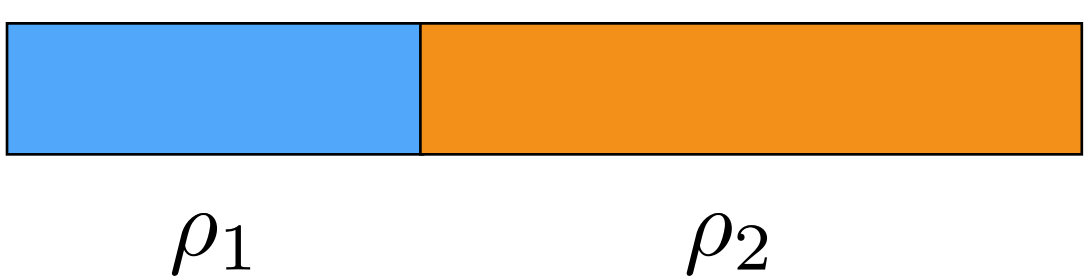
- Don't know μ_1, μ_2

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

- Don't know ρ_1, ρ_2

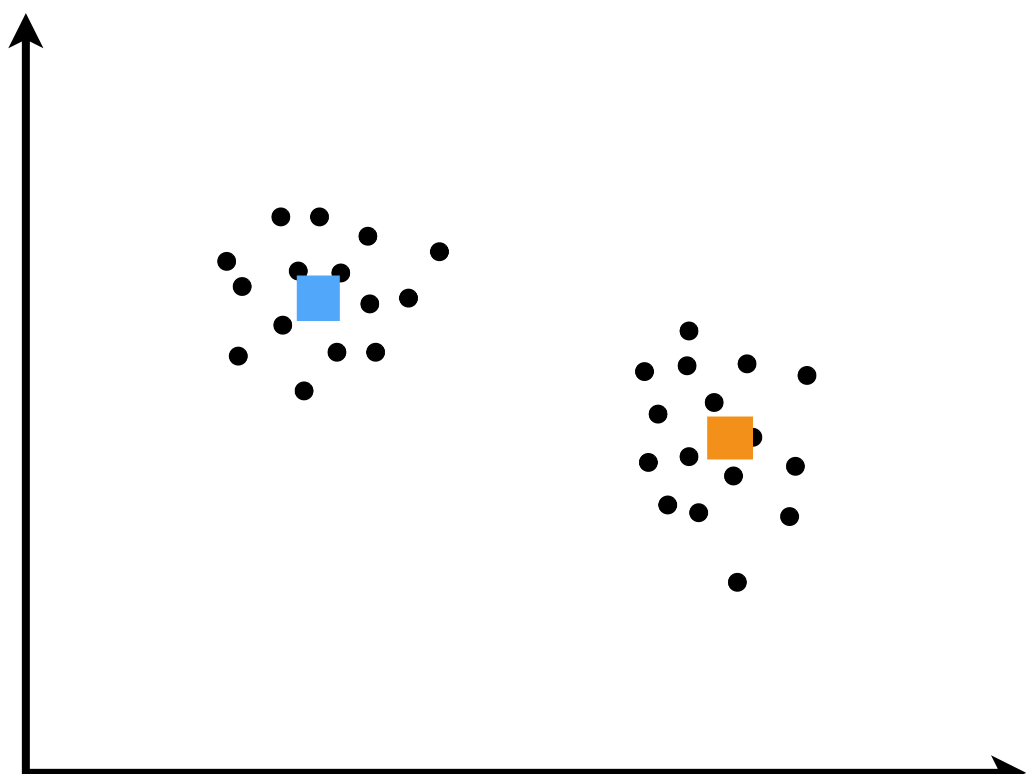
$$\rho_1 \sim \text{Beta}(a_1, a_2)$$
$$\rho_2 = 1 - \rho_1$$

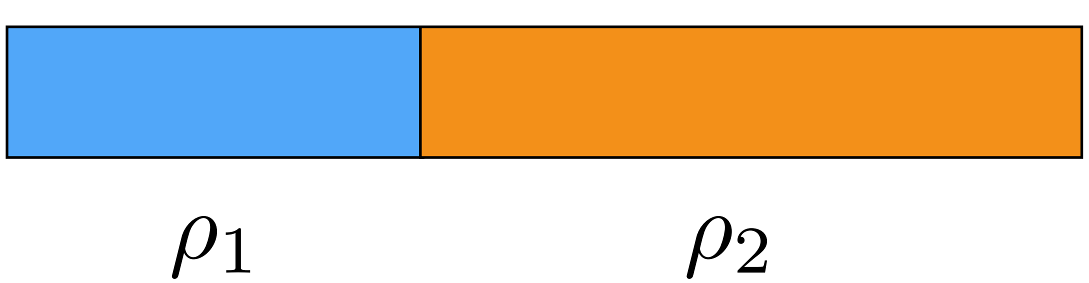
- Inference goal: assignments of data points to clusters, cluster parameters



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 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 μ_1, μ_2

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

- Don't know ρ_1, ρ_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

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

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

$$a_1, a_2 > 0$$

Beta distribution review

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

$$\mathrm{Beta}(\rho_1|a_1,a_2) = \frac{\Gamma(a_1+a_2)}{\Gamma(a_1)\Gamma(a_2)}\rho_1^{a_1-1}(1-\rho_1)^{a_2-1} \qquad a_1, a_2 > 0$$

- Gamma function Γ

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 Γ
  - integer *m*: $\Gamma(m + 1) = m!$

Beta distribution review

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

$\rho_1 \in (0, 1)$

$a_1, a_2 > 0$

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

Beta distribution review

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

$\rho_1 \in (0, 1)$

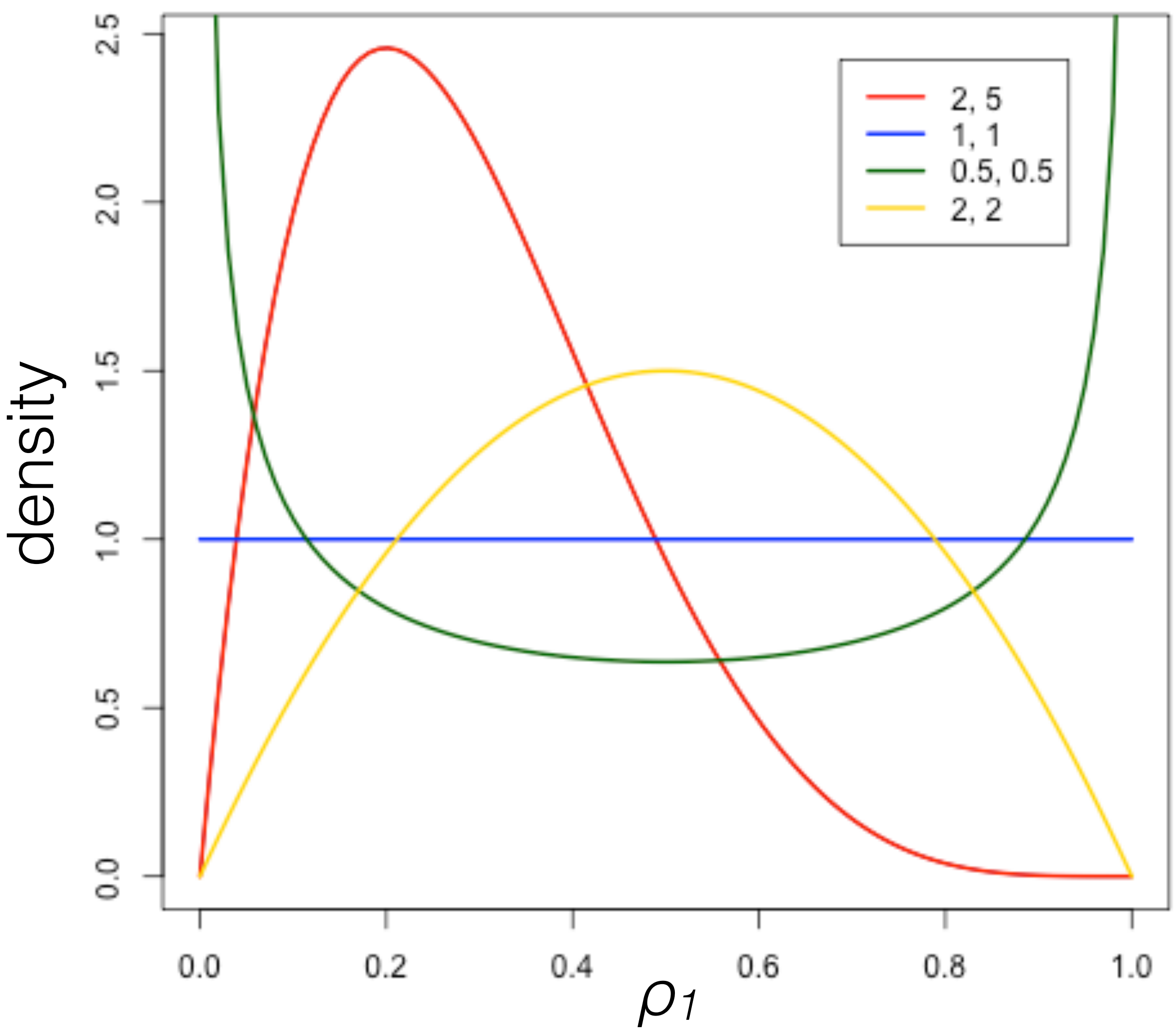
$a_1, a_2 > 0$

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

Beta distribution review

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

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



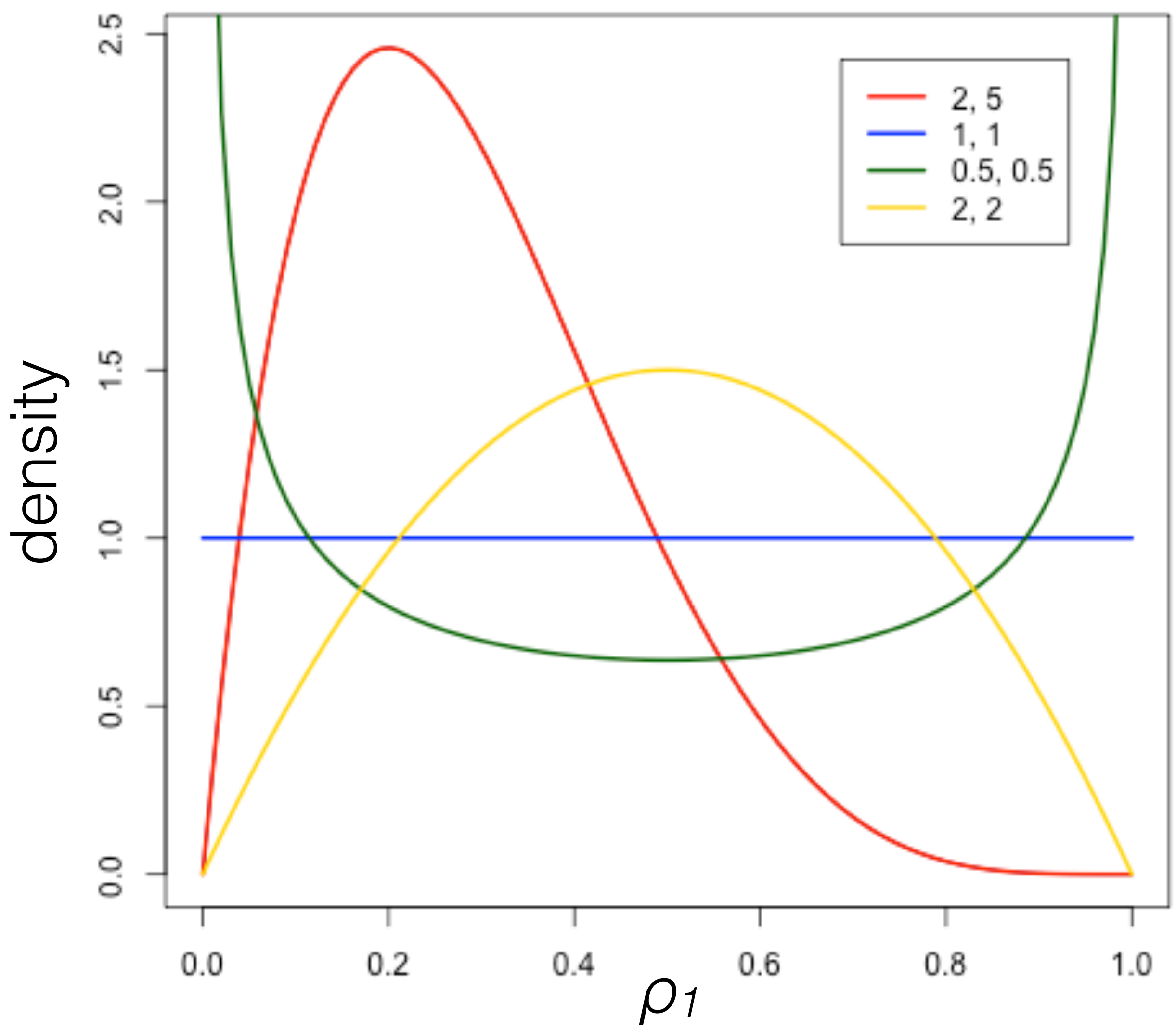
- Gamma function Γ
  - integer *m*: $\Gamma(m+1) = m!$
  - for *x* > 0: $\Gamma(x+1) = x\Gamma(x)$
- What happens?

Beta distribution review

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

$\rho_1 \in (0, 1)$

$a_1, a_2 > 0$



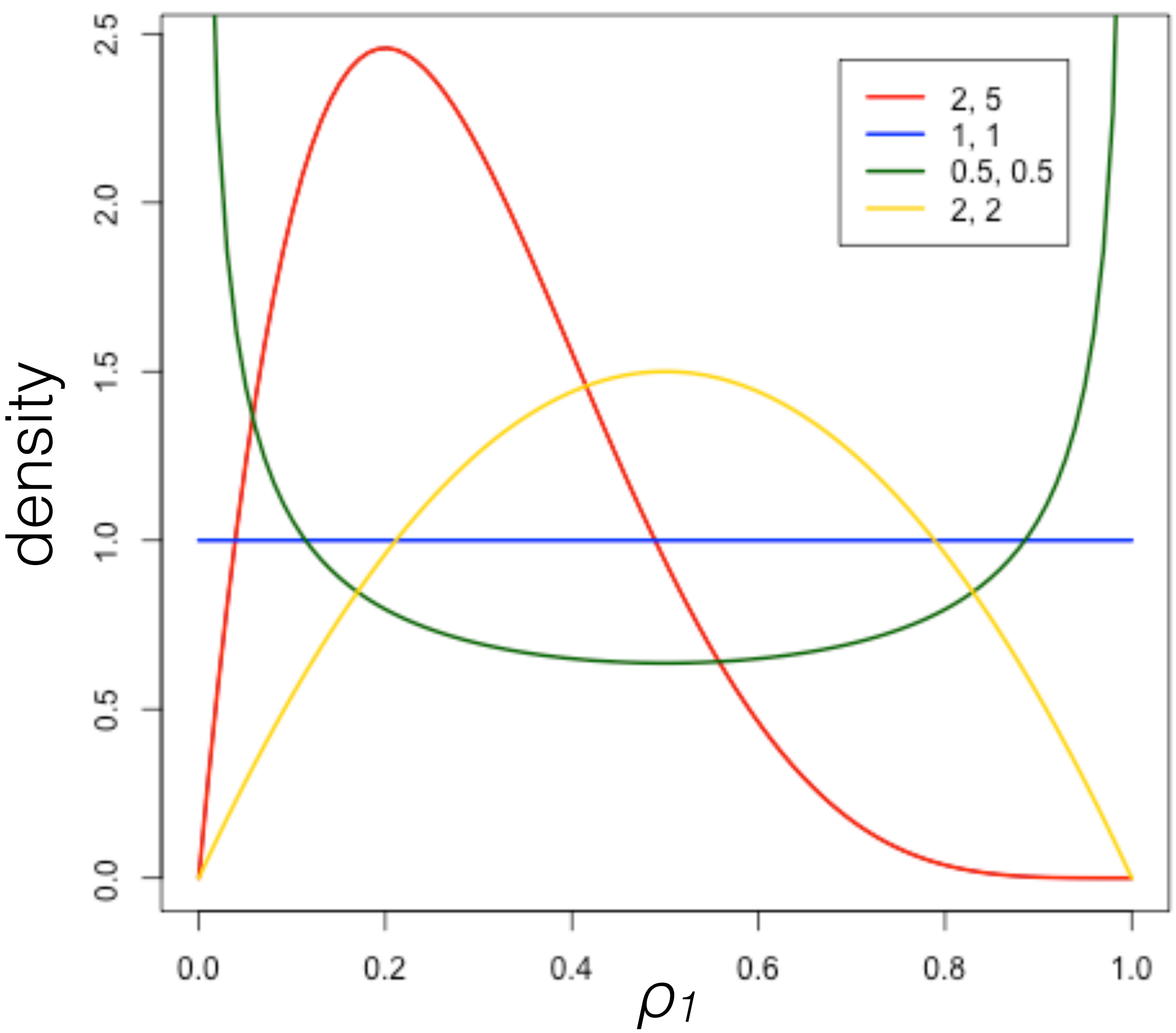
- Gamma function Γ
  - integer *m*: $\Gamma(m+1) = m!$
  - for *x* > 0: $\Gamma(x+1) = x\Gamma(x)$
- What happens?
  - $a = a_1 = a_2 \to 0$

Beta distribution review

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

$\rho_1 \in (0,1)$

$a_1, a_2 > 0$



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

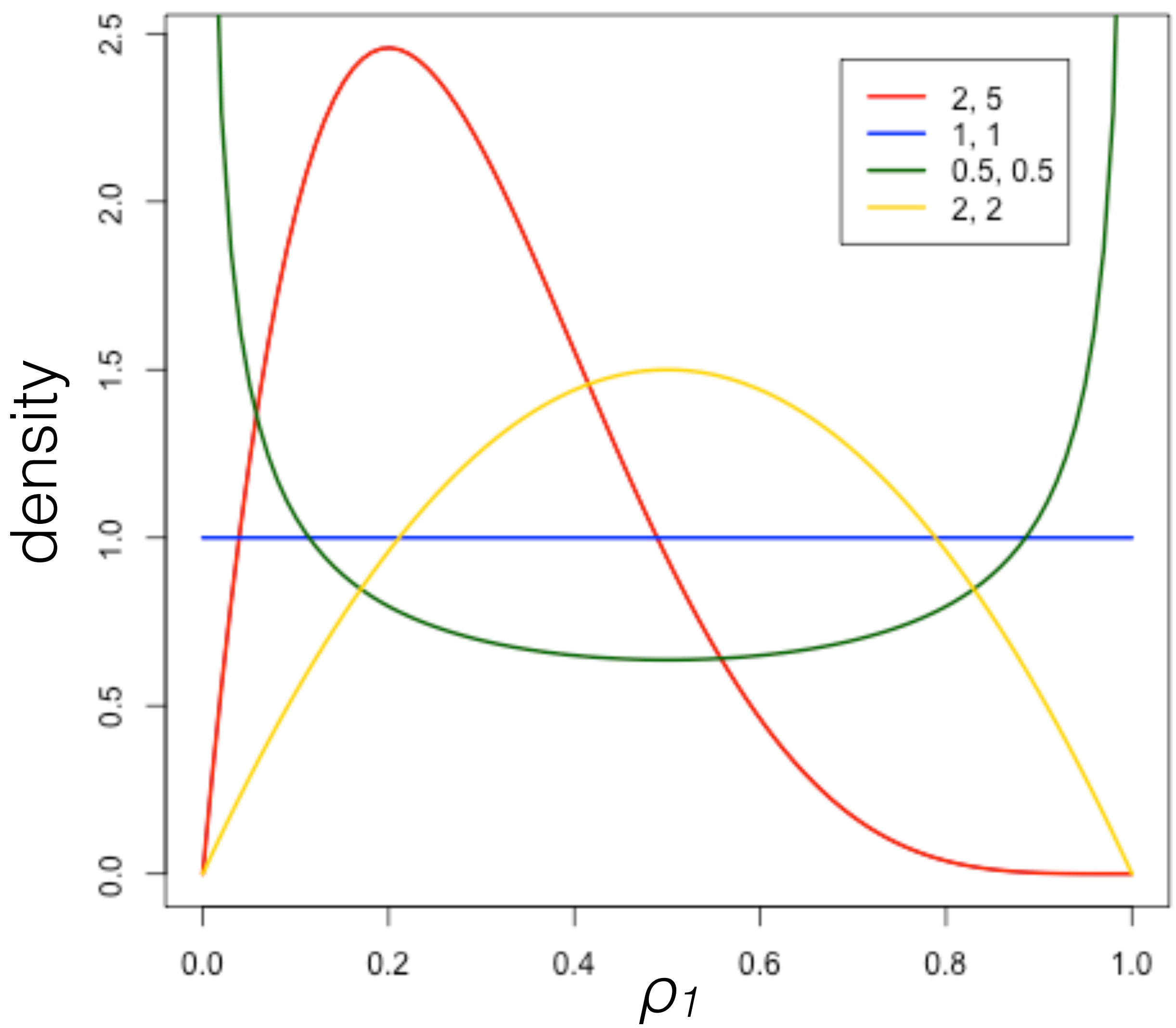
[demo]

Beta distribution review

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

$\rho_1 \in (0, 1)$

$a_1, a_2 > 0$



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

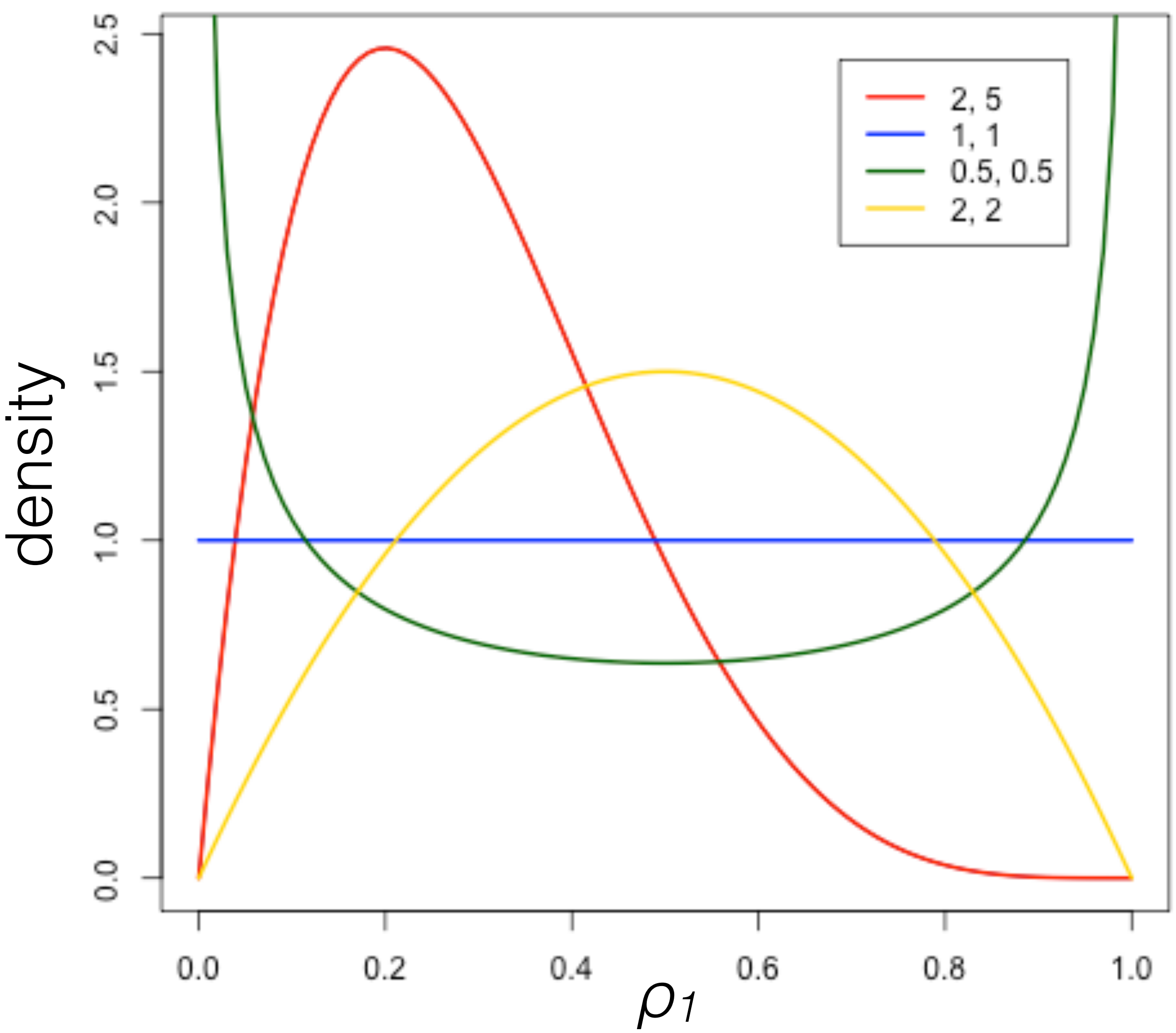
[demo]

Beta distribution review

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

$\rho_1 \in (0,1)$

$a_1, a_2 > 0$



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

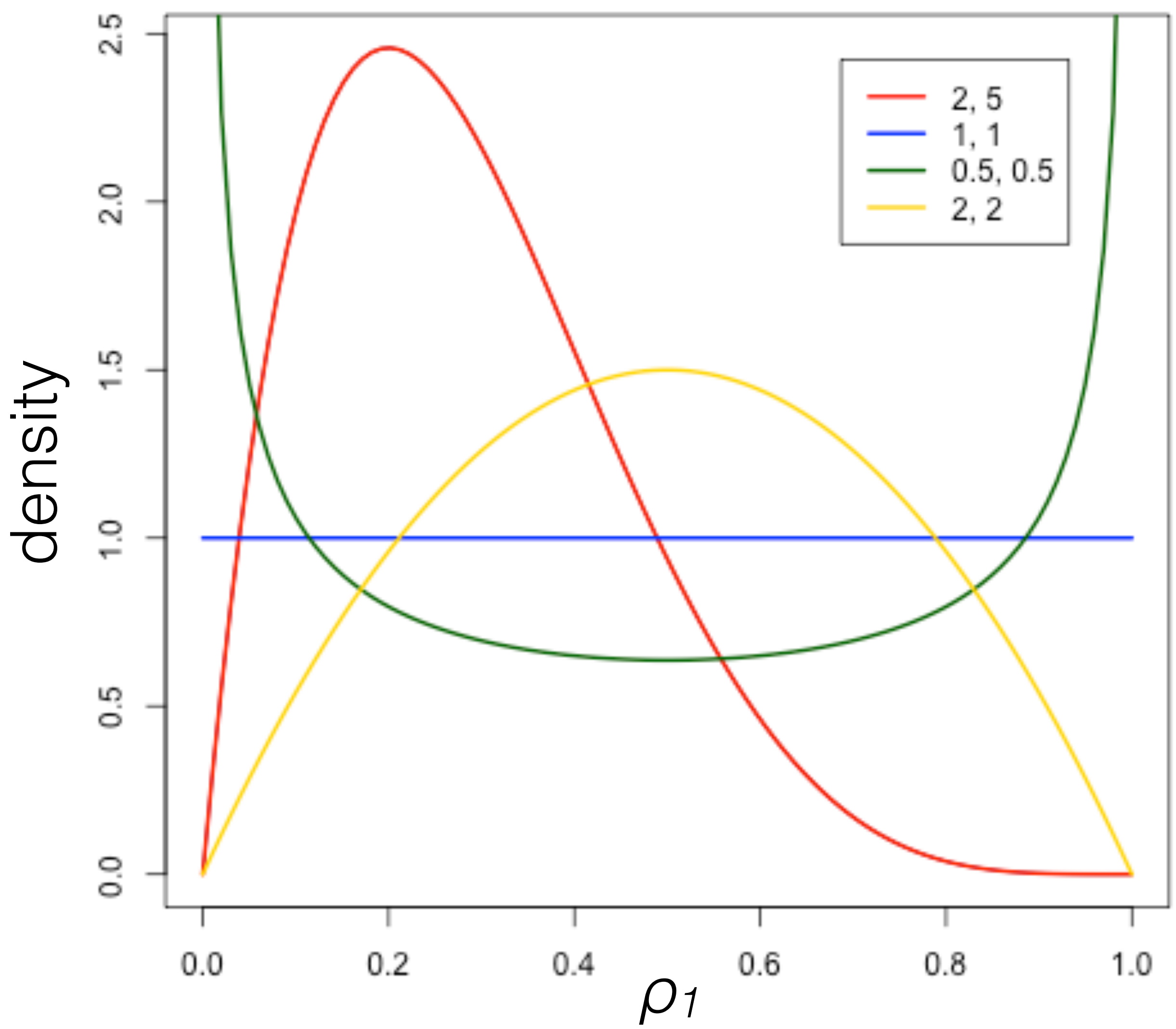
[demo]

Beta distribution review

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

$\rho_1 \in (0,1)$

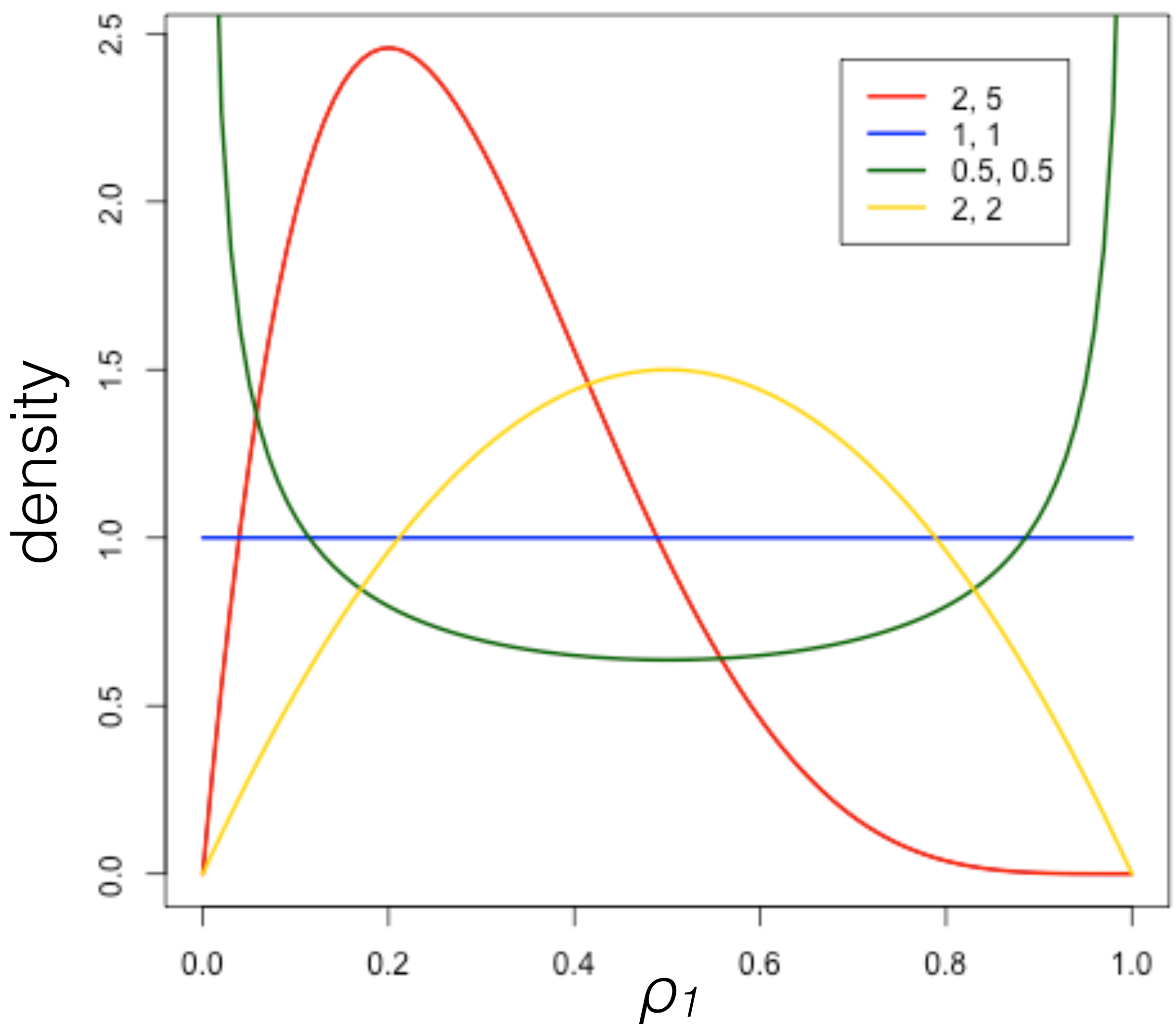
$a_1, a_2 > 0$



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

Beta distribution review

$$\text{Beta}(\rho_1|a_1,a_2) = \frac{\Gamma(a_1+a_2)}{\Gamma(a_1)\Gamma(a_2)}\rho_1^{a_1-1}(1-\rho_1)^{a_2-1} \qquad \begin{array}{c}\rho_1 \in (0,1)\\ a_1, a_2 > 0\end{array}$$



- Gamma function Γ
  - 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

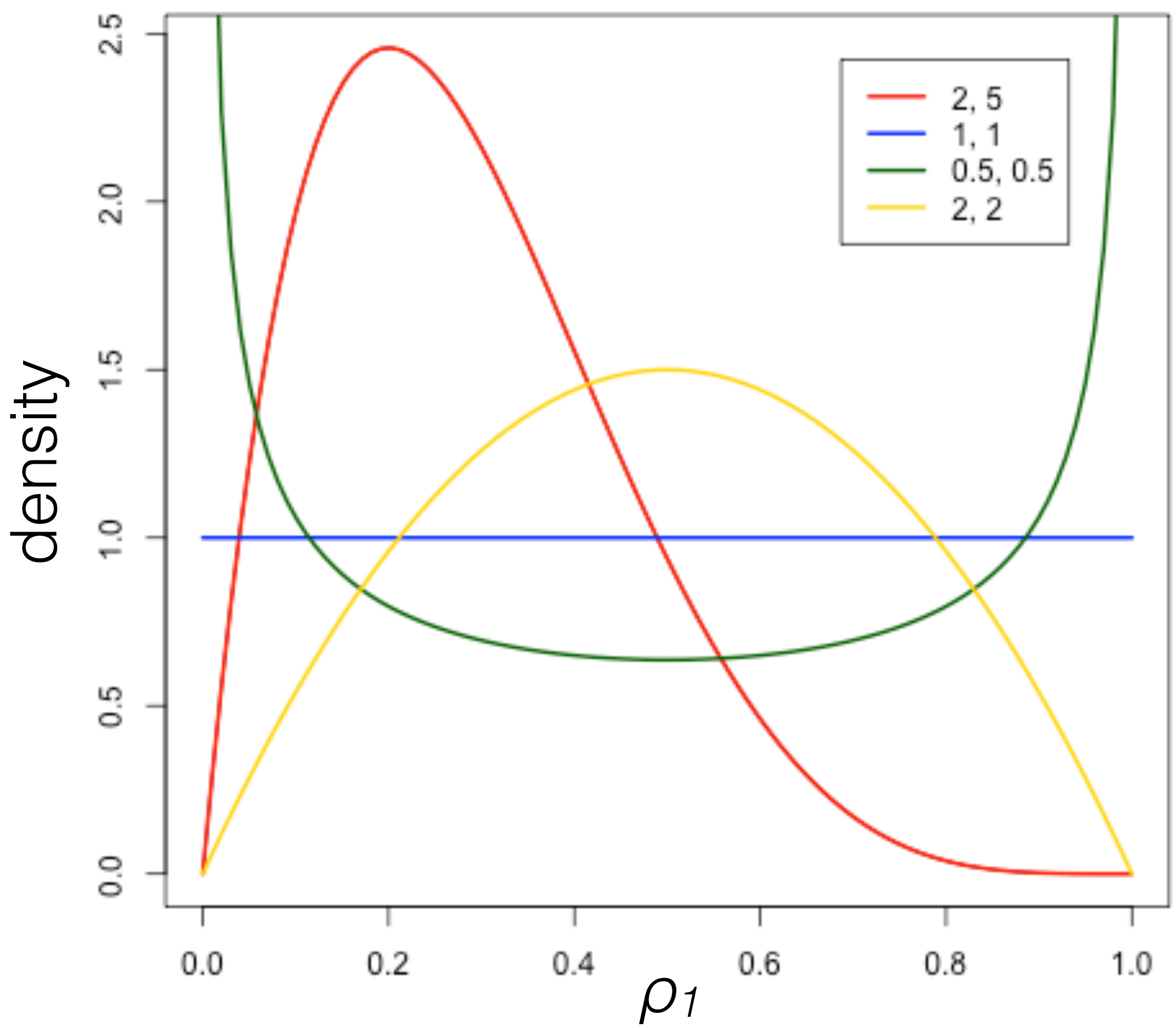
$$\rho_1 \sim \text{Beta}(a_1, a_2), z \sim \text{Cat}(\rho_1, \rho_2)$$

Beta distribution review

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

$\rho_1 \in (0,1)$

$a_1, a_2 > 0$



- Gamma function Γ
  - 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

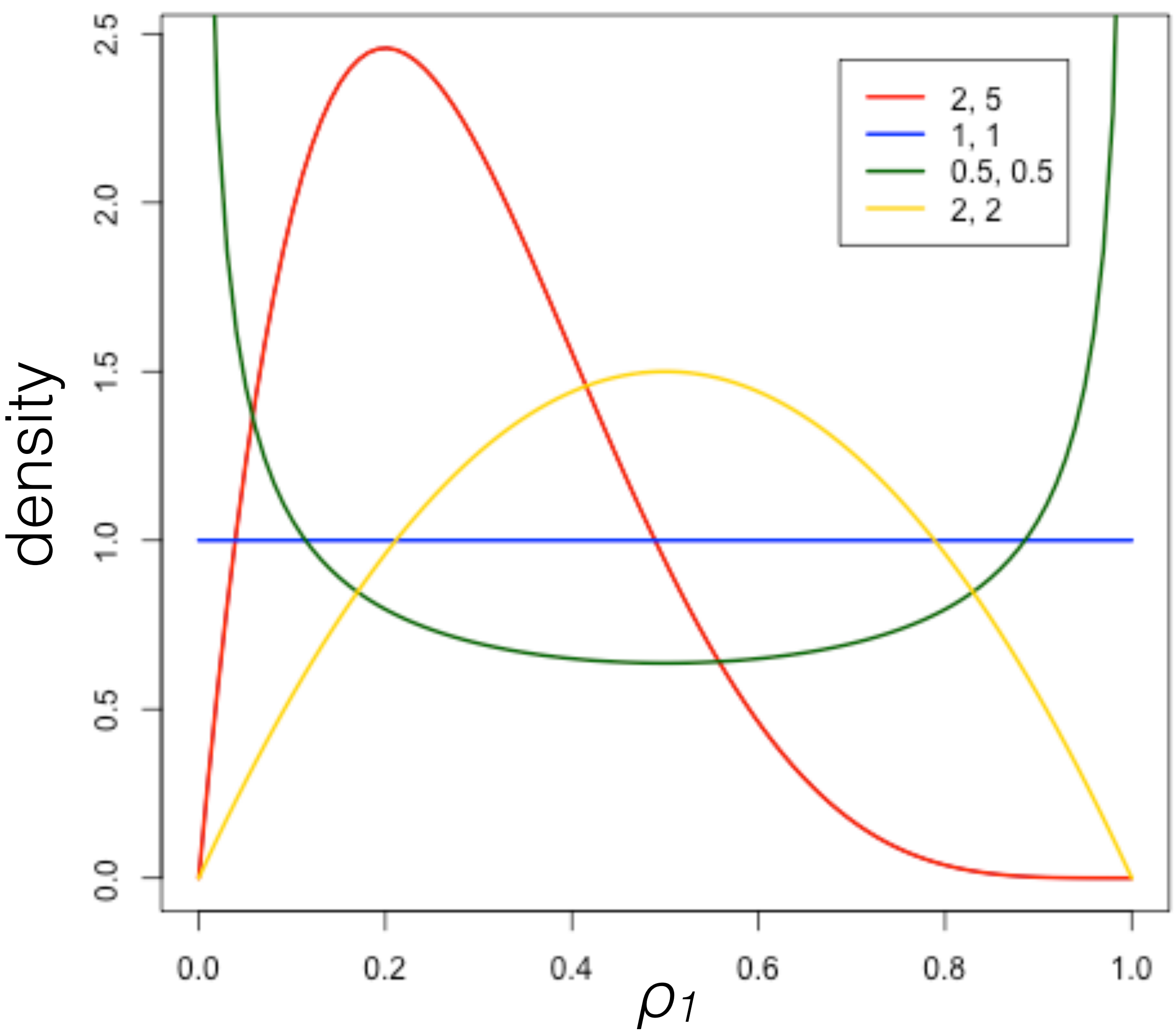
$\rho_1 \sim \mathrm{Beta}(a_1, a_2), z \sim \mathrm{Cat}(\rho_1, \rho_2)$

$p(\rho_1, z) \propto$

Beta distribution review

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

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



- Gamma function Γ
  - 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

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

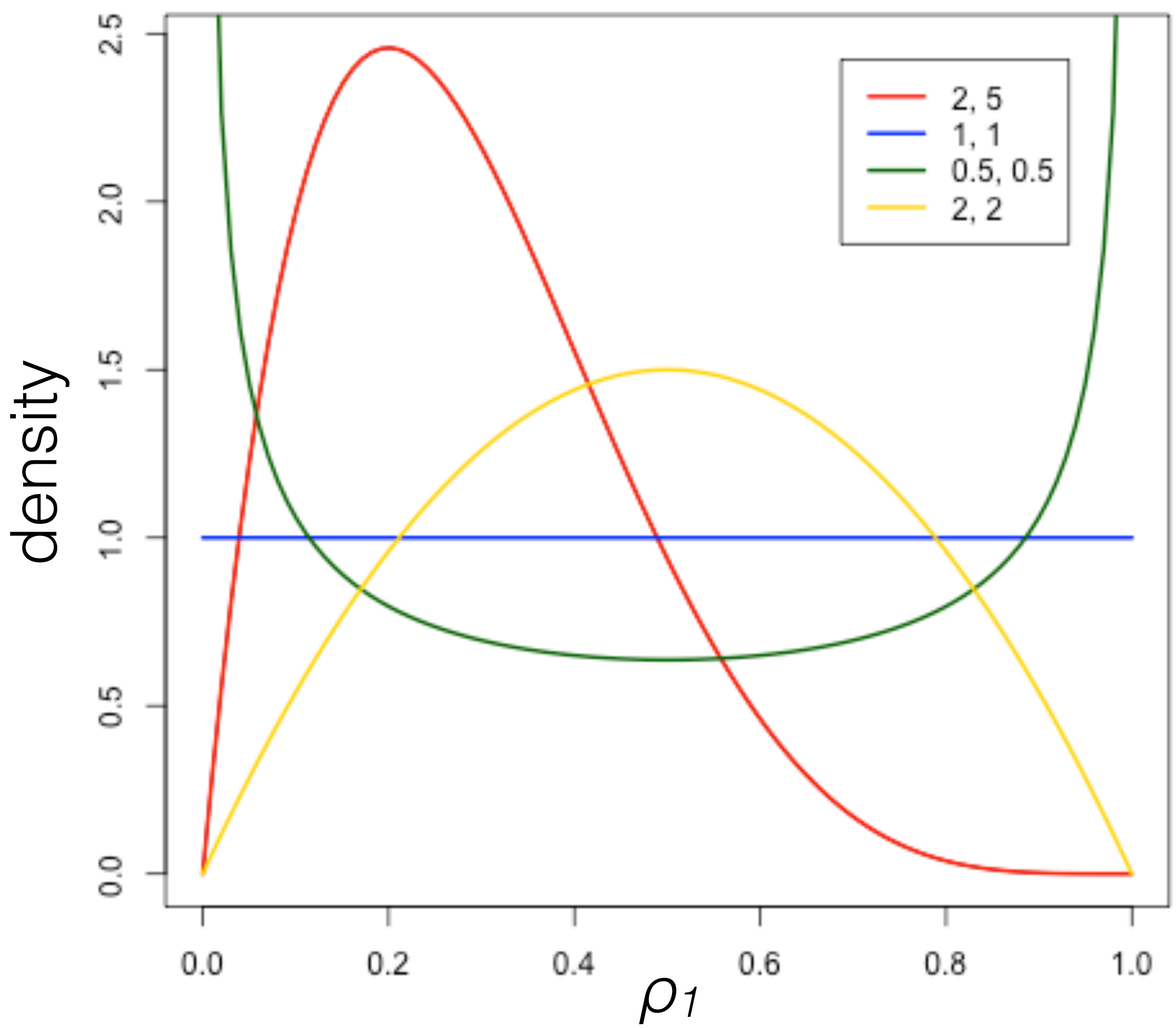
$$p(\rho_1, z) \propto \rho_1^{\mathbf{1}\{z=1\}}(1-\rho_1)^{\mathbf{1}\{z=2\}}$$

Beta distribution review

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

$\rho_1 \in (0,1)$

$a_1, a_2 > 0$



- Gamma function Γ
  - 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

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

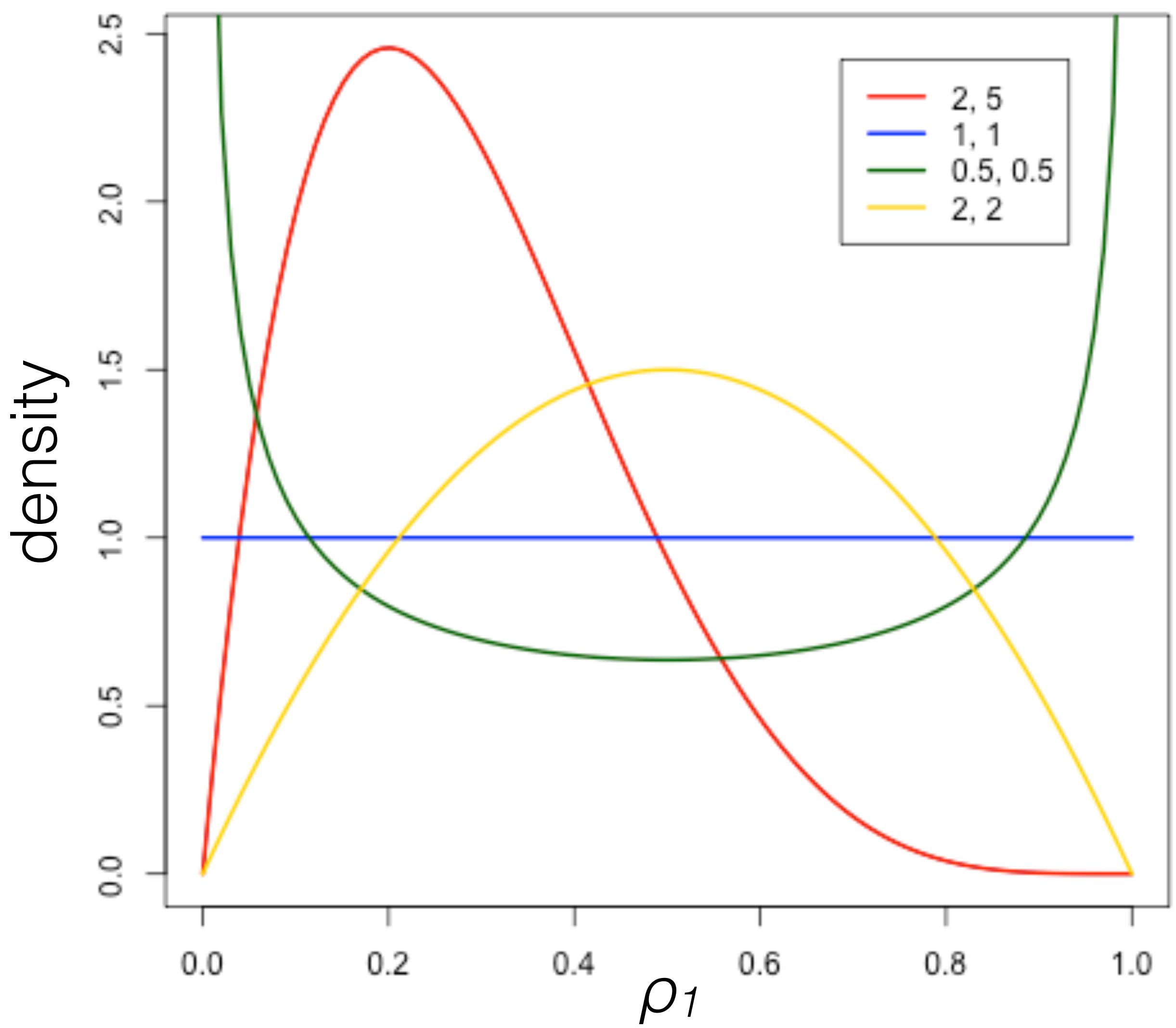
$$p(\rho_1, z) \propto \rho_1^{\mathbf{1}\{z=1\}}(1-\rho_1)^{\mathbf{1}\{z=2\}}\rho_1^{a_1-1}(1-\rho_1)^{a_2-1}$$

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 Γ
  - 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

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

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

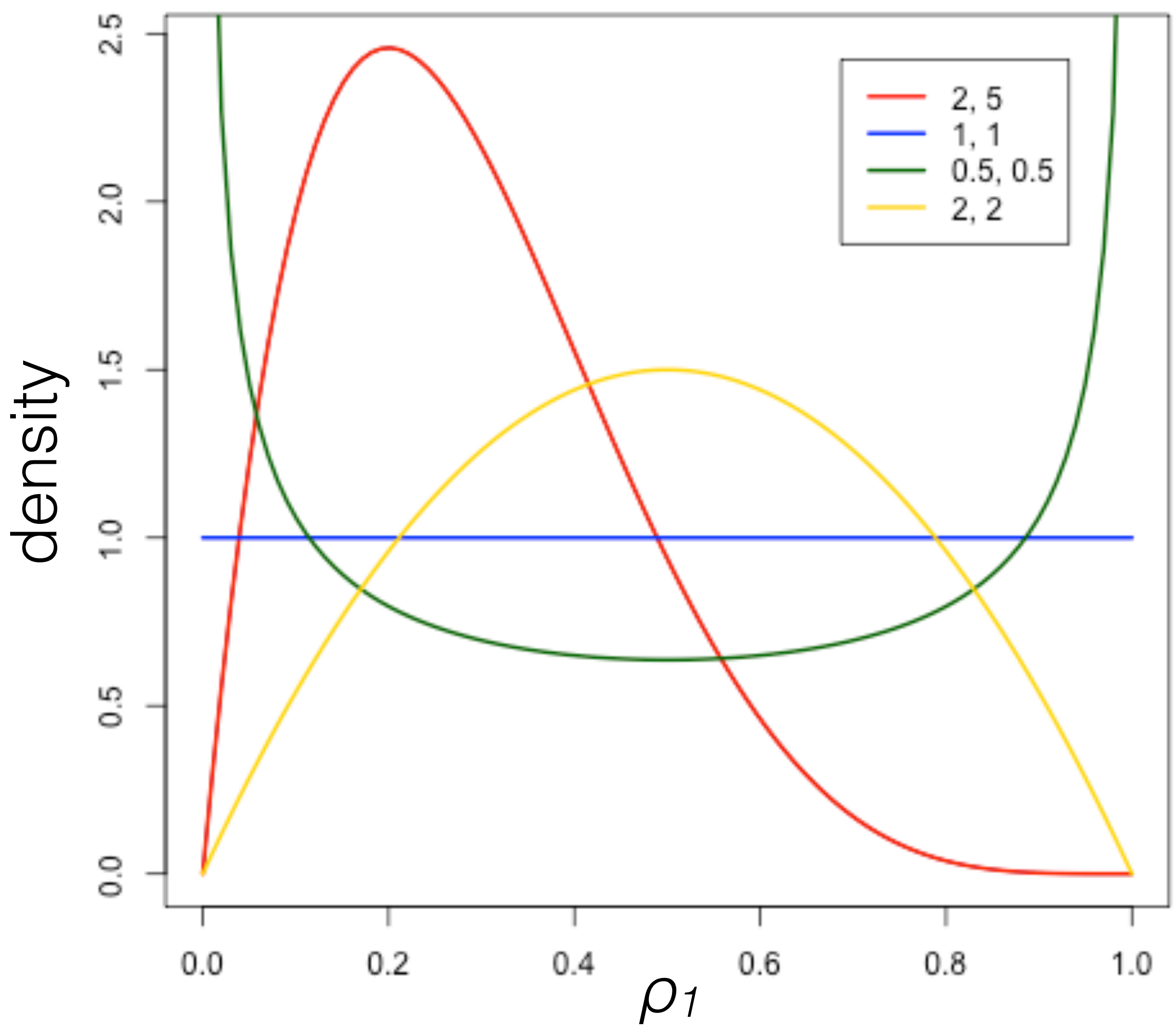
$$p(\rho_1|z) \propto$$

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 Γ
  - 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

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

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

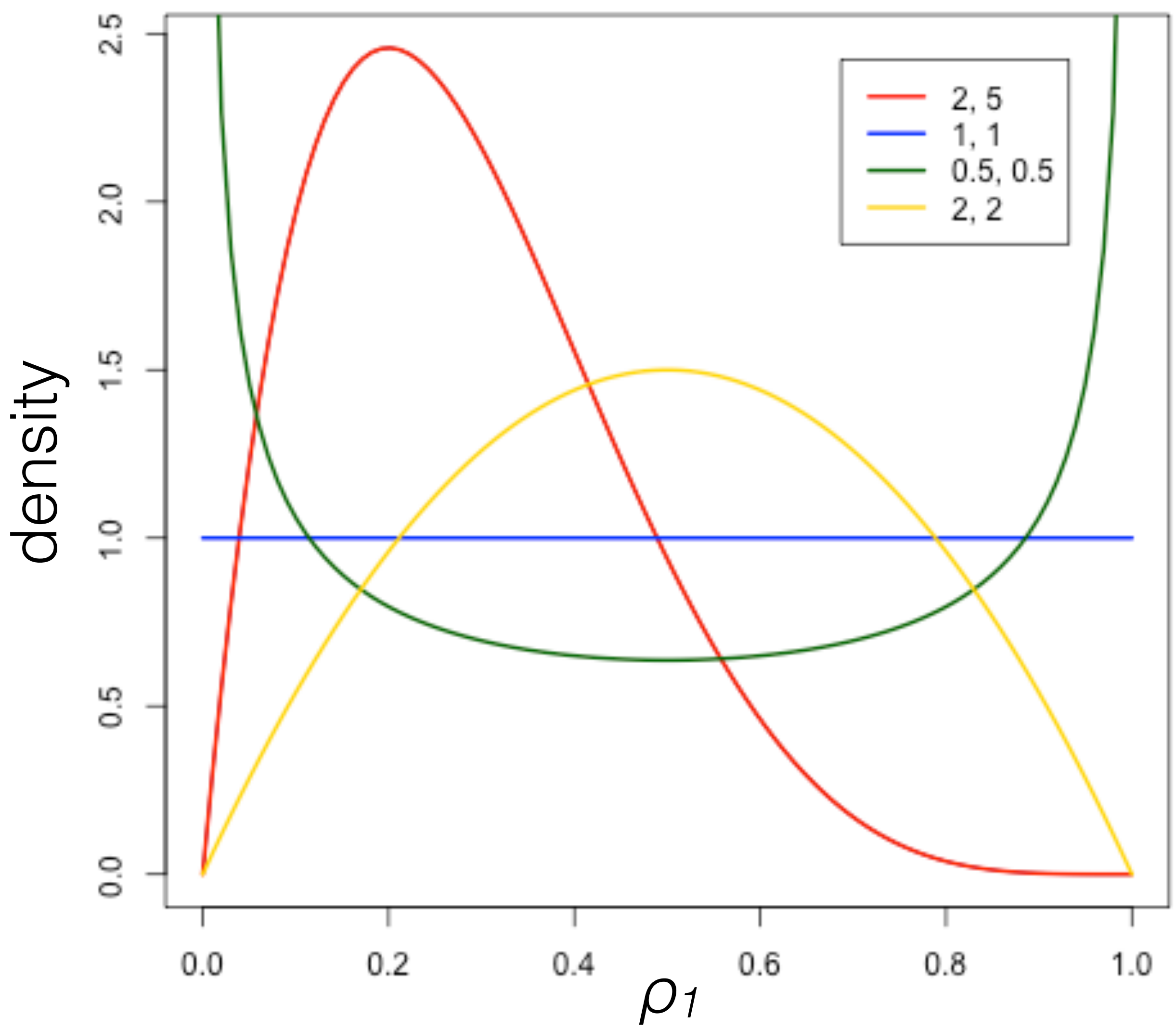
$$p(\rho_1 | z) \propto \rho_1^{a_1 + \mathbf{1}\{z=1\} - 1}(1 - \rho_1)^{a_2 + \mathbf{1}\{z=2\} - 1}$$

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 Γ
  - 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

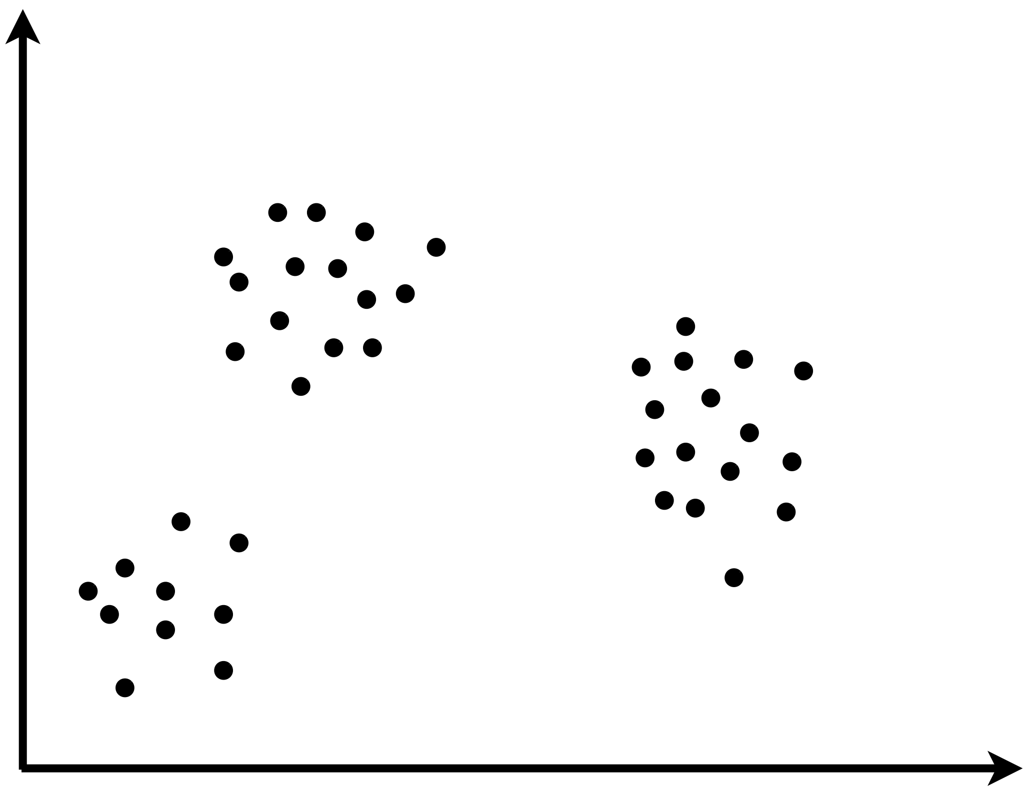
$$\rho_1 \sim \text{Beta}(a_1, a_2), z \sim \text{Cat}(\rho_1, \rho_2)$$

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

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

Generative model

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



- Finite Gaussian mixture model (*K* clusters)

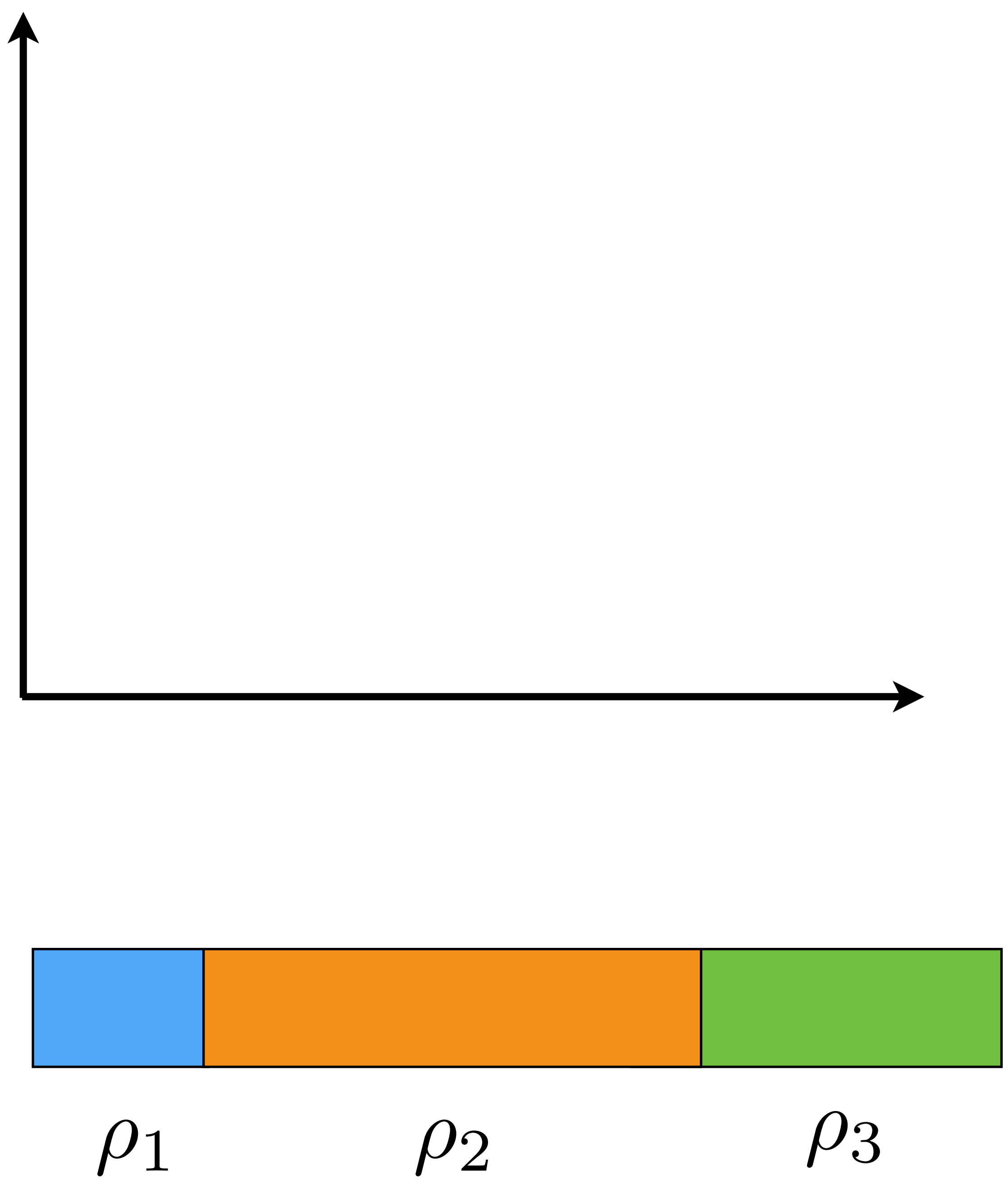
Generative model

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

- Finite Gaussian mixture model (*K* clusters)

Generative model

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

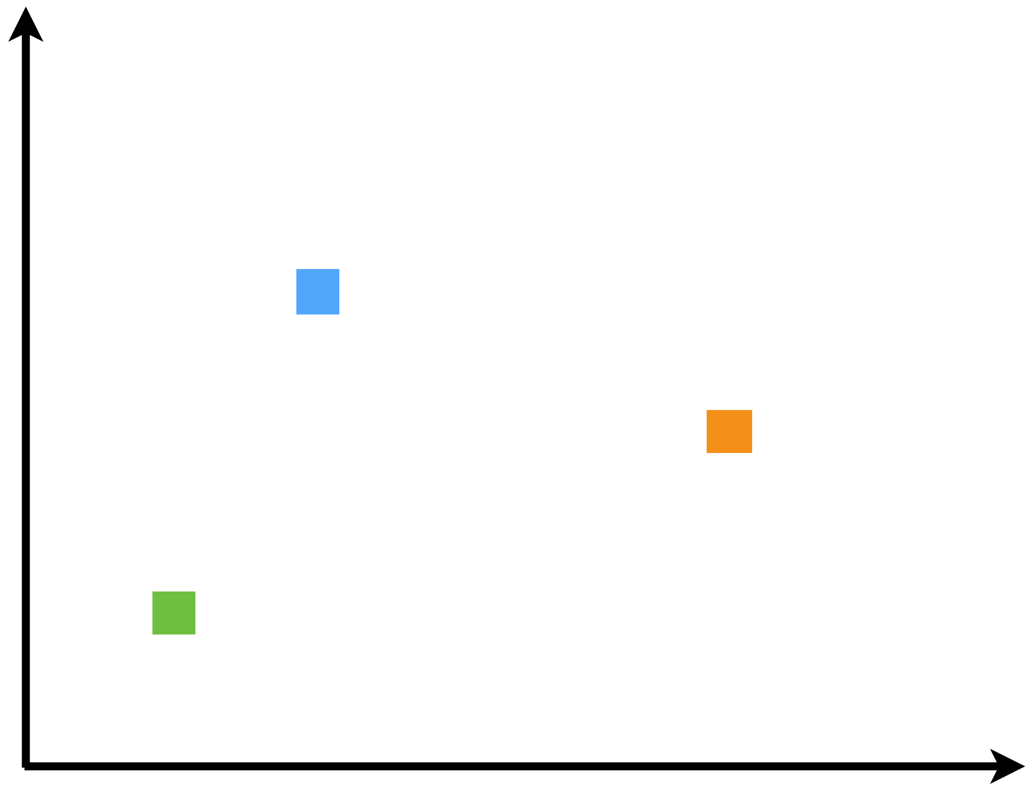


- Finite Gaussian mixture model (*K* clusters)

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

Generative model

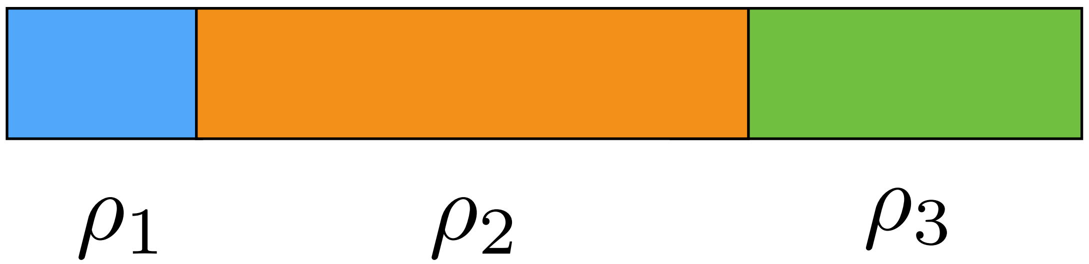
$$\mathbb{P}(\text{parameters}|\text{data}) \propto \mathbb{P}(\text{data}|\text{parameters})\mathbb{P}(\text{parameters})$$



- Finite Gaussian mixture model (*K* clusters)

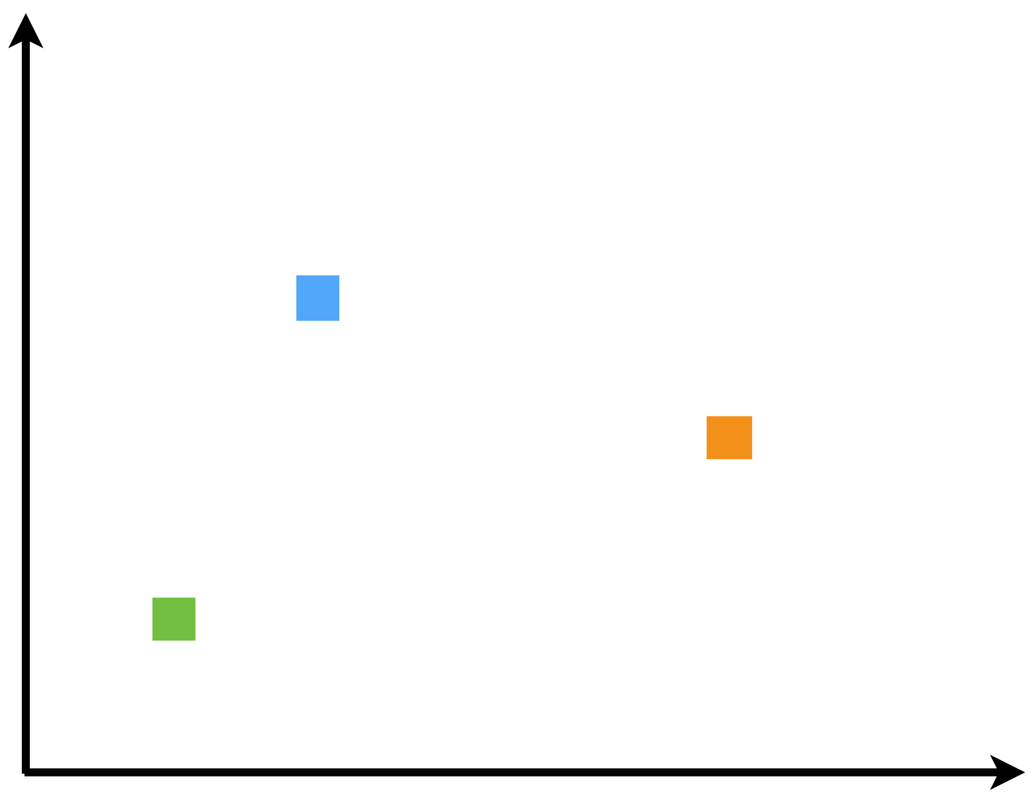
$$\rho_{1:K} \sim \text{Dirichlet}(a_{1:K})$$

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



Generative model

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

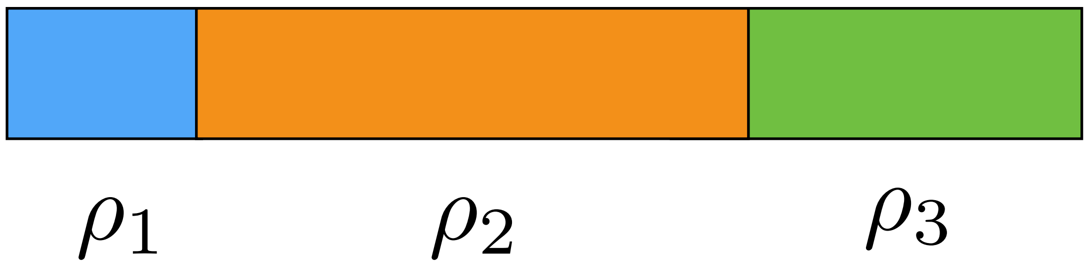


- Finite Gaussian mixture model (*K* clusters)

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

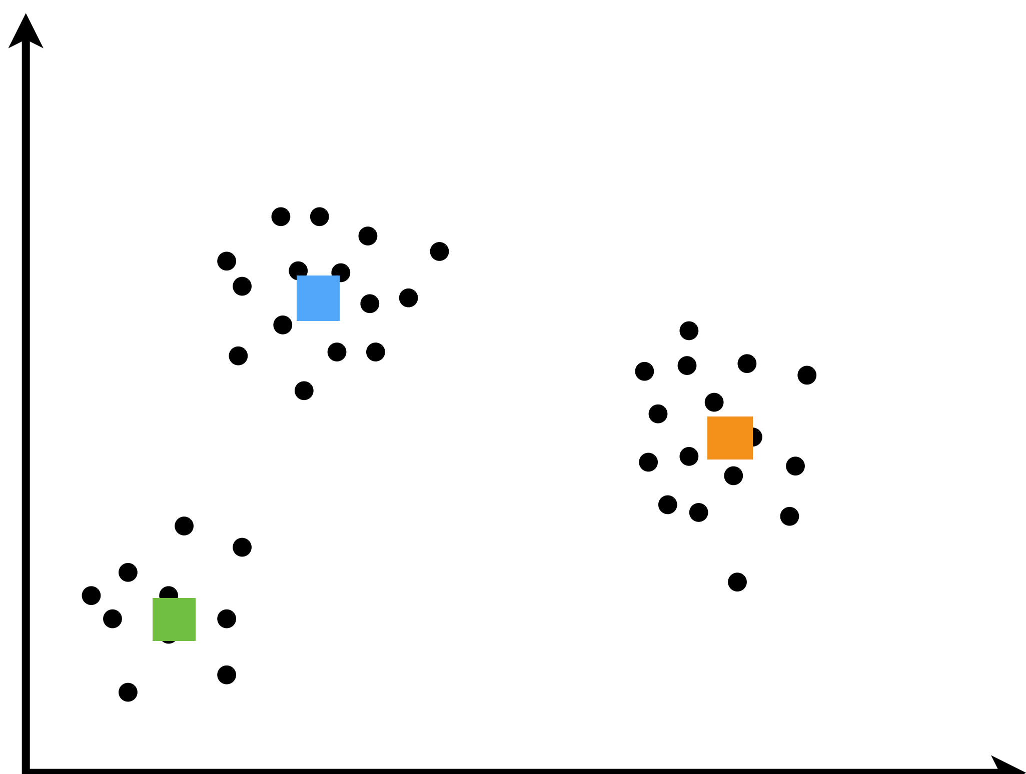
$$\mu_k \overset{iid}{\sim} \mathcal{N}(\mu_0, \Sigma_0)$$

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



Generative model

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



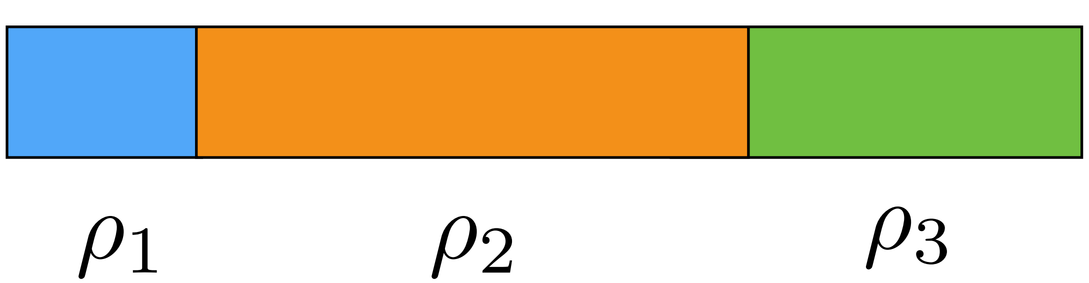
- Finite Gaussian mixture model (*K* clusters)

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

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

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

$$x_n \overset{indep}{\sim} \mathcal{N}(\mu_{z_n}, \Sigma)$$



Dirichlet distribution review

$$\text{Dirichlet}(\rho_{1:K}|a_{1:K}) = \frac{\Gamma(\sum_{k=1}^{K} a_k)}{\prod_{k=1}^{K} \Gamma(a_k)} \prod_{k=1}^{K} \rho_k^{a_k - 1} \qquad a_k > 0$$

Dirichlet distribution review

$$a_k > 0$$

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

$$\sum_k \rho_k = 1$$

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

Dirichlet distribution review

$$a_k > 0$$

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

$$\sum_k \rho_k = 1$$

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

- What happens?

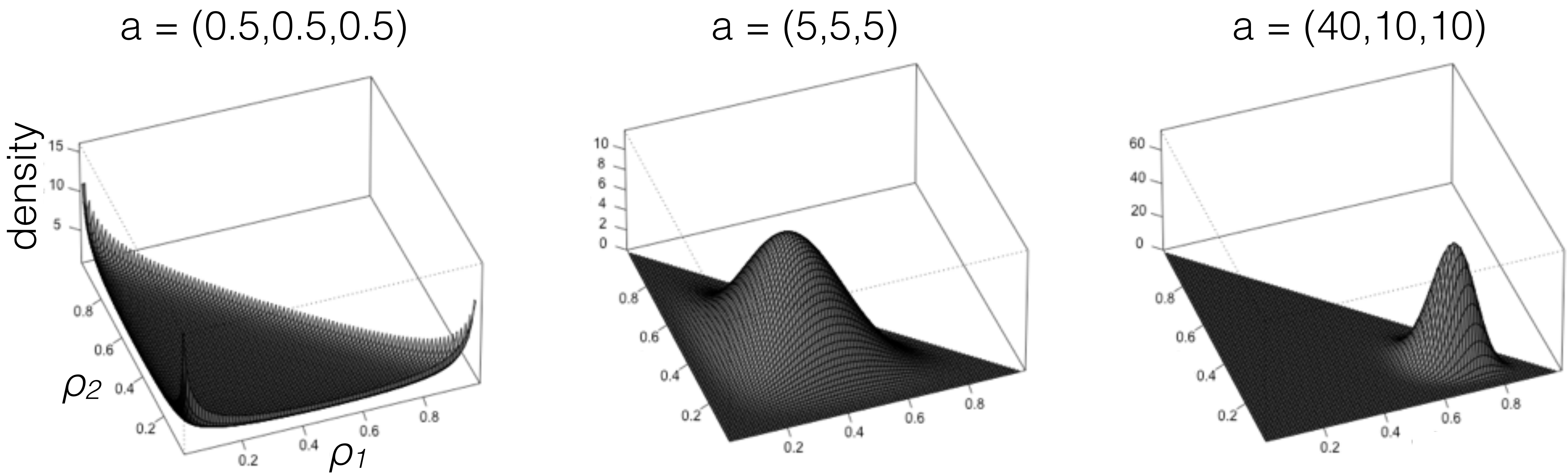
Dirichlet distribution review

$$a_k > 0$$

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

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

$$\sum_k \rho_k = 1$$

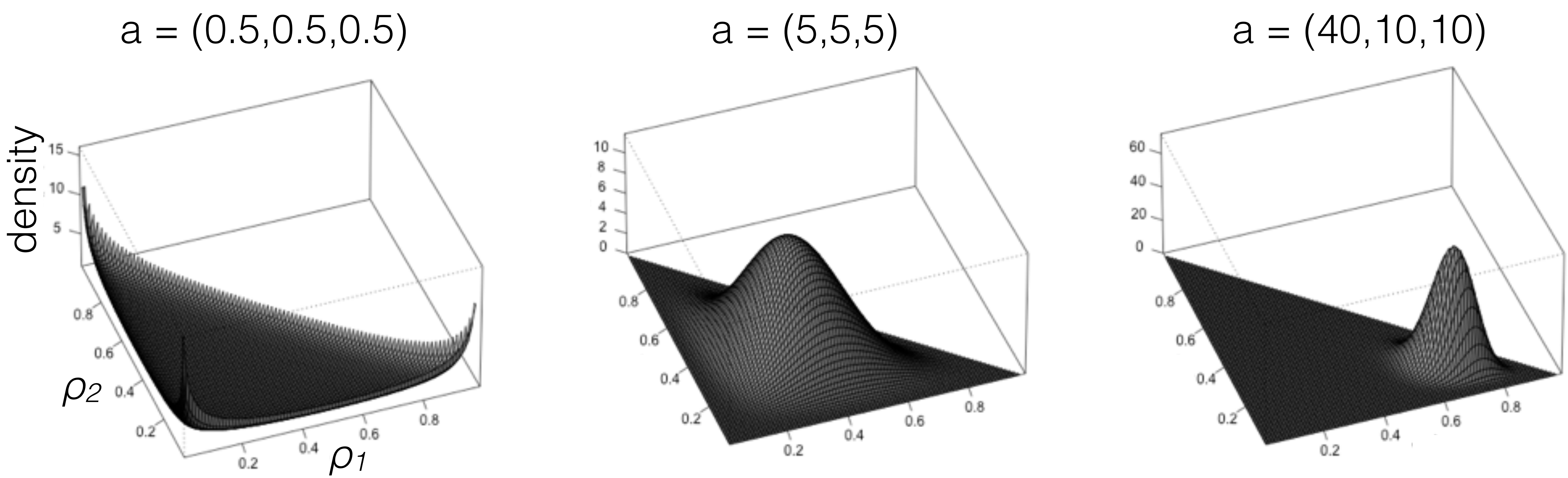


- What happens?

Dirichlet distribution review

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

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



- What happens? $a = a_k = 1$

Dirichlet distribution review

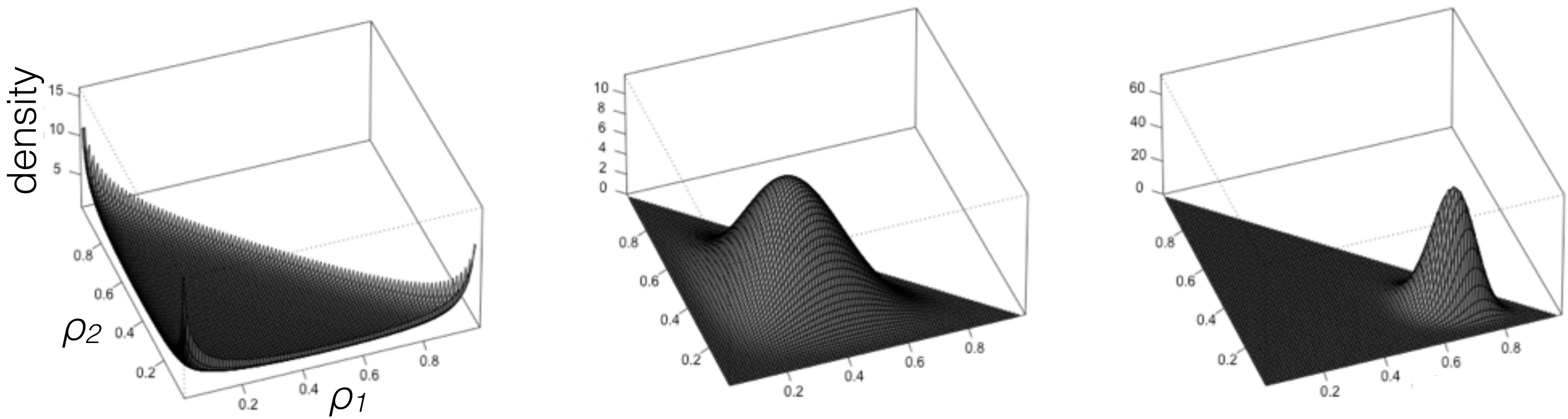
$$a_k > 0$$

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

$$\sum_k \rho_k = 1$$

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

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



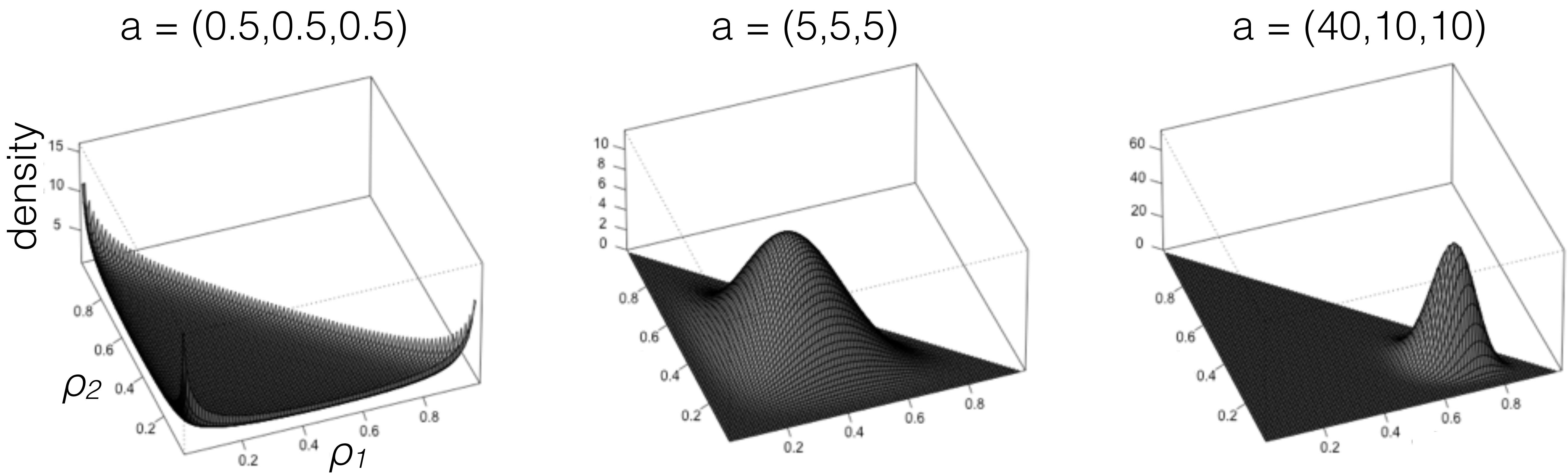
- What happens? $a = a_k = 1$

[demo]

Dirichlet distribution review

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

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



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

[demo]

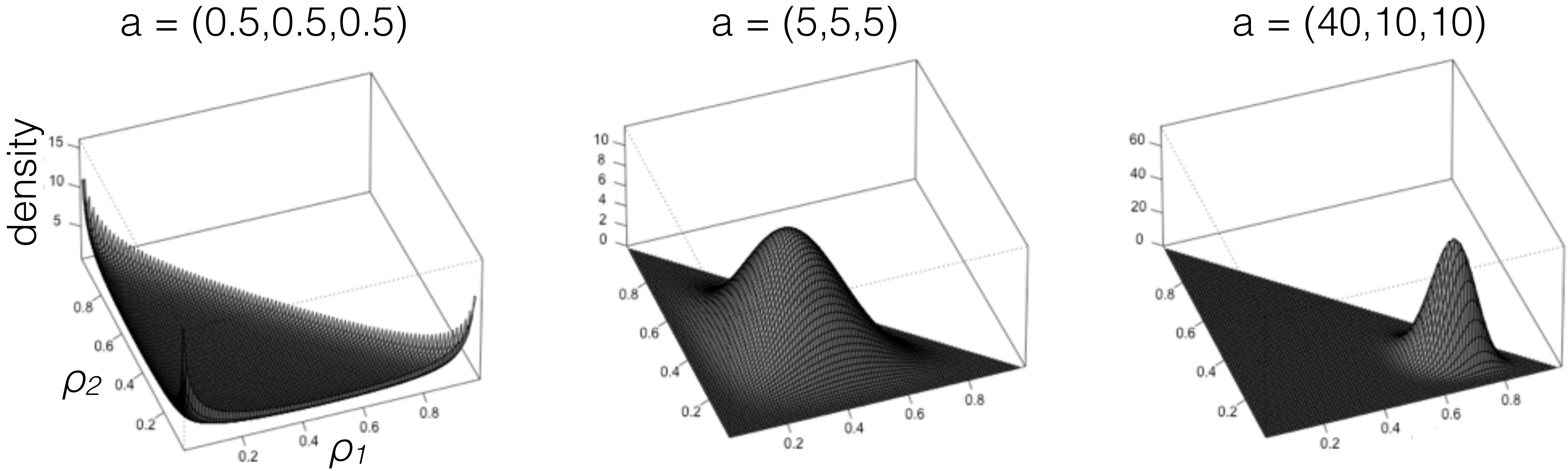
Dirichlet distribution review

$$a_k > 0$$

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

$$\sum_k \rho_k = 1$$

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



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

[demo]

Dirichlet distribution review

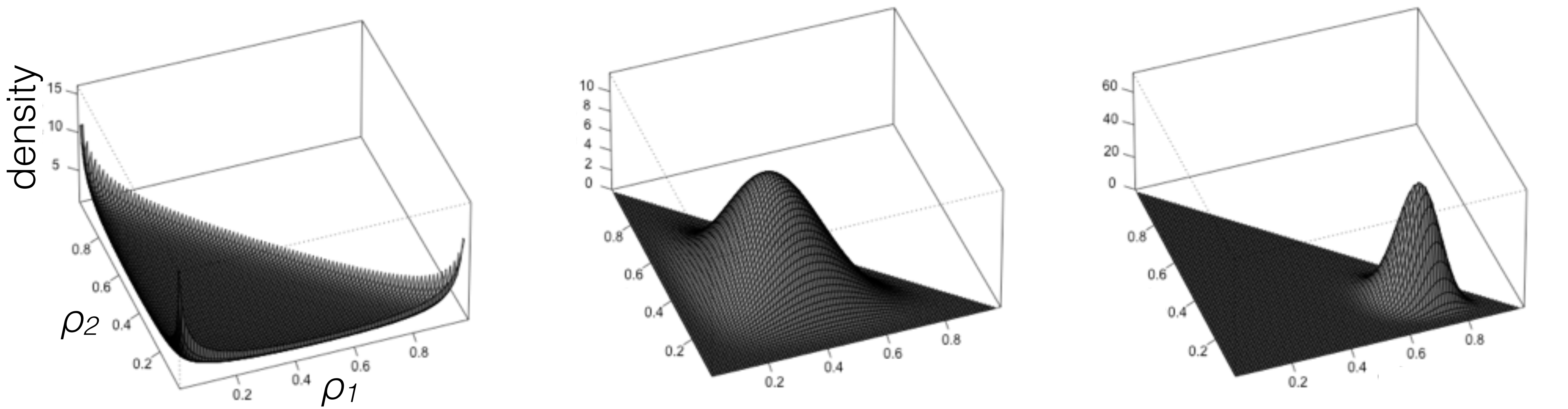
$$\text{Dirichlet}(\rho_{1:K}|a_{1:K}) = \frac{\Gamma(\sum_{k=1}^K a_k)}{\prod_{k=1}^K \Gamma(a_k)} \prod_{k=1}^K \rho_k^{a_k - 1}$$

$$a_k > 0$$

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

$$\sum_k \rho_k = 1$$

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

[demo]

Dirichlet distribution review

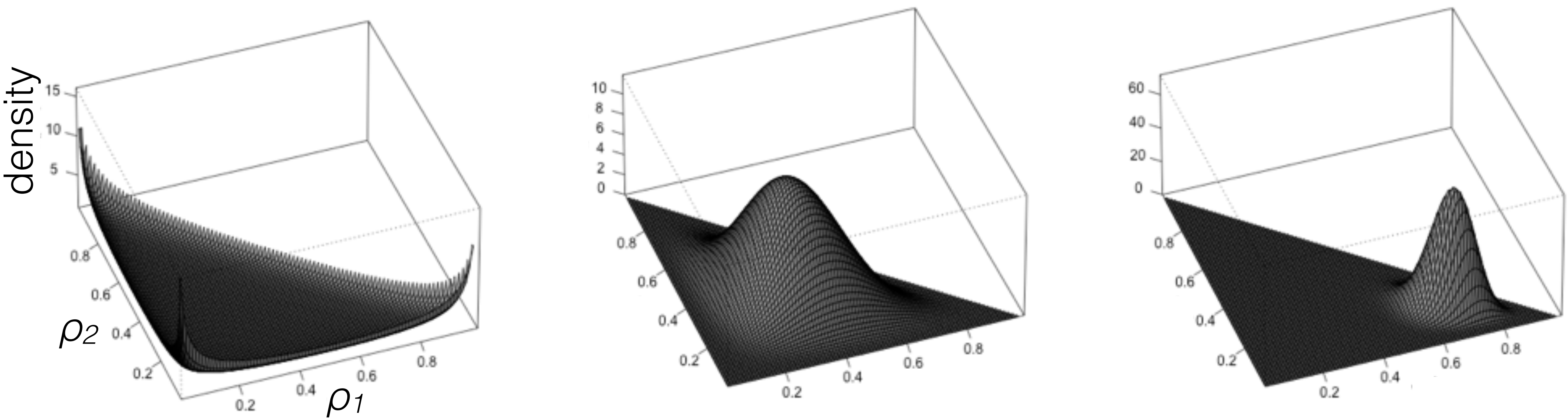
$$a_k > 0$$

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

$$\sum_k \rho_k = 1$$

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

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



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

  [demo]
- Dirichlet is conjugate to Categorical

  $\rho_{1:K} \sim \mathrm{Dirichlet}(a_{1:K}), z \sim \mathrm{Cat}(\rho_{1:K})$

Dirichlet distribution review

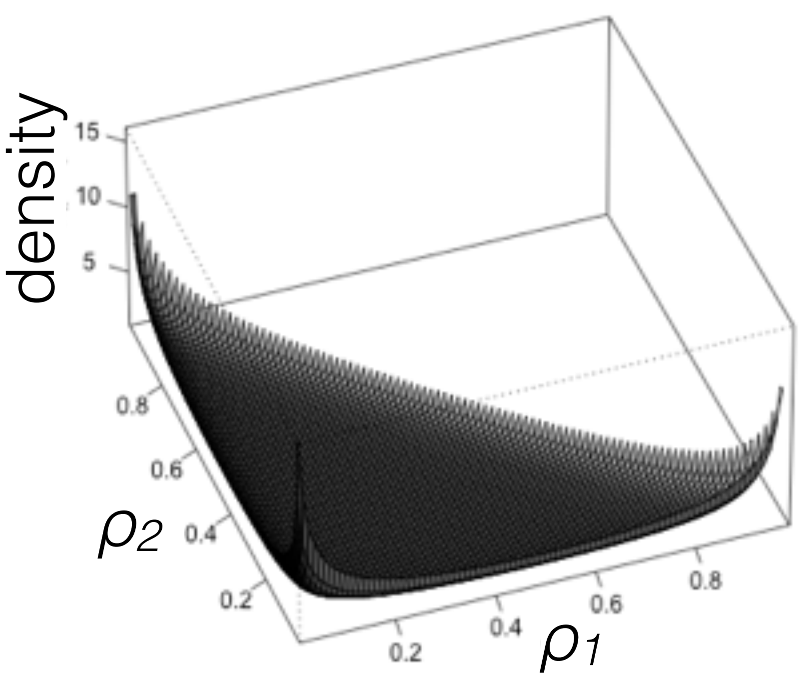
$a_k > 0$

$\rho_k \in (0, 1)$

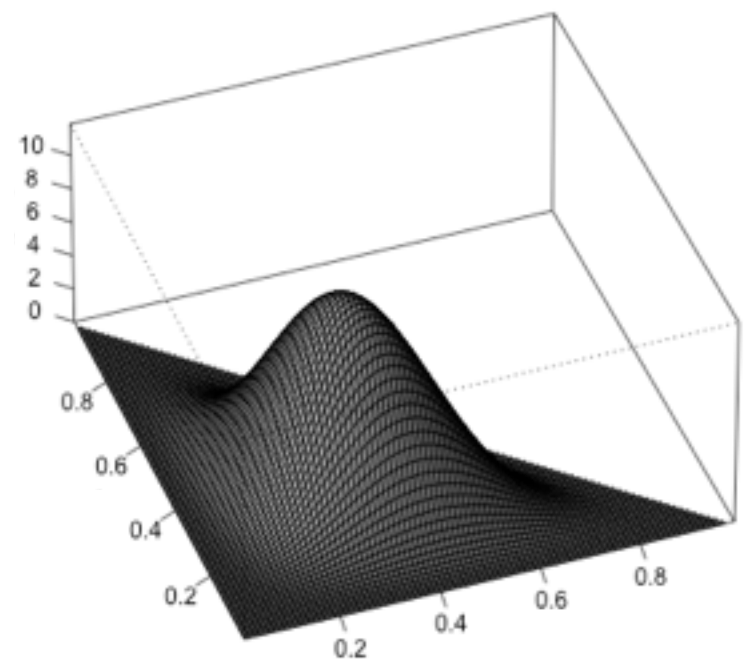
$\sum_k \rho_k = 1$

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

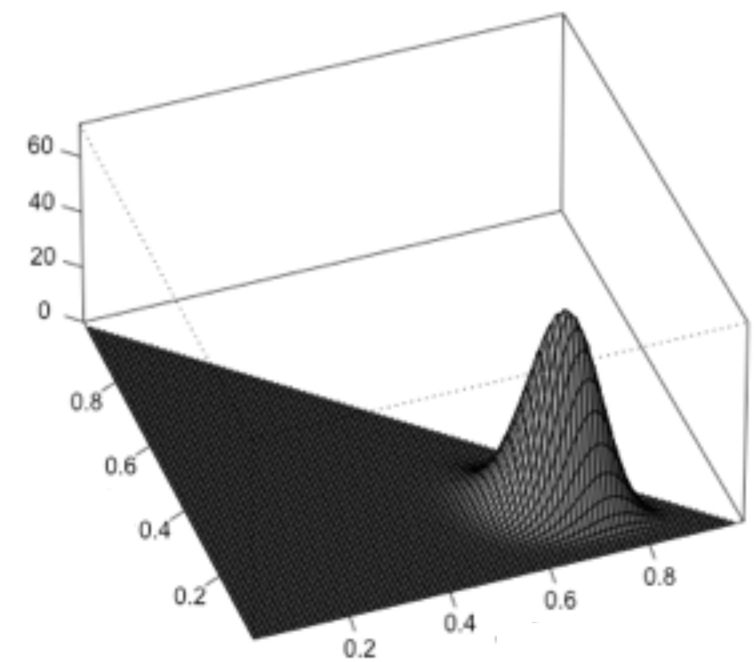
a = (0.5,0.5,0.5)



a = (5,5,5)



a = (40,10,10)



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

[demo]

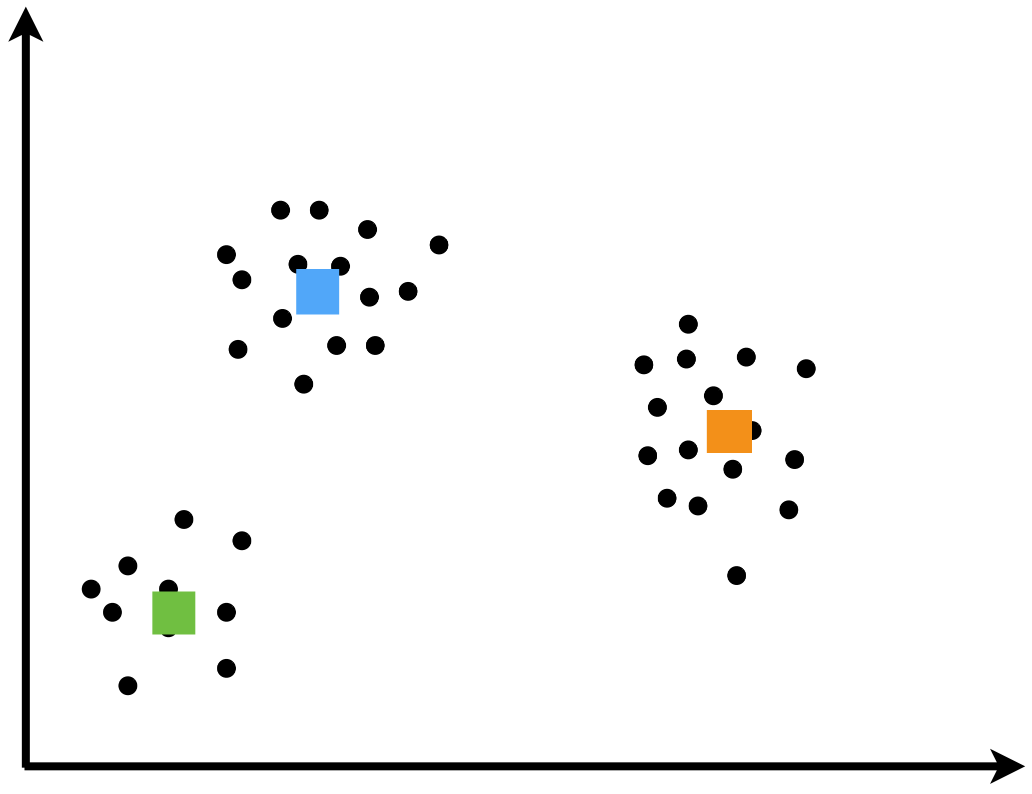
- Dirichlet is conjugate to Categorical

$$\rho_{1:K} \sim \mathrm{Dirichlet}(a_{1:K}), z \sim \mathrm{Cat}(\rho_{1:K})$$

$$\rho_{1:K} | z \overset{d}{=} \mathrm{Dirichlet}(a'_{1:K}), a'_k = a_k + \mathbf{1}\{z = k\}$$

What if $K > N$?

What if $K > N$?



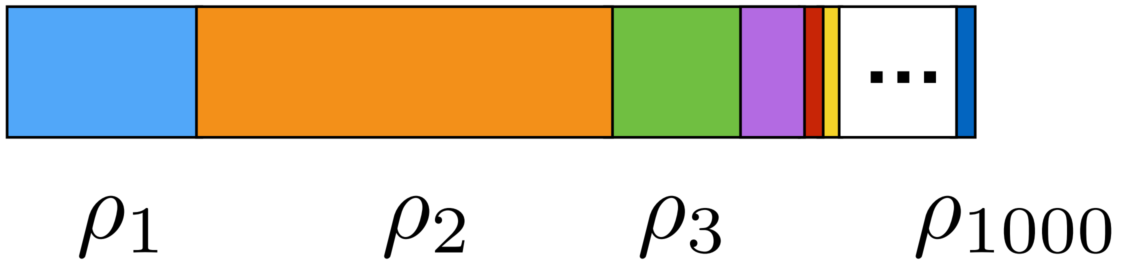
What if $K > N$?

What if $K > N$?

ρ_1 ρ_2 ρ_3 ρ_{1000}

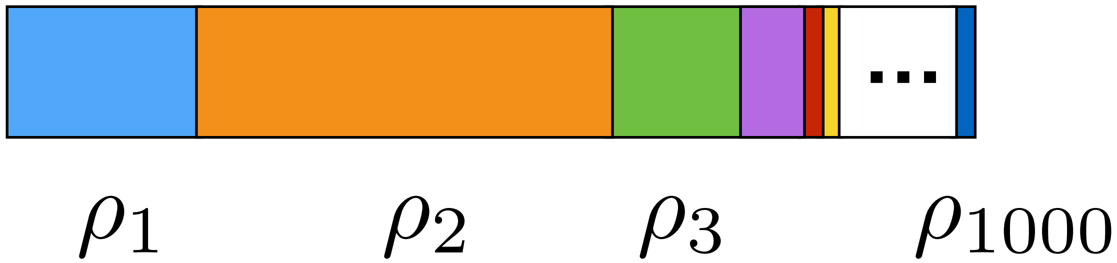
What if $K > N$?

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



What if $K > N$?

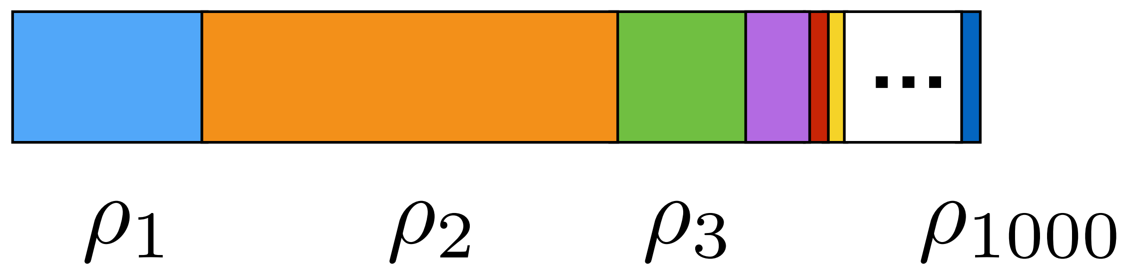
- e.g. species sampling, topic modeling, groups on a social network, etc.



- Components: number of latent groups

What if $K > N$?

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

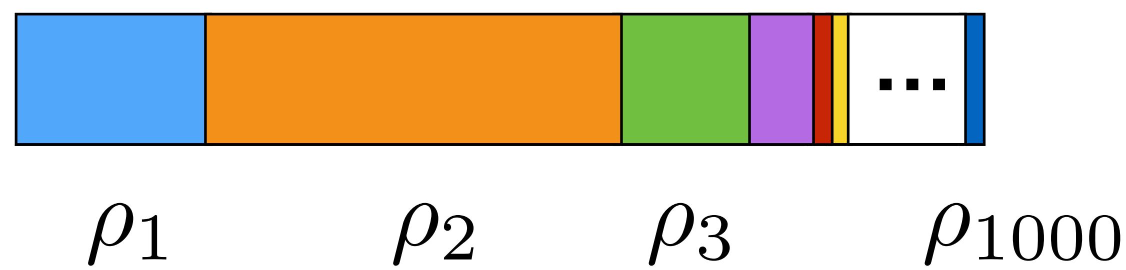


- Components: number of latent groups

- Clusters: number of components represented in the data

What if $K > N$?

- 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.

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

- Components: number of latent groups

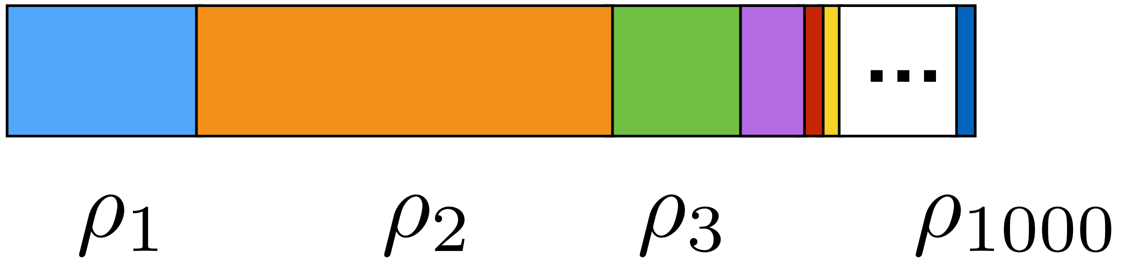
- Clusters: number of components represented in the data

- [demo 1, demo 2]

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

What if $K > N$?

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

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

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

$$\Leftrightarrow \rho_1 \stackrel{d}{=} \mathrm{Beta}(a_1, \sum_{k=1}^{K} a_k - a_1)$$

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

$$\Leftrightarrow \rho_1 \stackrel{d}{=} \mathrm{Beta}(a_1, \sum_{k=1}^{K} a_k - a_1)$$

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

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

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

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

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

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

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

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

- "Stick breaking"

Choosing $K = \infty$

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

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

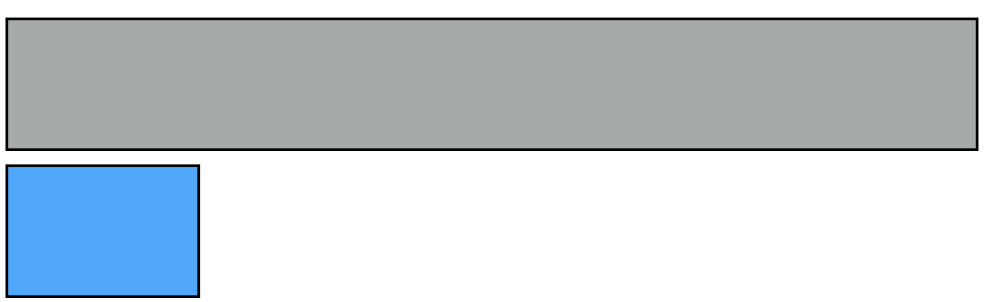
- "Stick breaking"

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

Choosing $K = \infty$

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

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



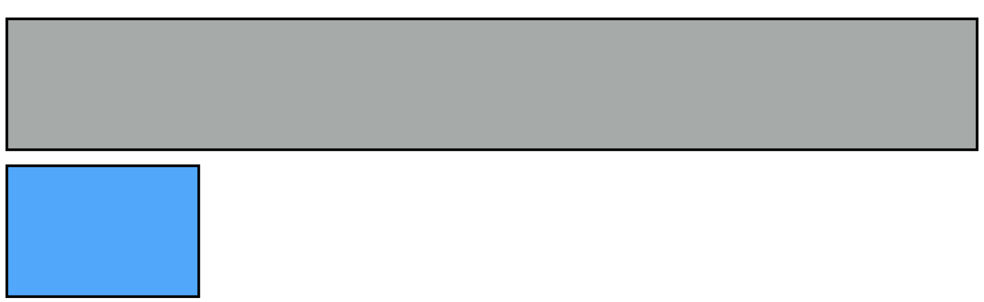
- "Stick breaking"

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

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

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



- "Stick breaking"

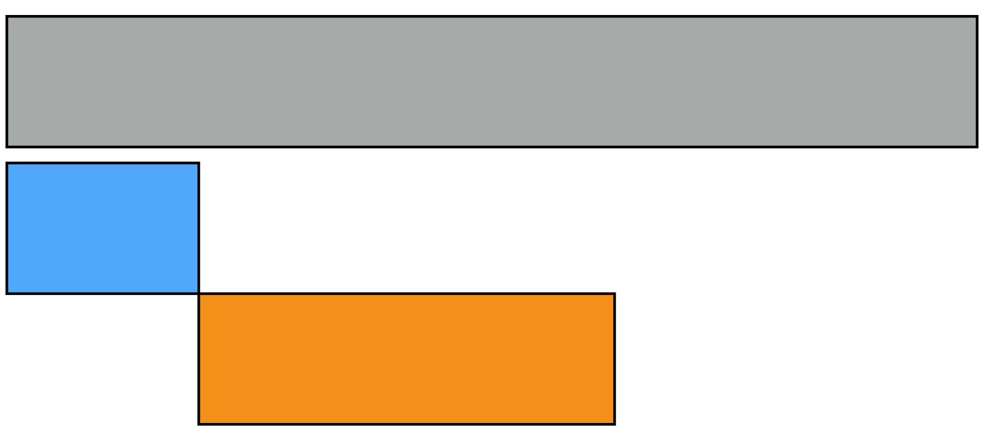
$$V_1 \sim \text{Beta}(a_1, a_2 + a_3 + a_4) \qquad \rho_1 = V_1$$

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

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

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



- "Stick breaking"

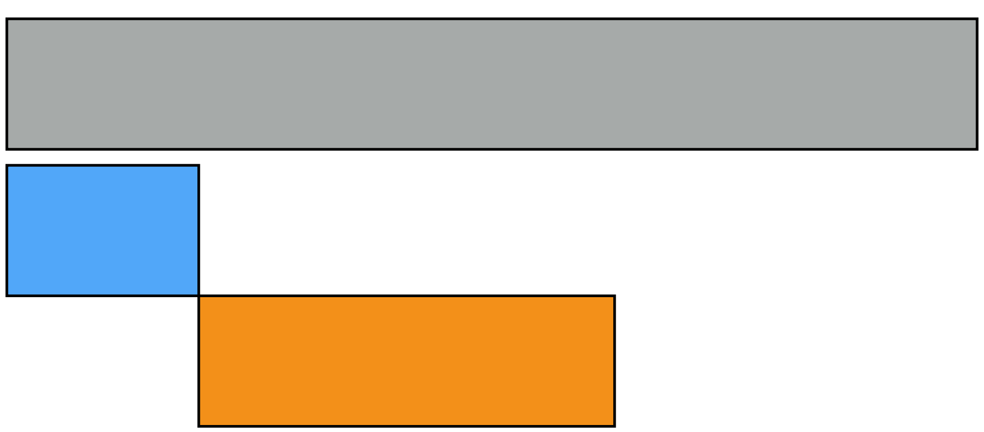
$$V_1 \sim \mathrm{Beta}(a_1, a_2 + a_3 + a_4) \qquad \rho_1 = V_1$$

$$V_2 \sim \mathrm{Beta}(a_2, a_3 + a_4) \qquad \rho_2 = (1 - V_1)V_2$$

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

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



- "Stick breaking"

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

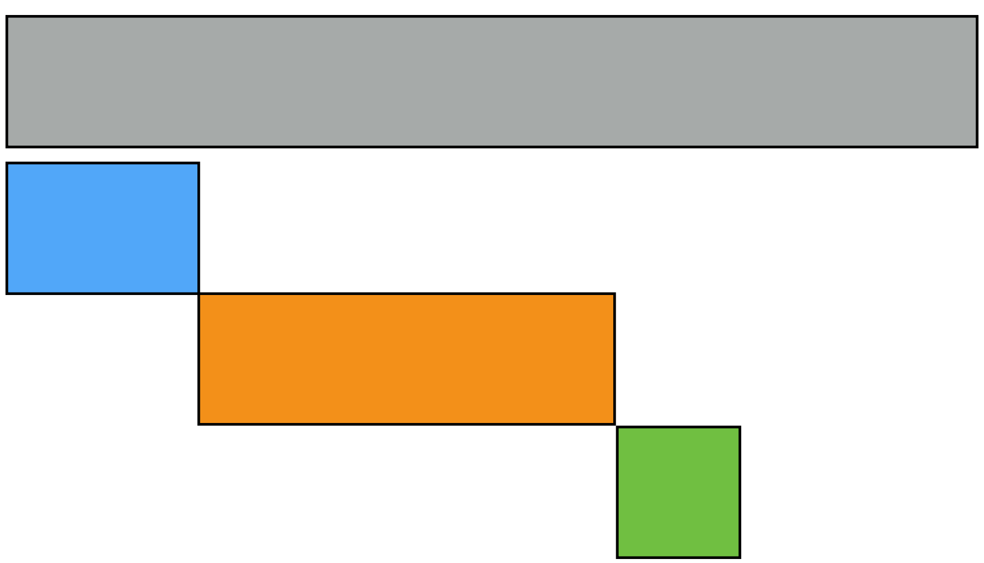
$V_2 \sim \text{Beta}(a_2, a_3 + a_4)$ $\quad \rho_2 = (1 - V_1)V_2$

$V_3 \sim \text{Beta}(a_3, a_4)$

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

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



- "Stick breaking"

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

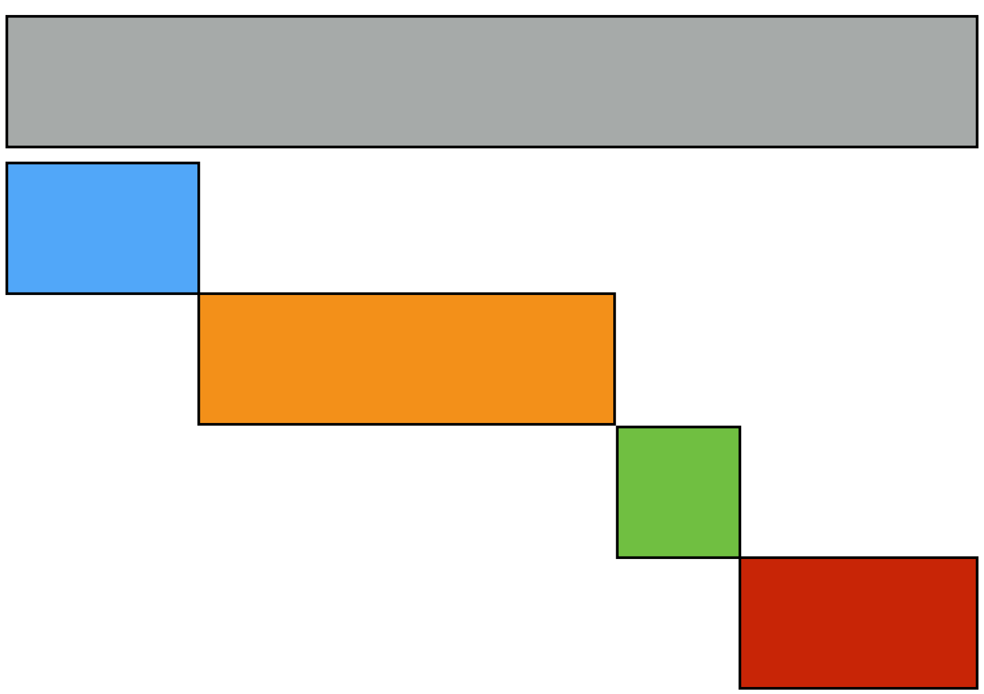
$V_2 \sim \text{Beta}(a_2, a_3 + a_4)$ $\quad \rho_2 = (1 - V_1)V_2$

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

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

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



- "Stick breaking"

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

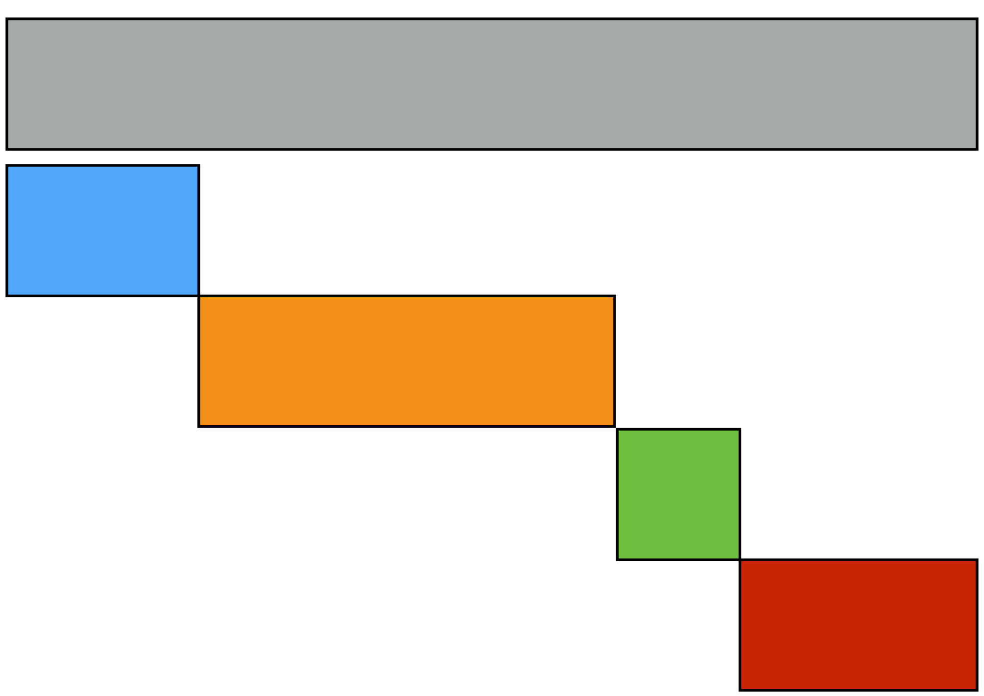
$$V_2 \sim \mathrm{Beta}(a_2, a_3 + a_4) \qquad \rho_2 = (1 - V_1)V_2$$

$$V_3 \sim \mathrm{Beta}(a_3, a_4) \qquad \rho_3 = (1 - V_1)(1 - V_2)V_3$$

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

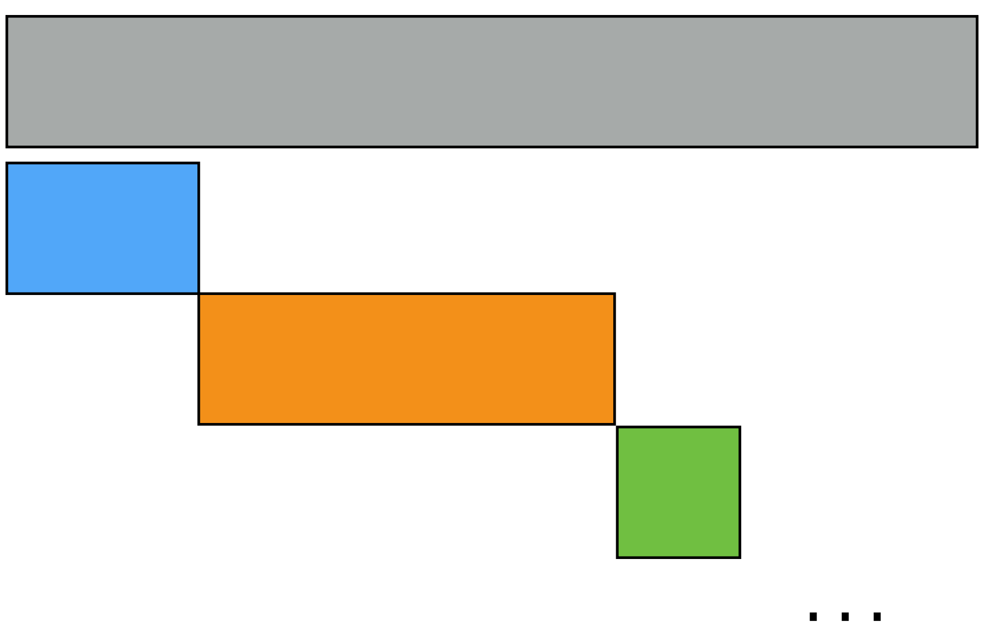
Choosing $K = \infty$

- 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?

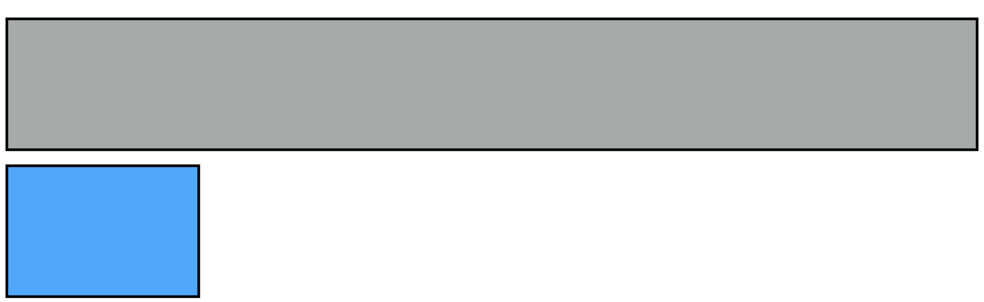
Choosing $K = \infty$

- 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?

$$V_1 \sim \mathrm{Beta}(a_1, b_1)$$

Choosing $K = \infty$

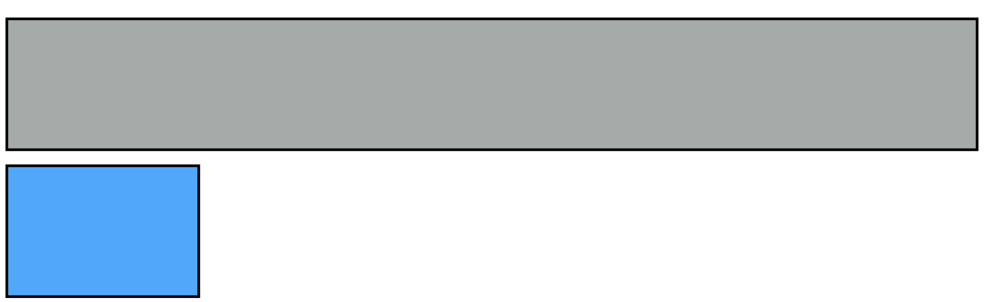
- Here, difficult to choose finite K in advance (contrast with small K): don't know K, difficult to infer, streaming data
- How to generate $K = \infty$ strictly positive frequencies that sum to one?



$$V_1 \sim \text{Beta}(a_1, b_1) \qquad \rho_1 = V_1$$

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?

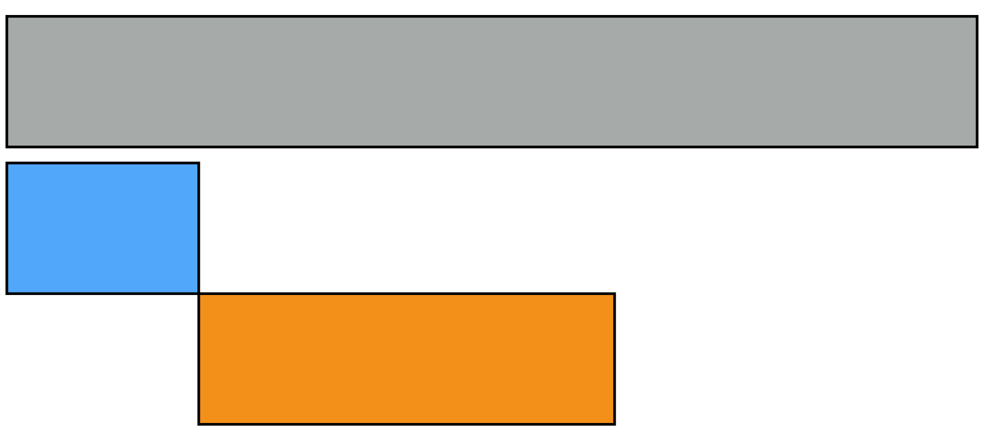


$$V_1 \sim \mathrm{Beta}(a_1, b_1) \qquad \rho_1 = V_1$$

$$V_2 \sim \mathrm{Beta}(a_2, b_2)$$

Choosing $K = \infty$

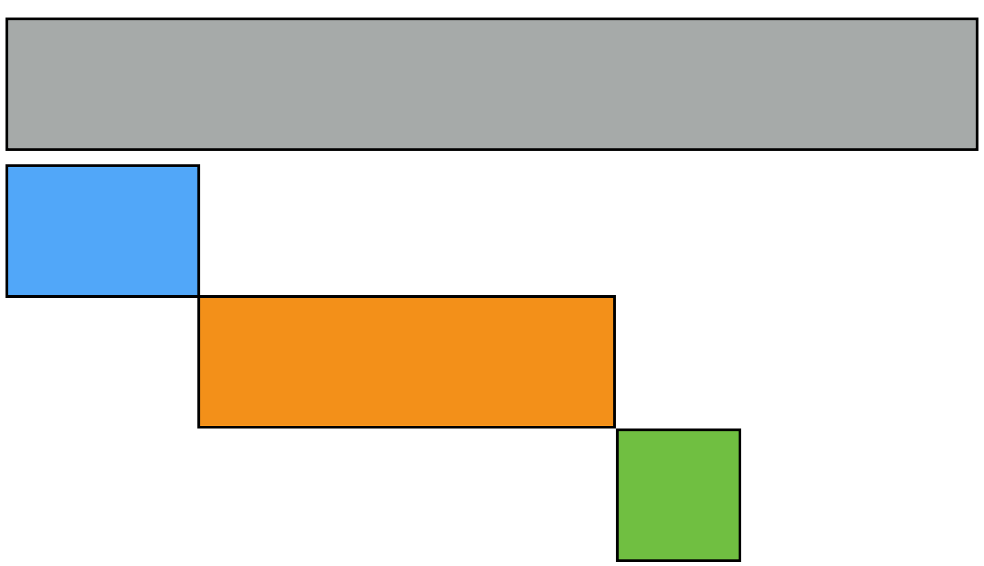
- Here, difficult to choose finite K in advance (contrast with small K): don't know K, difficult to infer, streaming data
- How to generate $K = \infty$ strictly positive frequencies that sum to one?



$$V_1 \sim \mathrm{Beta}(a_1, b_1) \qquad \rho_1 = V_1$$
$$V_2 \sim \mathrm{Beta}(a_2, b_2) \qquad \rho_2 = (1 - V_1)V_2$$

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?

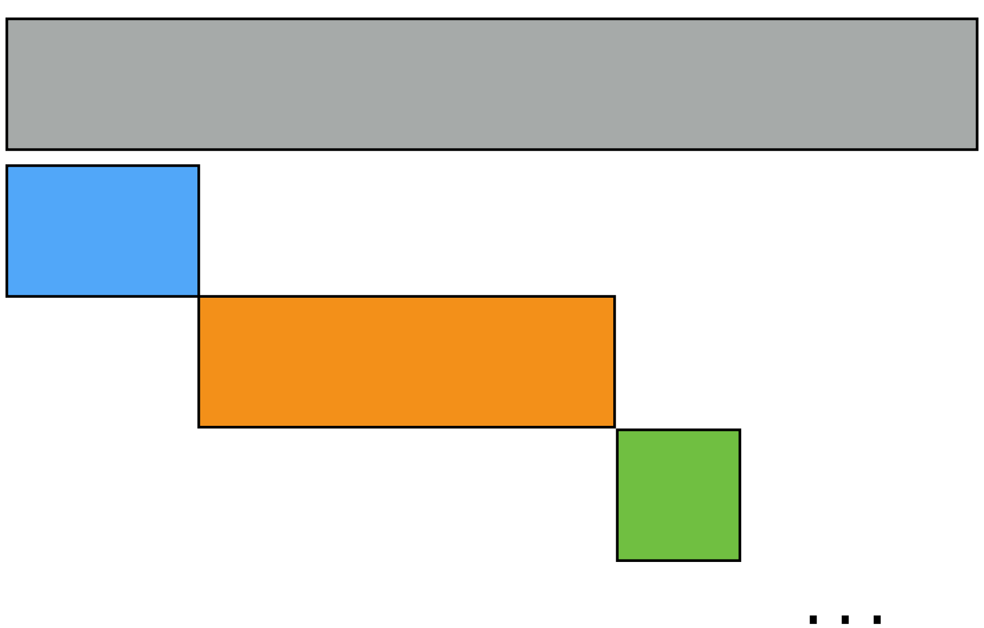


$$V_1 \sim \text{Beta}(a_1, b_1) \qquad \rho_1 = V_1$$

$$V_2 \sim \text{Beta}(a_2, b_2) \qquad \rho_2 = (1 - V_1)V_2$$

Choosing $K = \infty$

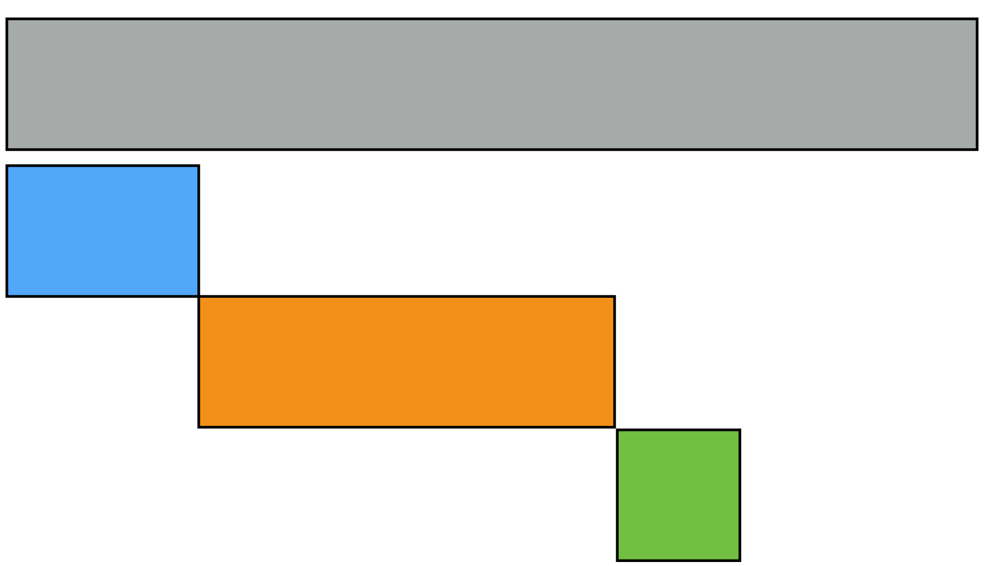
- Here, difficult to choose finite K in advance (contrast with small K): don't know K, difficult to infer, streaming data
- How to generate $K = \infty$ strictly positive frequencies that sum to one?



$$V_1 \sim \text{Beta}(a_1, b_1) \qquad \rho_1 = V_1$$
$$V_2 \sim \text{Beta}(a_2, b_2) \qquad \rho_2 = (1 - V_1)V_2$$

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?



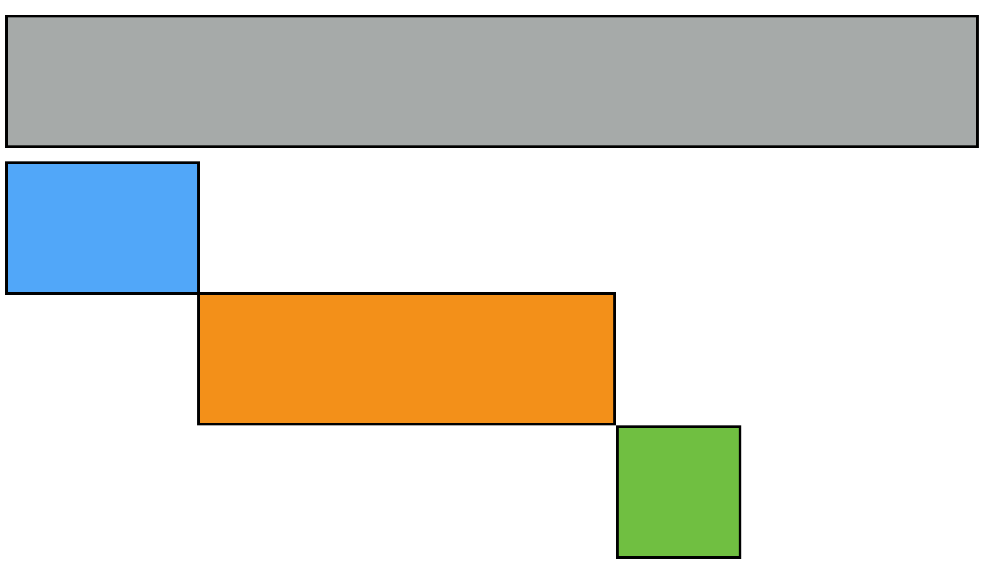
$$V_1 \sim \mathrm{Beta}(a_1, b_1) \qquad \rho_1 = V_1$$

$$V_2 \sim \mathrm{Beta}(a_2, b_2) \qquad \rho_2 = (1 - V_1)V_2$$

$$\cdots \quad V_k \sim \mathrm{Beta}(a_k, b_k)$$

Choosing $K = \infty$

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



$$V_1 \sim \mathrm{Beta}(a_1, b_1) \qquad \rho_1 = V_1$$

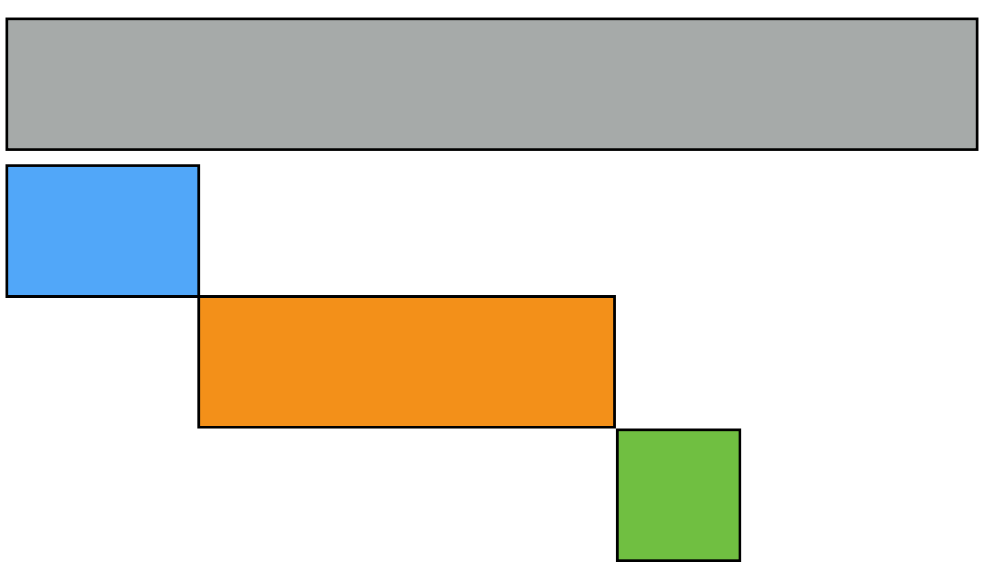
$$V_2 \sim \mathrm{Beta}(a_2, b_2) \qquad \rho_2 = (1 - V_1)V_2$$

$$\cdots$$

$$V_k \sim \mathrm{Beta}(a_k, b_k) \qquad \rho_k = \left[\prod_{j=1}^{k-1}(1 - V_j)\right] V_k$$

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?



$$V_1 \sim \mathrm{Beta}(a_1, b_1) \qquad \rho_1 = V_1$$

$$V_2 \sim \mathrm{Beta}(a_2, b_2) \qquad \rho_2 = (1 - V_1)V_2$$

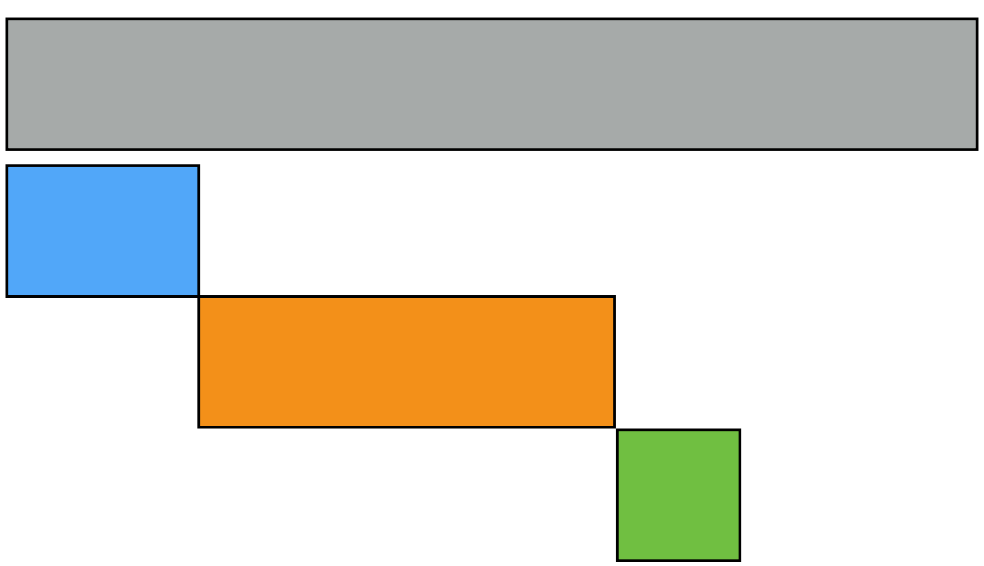
$$\cdots$$

$$V_k \sim \mathrm{Beta}(a_k, b_k) \qquad \rho_k = \left[\prod_{j=1}^{k-1}(1 - V_j)\right] V_k$$

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



$$V_1 \sim \text{Beta}(a_1, b_1) \qquad \rho_1 = V_1$$

$$V_2 \sim \text{Beta}(a_2, b_2) \qquad \rho_2 = (1 - V_1)V_2$$

$$\cdots$$

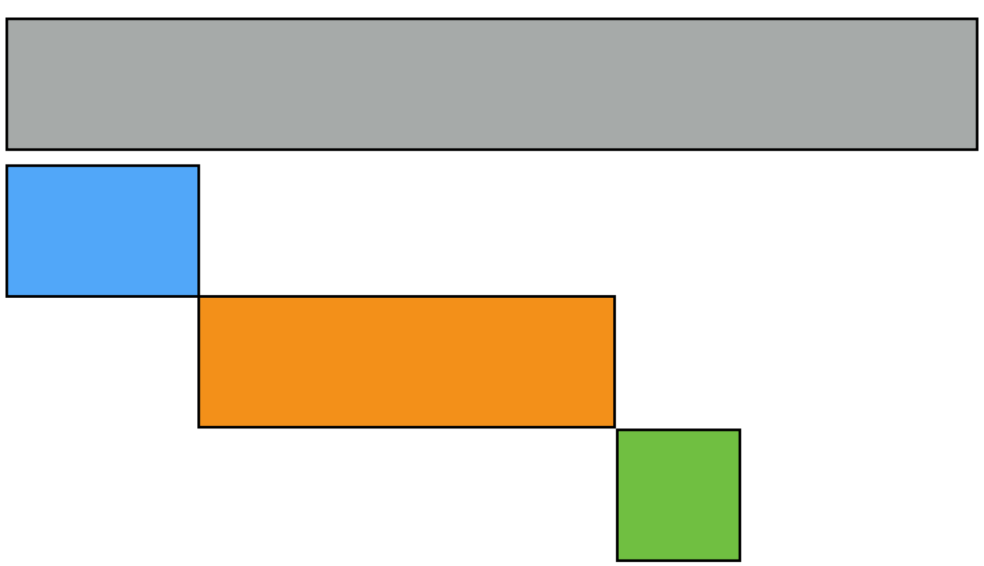
$$V_k \sim \text{Beta}(a_k, b_k) \qquad \rho_k = \left[\prod_{j=1}^{k-1}(1 - V_j)\right] V_k$$

[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:

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



$$V_1 \sim \mathrm{Beta}(a_1, b_1) \qquad \rho_1 = V_1$$

$$V_2 \sim \mathrm{Beta}(a_2, b_2) \qquad \rho_2 = (1 - V_1)V_2$$

$$\cdots$$

$$V_k \sim \mathrm{Beta}(a_k, b_k) \qquad \rho_k = \left[\prod_{j=1}^{k-1}(1 - V_j)\right] V_k$$

[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

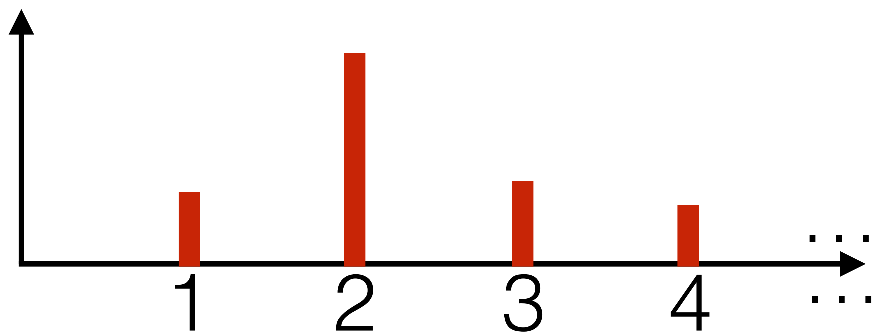
- Example problem: clustering
- Example NPBayes model: Dirichlet process
- Chinese restaurant process
- Inference
- Venture further into the wild world of Nonparametric Bayes

- 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

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

- 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 $\rho_{1:K} \sim \mathrm{Dirichlet}(a_{1:K})$; prove equivalence to

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

References

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