Statistical Machine Learning

Lecture 01
Non-Parametric Bayesian Introduction

Spring 2021

Sharif University of Technology

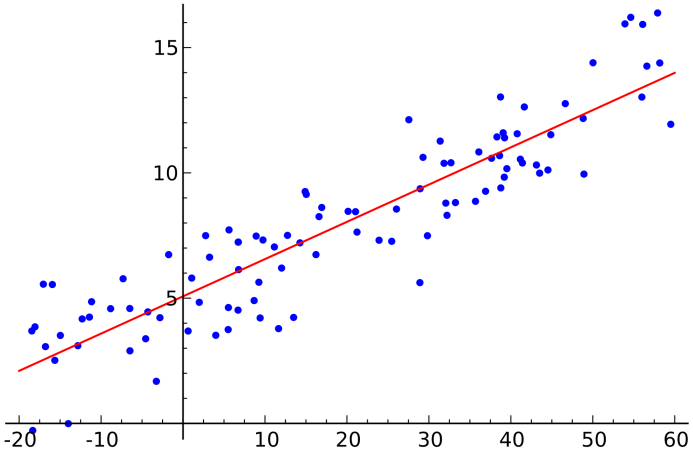
Bayesian Nonparametrics

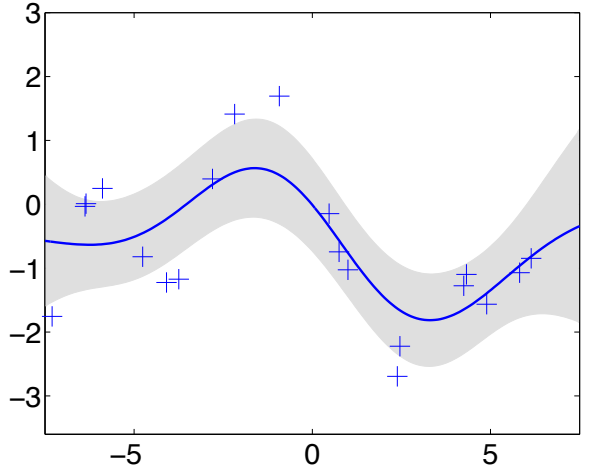
Part I

Peter Orbanz

Parameters

$$P(X|\theta) \quad = \quad \text{Probability}[\text{data}|\text{pattern}]$$





Inference idea

$$\text{data} \quad = \quad \text{underlying pattern} \quad + \quad \text{independent noise}$$

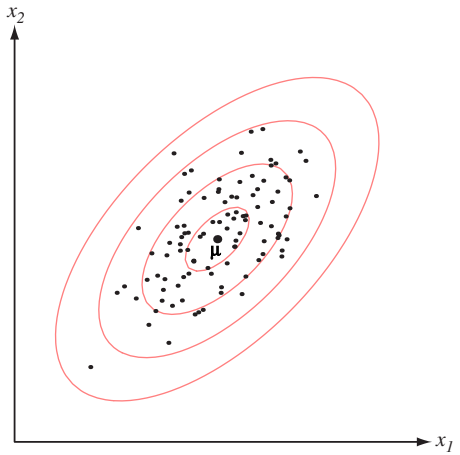
Parametric model

- Number of parameters fixed (or constantly bounded) w.r.t. sample size

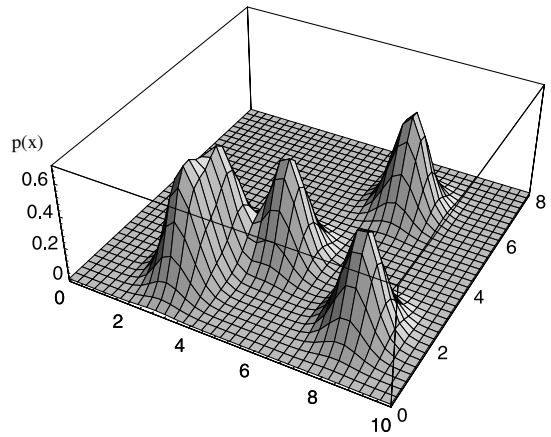
Nonparametric model

- Number of parameters grows with sample size
- ∞-dimensional parameter space

Example: Density estimation



Parametric



Nonparametric

Definition

A nonparametric Bayesian model is a Bayesian model on an ∞-dimensional parameter space.

Interpretation

Parameter space $\mathcal{T}$ = set of possible patterns, for example:

Problem	$\mathcal{T}$
Density estimation	Probability distributions
Regression	Smooth functions
Clustering	Partitions

Solution to Bayesian problem = posterior distribution on patterns

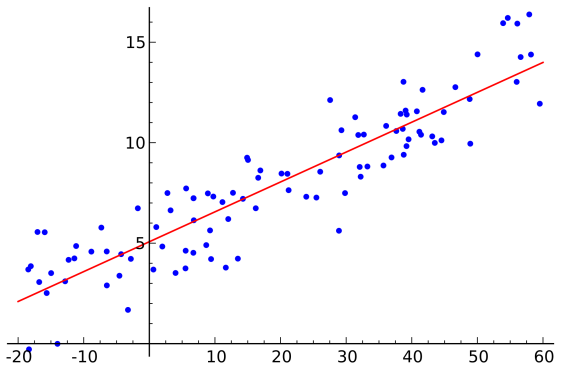
Can we justify our assumptions?

Recall:

$$\text{data} \;=\; \text{pattern} \;+\; \text{noise}$$

In Bayes' theorem:

$$Q(d\theta|x_1,\ldots,x_n) = \frac{\prod_{j=1}^{n} p(x_j|\theta)}{p(x_1,\ldots,x_n)} Q(d\theta)$$



Definition

$X_1, X_2, \ldots$ are *exchangeable* if $P(X_1, X_2, \ldots)$ is invariant under any permutation σ:

$$P(X_1 = x_1, X_2 = x_2, \ldots) = P(X_1 = x_{\sigma(1)}, X_2 = x_{\sigma(2)}, \ldots)$$

In words:

Order of observations does not matter.

De Finetti's Theorem

$$P(X_1 = x_1, X_2 = x_2, \ldots) = \int_{\mathbf{M}(\mathcal{X})} \Big(\prod_{j=1}^{\infty} \theta(X_j = x_j) \Big) Q(d\theta)$$

$$\Updownarrow$$

$$X_1, X_2, \ldots \text{ exchangeable}$$

where:

- $\mathbf{M}(\mathcal{X})$ is the set of probability measures on $\mathcal{X}$
- θ are values of a random probability measure Θ with distribution Q

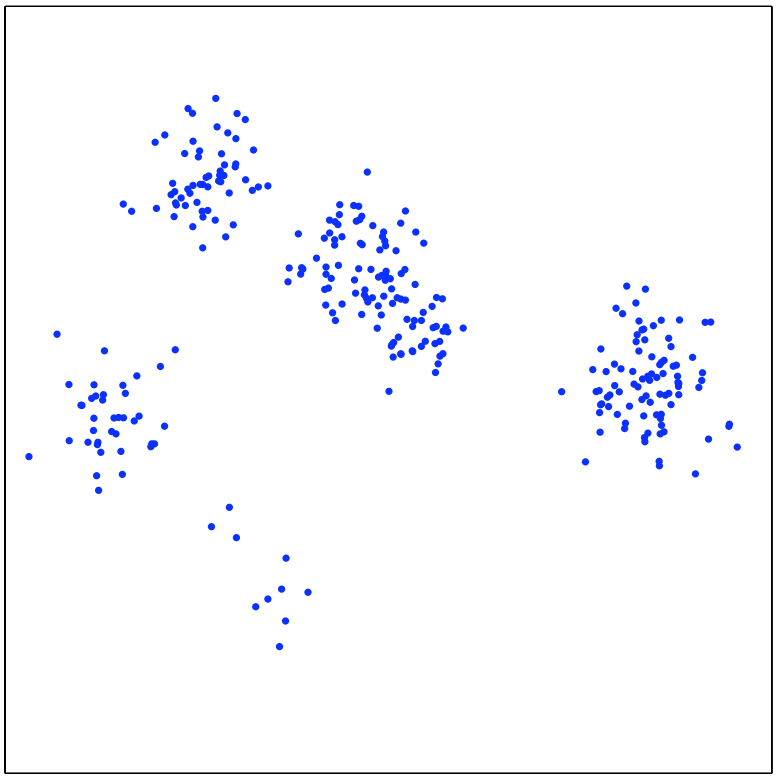
Implications

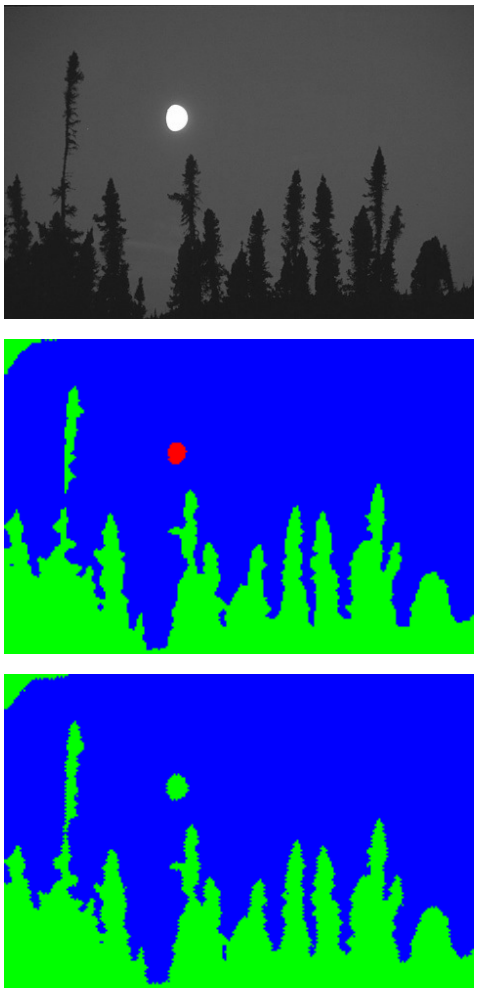
- Exchangeable data decomposes into pattern and noise
- More general than i.i.d.-assumption
- Caution: θ is in general an ∞-dimensional quantity

CLUSTERING

CLUSTERING

- Observations $X_1, X_2, \ldots$
- Each observation belongs to exactly one cluster
- Unknown pattern = partition of $\{1, \ldots, n\}$ or $\mathbb{N}$





Mixture models

$$p(x|m) = \int_{\Omega_\theta} p(x|\theta) m(d\theta)$$

m is called the *mixing measure*

Two-stage sampling

Sample $X \sim p(\,.\,|m)$ as:

1. $\Theta \sim m$
2. $X \sim p(\,.\,|\theta)$

Finite mixture model

$$p(x|\boldsymbol{\theta}, \mathbf{c}) = \int_{\Omega_\theta} p(x|\theta) m(d\theta) \qquad \text{with} \qquad m(\,.\,) = \sum_{k=1}^{K} c_k \delta_{\theta_k}(\,.\,)$$

Random mixing measure

$$M(\,.\,) = \sum_{k=1}^{K} C_k \delta_{\Theta_k}(\,.\,)$$

Conjugate priors

A Bayesian model is *conjugate* if the posterior is an element of the same class of distributions as the prior ("closure under sampling").

$p(x\|\theta)$	conjugate prior
$\frac{1}{Z(\theta)} h(x) \exp(\langle S(x), \theta \rangle)$	$\frac{1}{K(\lambda, y)} \exp(\langle \theta, y \rangle - \lambda \log Z(\theta))$
Gaussian	Gaussian/inverse Wishart
multinomial	Dirichlet
...	...

Choice of priors in BMM

- Choose conjugate prior for each parameter
- In particular: Dirichlet prior on $(C_1, \ldots, C_k)$

Dirichlet process

A Dirichlet process is a distribution on random probability measures of the form

$$M(\,.\,) = \sum_{k=1}^{\infty} C_k \delta_{\Theta_k}(\,.\,) \qquad \text{where} \qquad \sum_{k=1}^{\infty} C_k = 1$$

Constructive definition of DP (α, G_0)

$$\Theta_k \sim_{\text{iid}} G_0$$
$$V_k \sim_{\text{iid}} \text{Beta}(1, \alpha)$$

Compute C_k as

$$C_k := V_k \prod_{i=1}^{k-1} (1 - V_i)$$

"Stick-breaking construction"

Posterior distribution

DP Posterior

$$\theta_{n+1}|\theta_1,\ldots,\theta_n \sim \frac{1}{n+\alpha}\sum_{j=1}^{n}\delta_{\theta_j}(\theta_{n+1}) + \frac{\alpha}{n+\alpha}G_0(\theta_{n+1})$$

Mixture Posterior

$$p(x_{n+1}|x_1,\ldots,x_n) = \sum_{k=1}^{K_n}\frac{n_k}{n+\alpha}p(x_{n+1}|\theta_k^*) + \frac{\alpha}{n+\alpha}\int p(x_{n+1}|\theta)G_0(\theta)d\theta$$

Conjugacy

- The posterior of DP (α, G_0) is DP $\left(\alpha+n, \frac{1}{n+\alpha}\left(\sum_k n_k\delta_{\theta_k^*} + \alpha G_0\right)\right)$
- Hence: The Dirichlet process is conjugate.

Latent variables

$$p(x_{n+1}|x_1,\ldots,x_n) = \sum_{k=1}^{K_n} \frac{n_k}{n+\alpha} p(x_{n+1}|\theta_k^*) + \frac{\alpha}{n+\alpha} \int p(x_{n+1}|\theta) G_0(\theta) d\theta$$

We do not actually observe the Θ_j (they are latent). We observe X_j.

Assignment probabilities

$$\begin{pmatrix} q_{10} & q_{11} & \ldots & q_{1K_n} \\ \vdots & \vdots & & \vdots \\ q_{n0} & q_{n1} & \ldots & q_{nK_n} \end{pmatrix}$$

Where:

- $q_{jk} \propto n_k p(x_j|\theta_k^*)$
- $q_{j0} \propto \alpha \int p(x_j|\theta) G_0(\theta) d\theta$

Gibbs Sampling

Uses an assignment variable ϕ_j for each observation X_j.

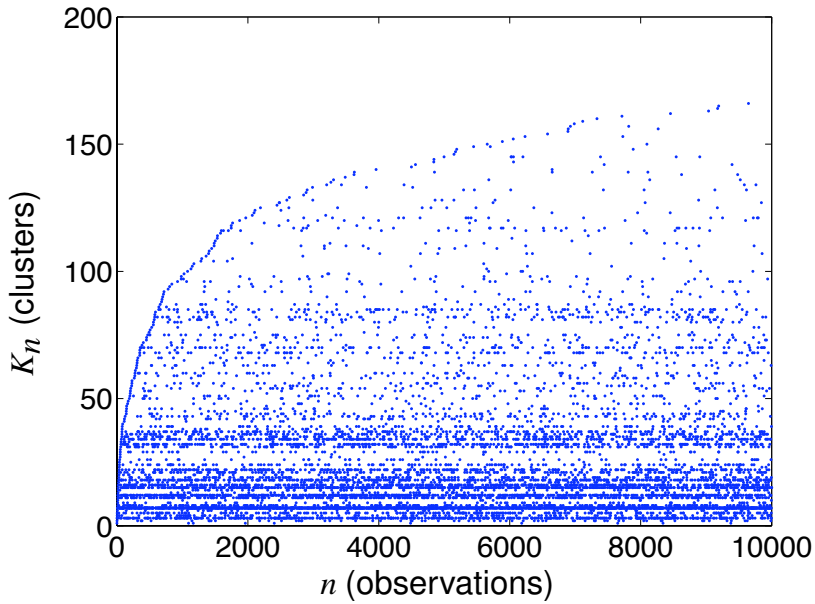
- Assignment step: Sample $\phi_j \sim \text{Multinomial}(q_{j0},\ldots,q_{jK_n})$
- Parameter sampling: $\theta_k^* \sim G_0(\theta_k^*) \prod_{x_j \in \text{ Cluster } k} p(x_j|\theta_k^*)$

Number of Clusters

Dirichlet process

$K_n = \#$ clusters in sample of size n

$$\mathbb{E}[K_n] = O(\log(n))$$



Modeling assumption

- Parametric clustering: K_∞ is *finite* (possibly unknown, but fixed).
- Nonparametric clustering: K_∞ is *infinite*

Rephrasing the question

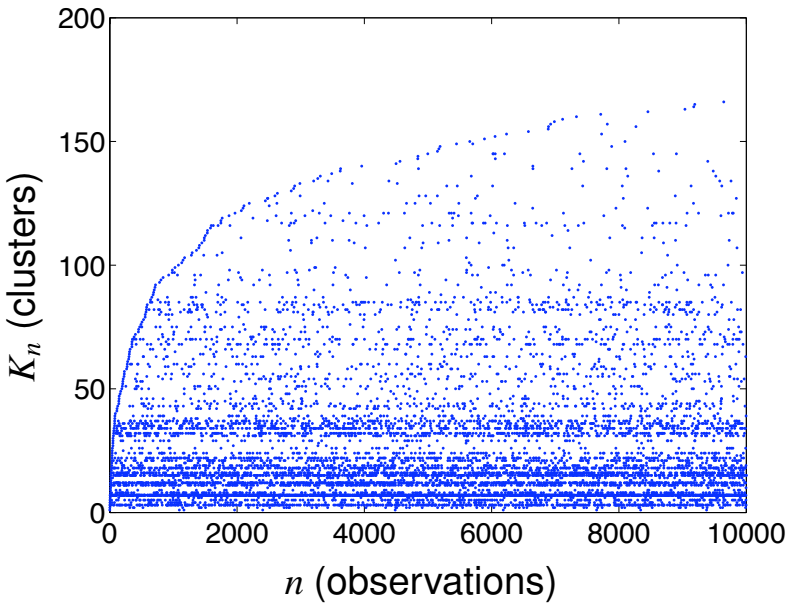
- Estimate of K_n is controlled by distribution of the cluster sizes C_k in $\sum_k C_k \delta_{\Theta_k}$.
- Ask instead: What should we assume about the distribution of C_k?

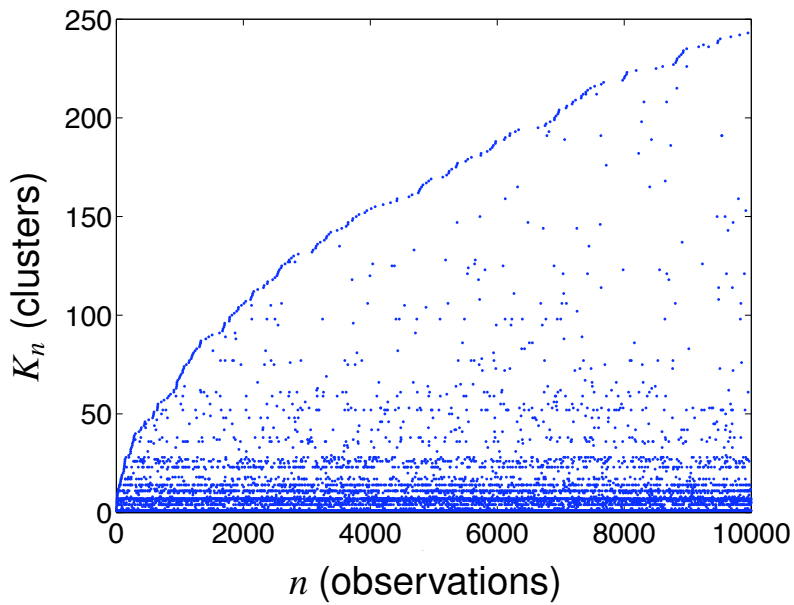
Pitman-Yor process

$$p(x_{n+1}|x_1,\ldots,x_n) = \sum_{k=1}^{K_n} \frac{n_k - d}{n+\alpha} p(x_{n+1}|\theta_k^*) + \frac{\alpha + K_n \cdot d}{n+\alpha} \int p(x_{n+1}|\theta) G_0(\theta) d\theta$$

Discount parameter $d \in [0, 1]$.

Cluster sizes





Power Laws

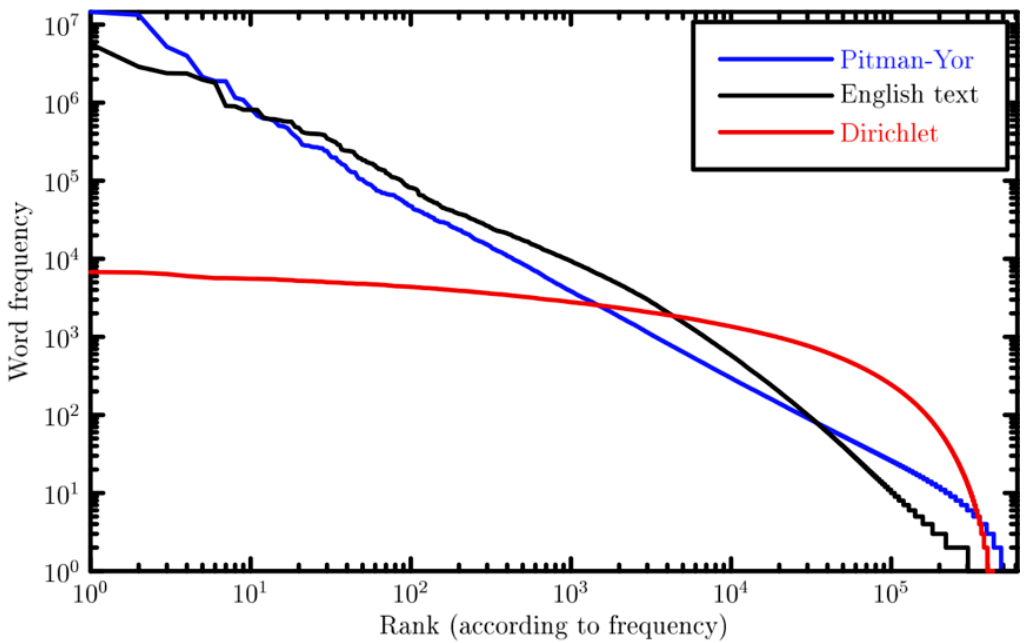
The distribution of cluster sizes is called a *power law* if

$$C_j \sim \gamma(\beta) \cdot j^{-\beta} \qquad \text{for some } \beta \in [0, 1] \ .$$

Examples of power laws

- Word frequencies
- Popularity (number of friends) in social networks

Pitman-Yor language model



Discrete measures and partitions

Sampling from a discrete measure determines a *partition* of $\mathbb{N}$ into blocks b_k:

$$\Theta_n \sim_{\text{iid}} \sum_{k=1}^{\infty} c_k \delta_{\theta_k^*} \qquad \text{and set} \qquad n \in b_k \quad \Leftrightarrow \quad \Theta_n = \theta_k^*$$

As $n \longrightarrow \infty$, the block proportions converge: $\frac{|b_k|}{n} \longrightarrow c_k$

Induced random partition

The distribution of a random discrete measure $M = \sum_{k=1}^{\infty} C_k \delta_{\Theta_k}$ induces the distribution of a *random partition* $\Pi = (B_1, B_2, \dots)$.

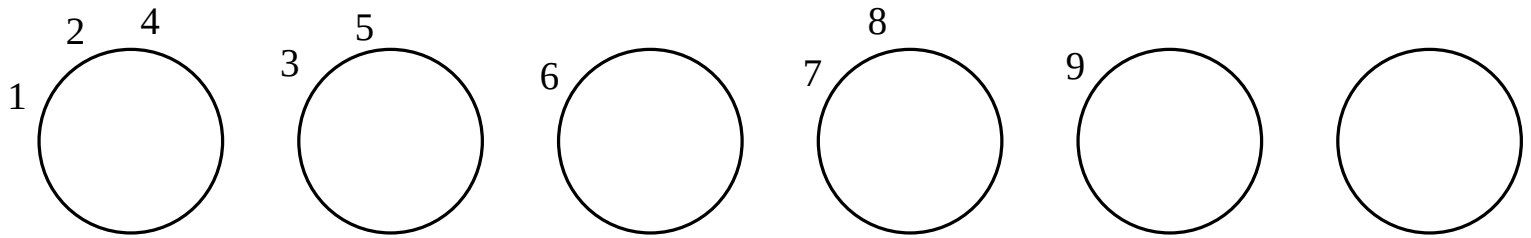
Exchangeable random partitions

- Π is called *exchangeable* if its distribution depends only on the sizes of its blocks.
- All exchangeable random parititions, and only those, can be represented by a random discrete distribution as above (Kingman's theorem).

Chinese Restaurant Process

The distribution of the random partition induced by the Dirichlet process is called the *Chinese Restaurant Process*.

"Customers and tables" analogy

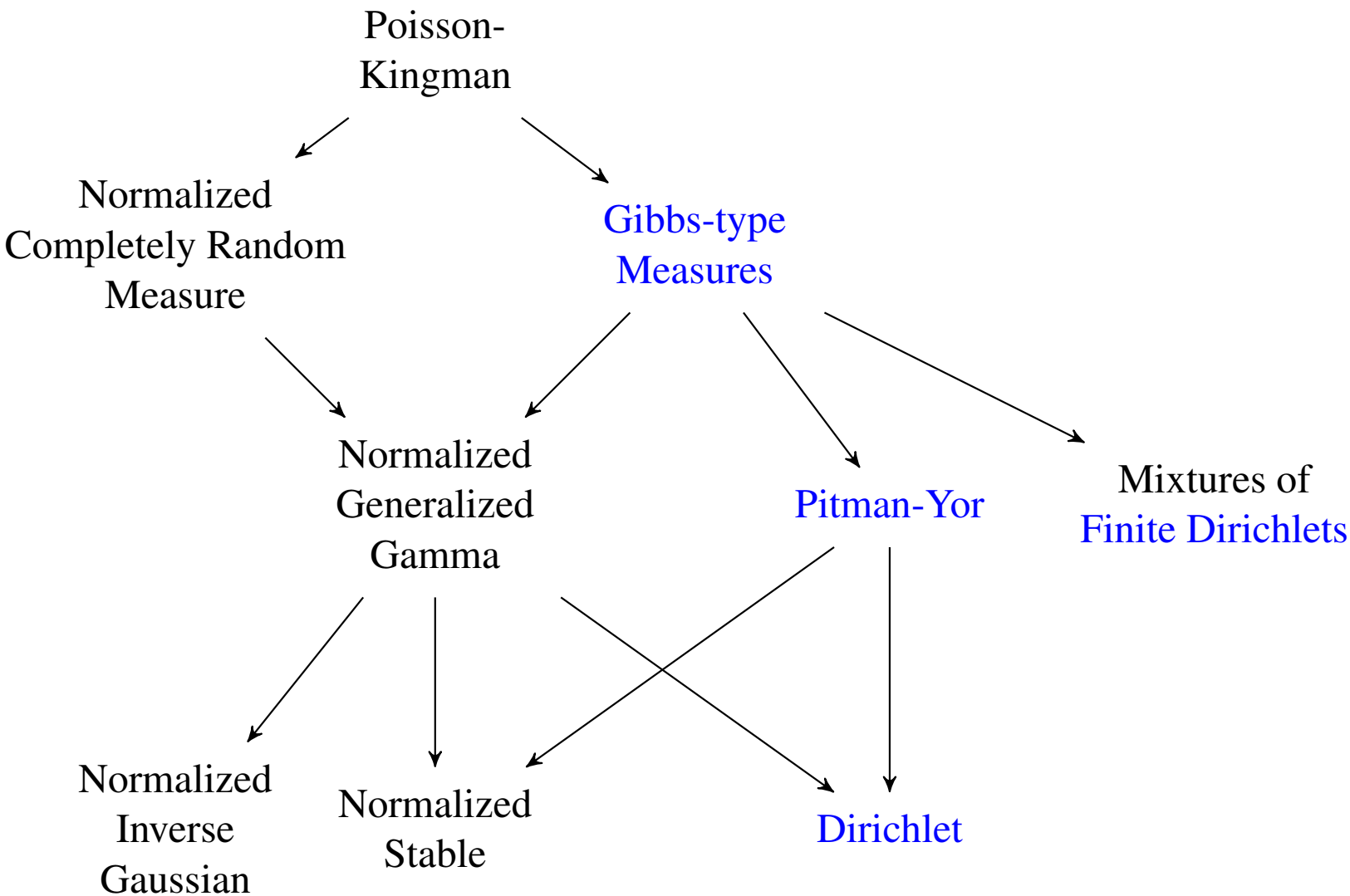


Customers = observations (indices in $\mathbb{N}$)
Tables = clusters (blocks)

Historical remark

- Originally introduced by Dubins & Pitman as a distribution on infinite permutations
- A permutation of n items defines a partition of $\{1, \ldots, n\}$ (regard cycles of permutation as blocks of partition)
- The induced distribution on partitions is the CRP we use in clustering

Families of Exchangeable Random Partitions



Classification (due to Prünster)

class	probability of new cluster	prior class
I	$\mathbb{P}\{\Theta_{n+1} \in \text{new cluster} \mid \Theta^{(n)}\} = f(n)$	Dirichlet processes
II	$\mathbb{P}\{\Theta_{n+1} \in \text{new cluster} \mid \Theta^{(n)}\} = f(n, K_n)$	Gibbs-type measures
III	$\mathbb{P}\{\Theta_{n+1} \in \text{new cluster} \mid \Theta^{(n)}\} = f(n, K_n, \mathbf{n})$	

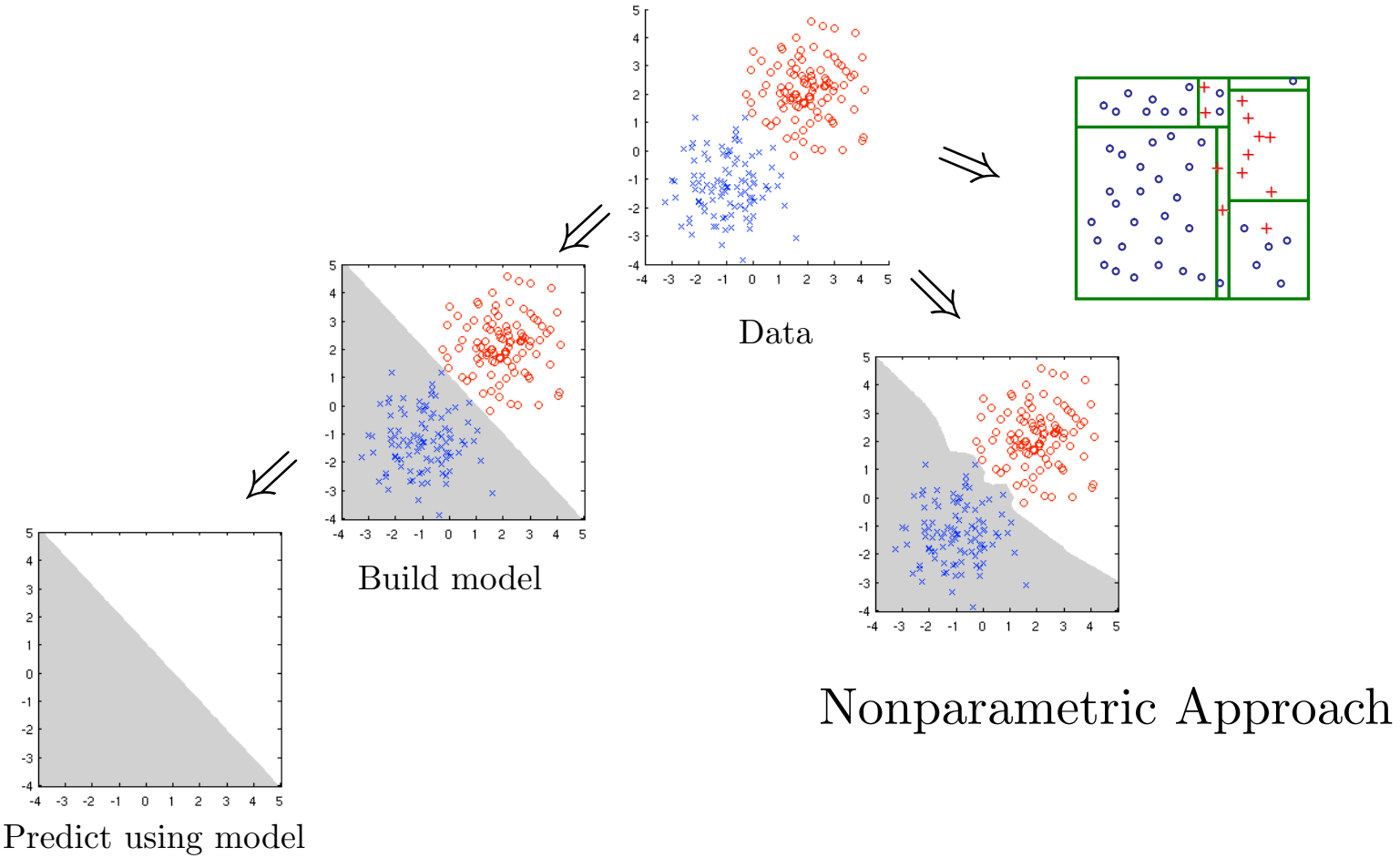
General partition priors

- Gibbs-type measures are completely classified [GP06b]
- Properties of some cases well-studied, e.g.:
  - Dirichlet process
  - Pitman-Yor process
  - Normalized inverse Gaussian process [LMP05b]
- In the future: We will have a range of models which express different prior assumptions on the distribution of cluster sizes.

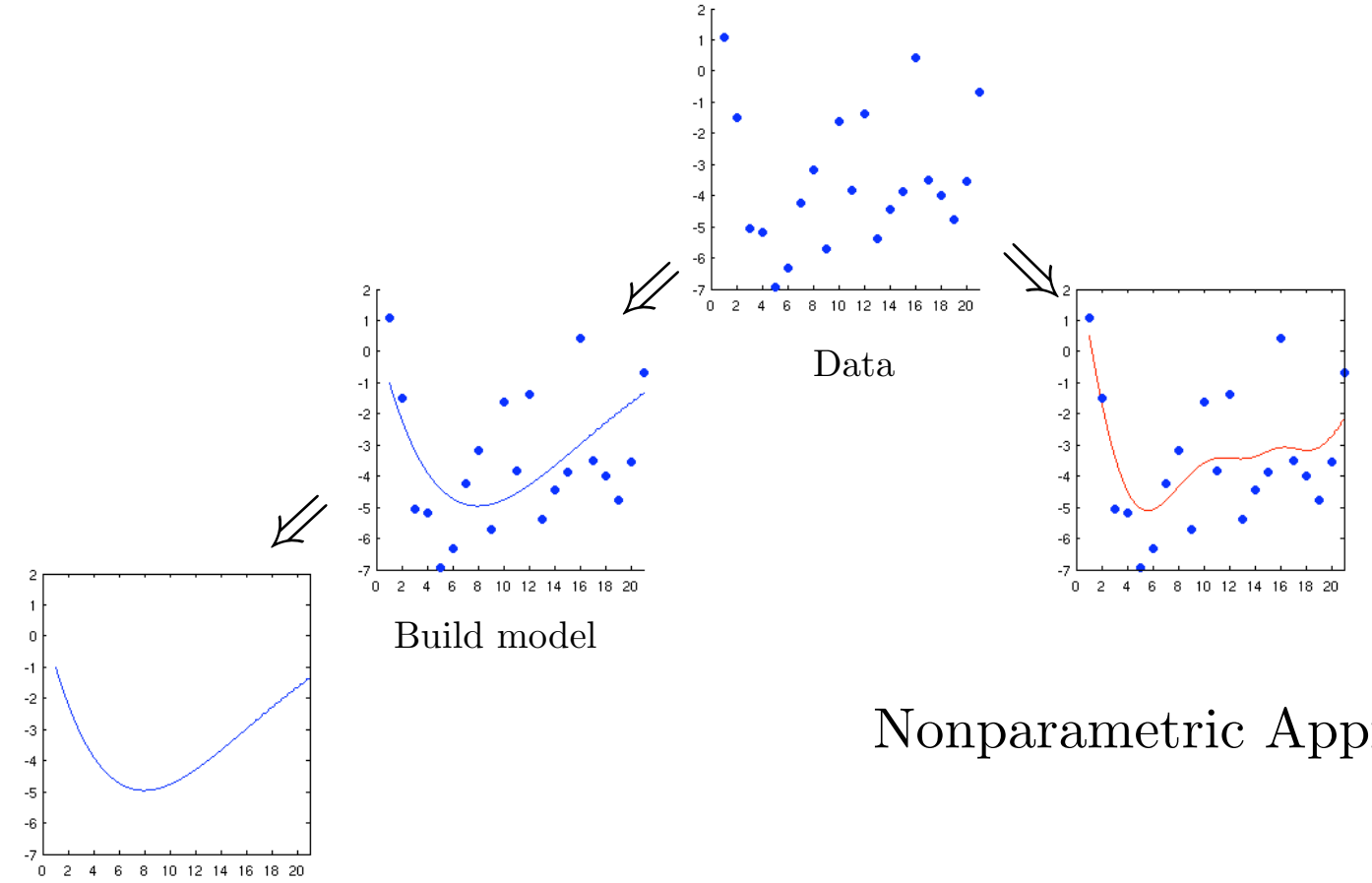
Nonparametric

Nonparametric: Does NOT mean there are no parameters.

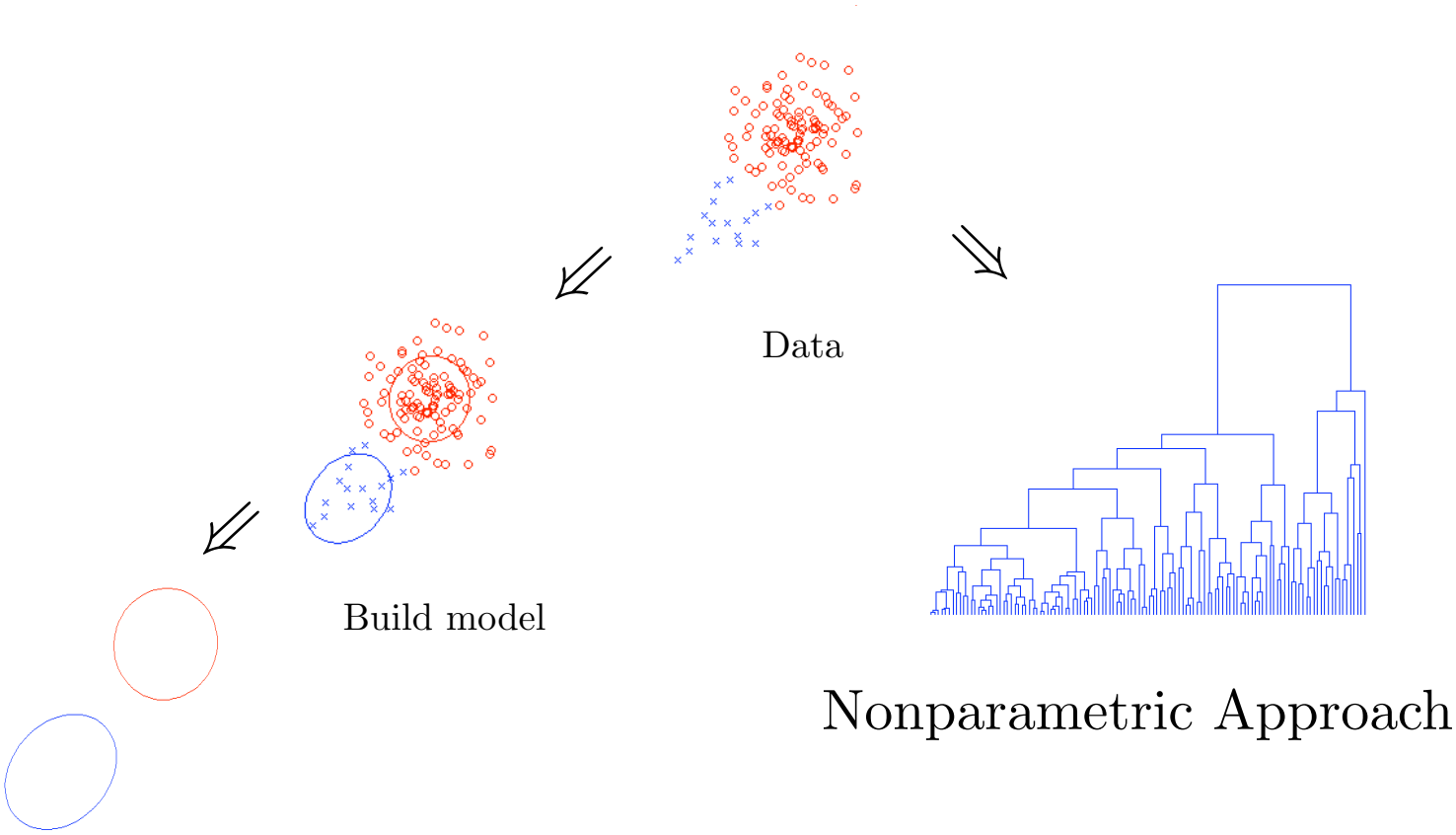
Example: Classification



Example: Regression



Example: Clustering



Why Be Bayesian?

You can take a course on this question.

Why Be Bayesian?

You can take a course on this question. One answer:

Infinite Exchangeability: $\forall n\ p(x_1, \ldots, x_n) = p(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$

Why Be Bayesian?

You can take a course on this question. One answer:

Infinite Exchangeability: $\forall n\ p(x_1, \ldots, x_n) = p(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$

De Finetti's Theorem (1955): If $(x_1, x_2, \ldots)$ are *infinitely exchangeable*, then $\forall n$

$$p(x_1, \ldots, x_n) = \int \left(\prod_{i=1}^{n} p(x_i|\theta) \right) dP(\theta)$$

for some random variable θ.

Simple Example

Task: Toss a (potentially biased) coin N times. Compute θ, the probability of heads.

Suppose we observe: $\{\text{T, H, H, T}\}$. What do we think θ is?

Simple Example

Task: Toss a (potentially biased) coin N times. Compute θ, the probability of heads.

Suppose we observe: $\{$T, H, H, T$\}$. What do we think θ is? The maximum likelihood estimate is $\theta = 1/2$. Seems reasonable.

Simple Example

Task: Toss a (potentially biased) coin N times. Compute θ, the probability of heads.

Suppose we observe: $\{$T, H, H, T$\}$. What do we think θ is? The maximum likelihood estimate is $\theta = 1/2$. Seems reasonable.

Now suppose we observe: $\{$H, H, H, H$\}$. What do we think θ is?

Simple Example

Task: Toss a (potentially biased) coin N times. Compute θ, the probability of heads.

Suppose we observe: $\{$T, H, H, T$\}$. What do we think θ is? The maximum likelihood estimate is $\theta = 1/2$. Seems reasonable.

Now suppose we observe: $\{$H, H, H, H$\}$. What do we think θ is? The maximum likelihood estimate is $\theta = 1$. Seem reasonable?

Simple Example

Task: Toss a (potentially biased) coin N times. Compute θ, the probability of heads.

Suppose we observe: $\{$T, H, H, T$\}$. What do we think θ is? The maximum likelihood estimate is $\theta = 1/2$. Seems reasonable.

Now suppose we observe: $\{$H, H, H, H$\}$. What do we think θ is? The maximum likelihood estimate is $\theta = 1$. Seem reasonable?

Not really. Why?

Simple Example

When we observe $\{\text{H, H, H, H}\}$, why does $\theta = 1$ seem unreasonable?

Simple Example

When we observe $\{\text{H, H, H, H}\}$, why does $\theta = 1$ seem unreasonable?

Prior knowledge! We believe coins generally have $\theta \approx 1/2$. How to encode this? By using a Beta *prior on* θ.

Bayesian Approach to Estimating θ

Place a Beta(a, b) prior on θ. This prior has the form

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}.$$

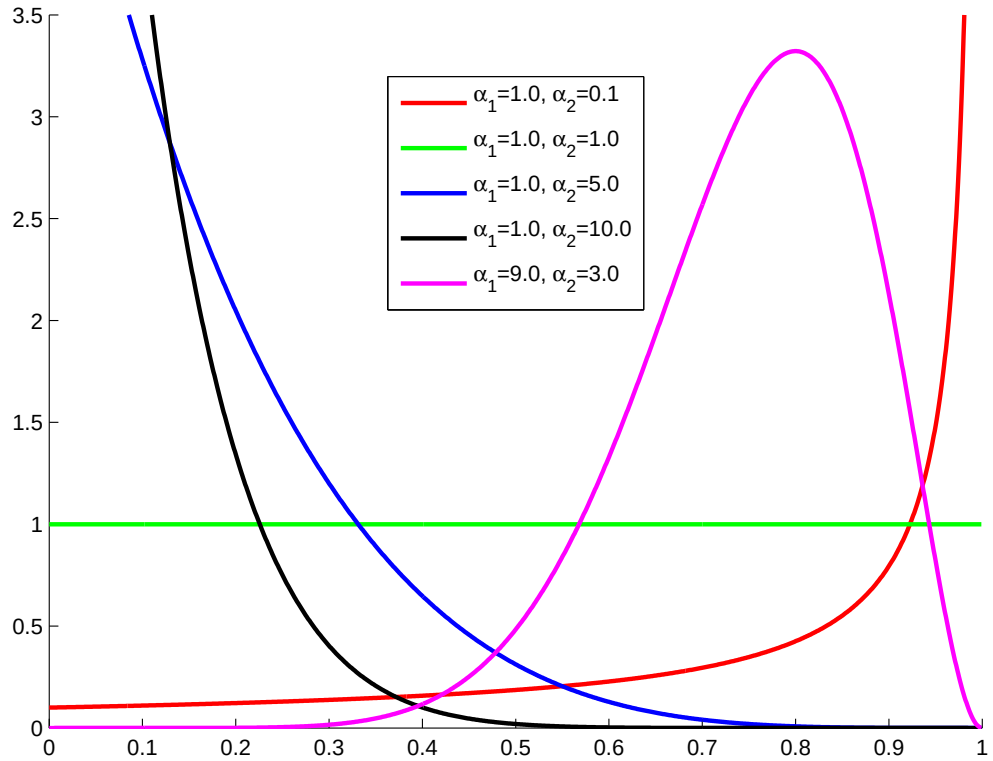
What does this distribution look like?

Bayesian Approach to Estimating θ

Place a Beta(a, b) prior on θ. This prior has the form

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}.$$

What does this distribution look like?



Bayesian Approach to Estimating θ

After observing X, a sequence with n heads and m tails, the posterior on θ is:

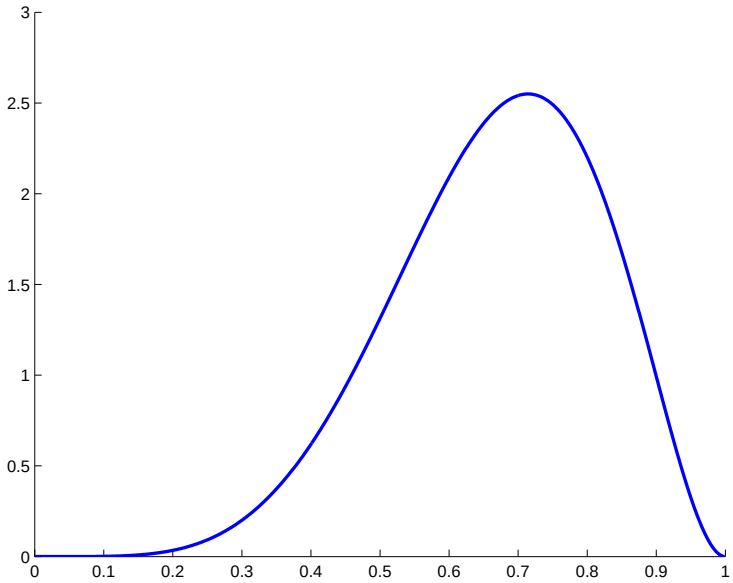
$$\begin{aligned} p(\theta|X) &\propto p(X|\theta)p(\theta) \\ &\propto \theta^{a+n-1}(1-\theta)^{b+m-1} \\ &\sim \text{Beta}(a+n, b+m). \end{aligned}$$

Bayesian Approach to Estimating θ

After observing X, a sequence with n heads and m tails, the posterior on θ is:

$$\begin{aligned} p(\theta|X) &\propto p(X|\theta)p(\theta) \\ &\propto \theta^{a+n-1}(1-\theta)^{b+m-1} \\ &\sim \text{Beta}(a+n, b+m). \end{aligned}$$

If $a = b = 1$ and we observe 5 heads and 2 tails, Beta$(6, 3)$ looks like



Nonparametric Bayesian Methods

Now we know what nonparametric and Bayesian mean. What should we expect from nonparametric Bayesian methods?

- Complexity of our model should be allowed to grow as we get more data.
- Place a prior on an unbounded number of parameters.