An Introduction to Bayesian Nonparametric Methods

Rebecca C. Steorts

Bayesian Methods and Modern Statistics: STA 360/601

Module 11

What is a nonparametric model?

1. A really large parametric model.
2. A parametric model where the number of parameters increases with data.
3. A family of distributions that is dense in some large space relevant to the problem at hand.

What is NOT a nonparametric model?

Gaussian model.

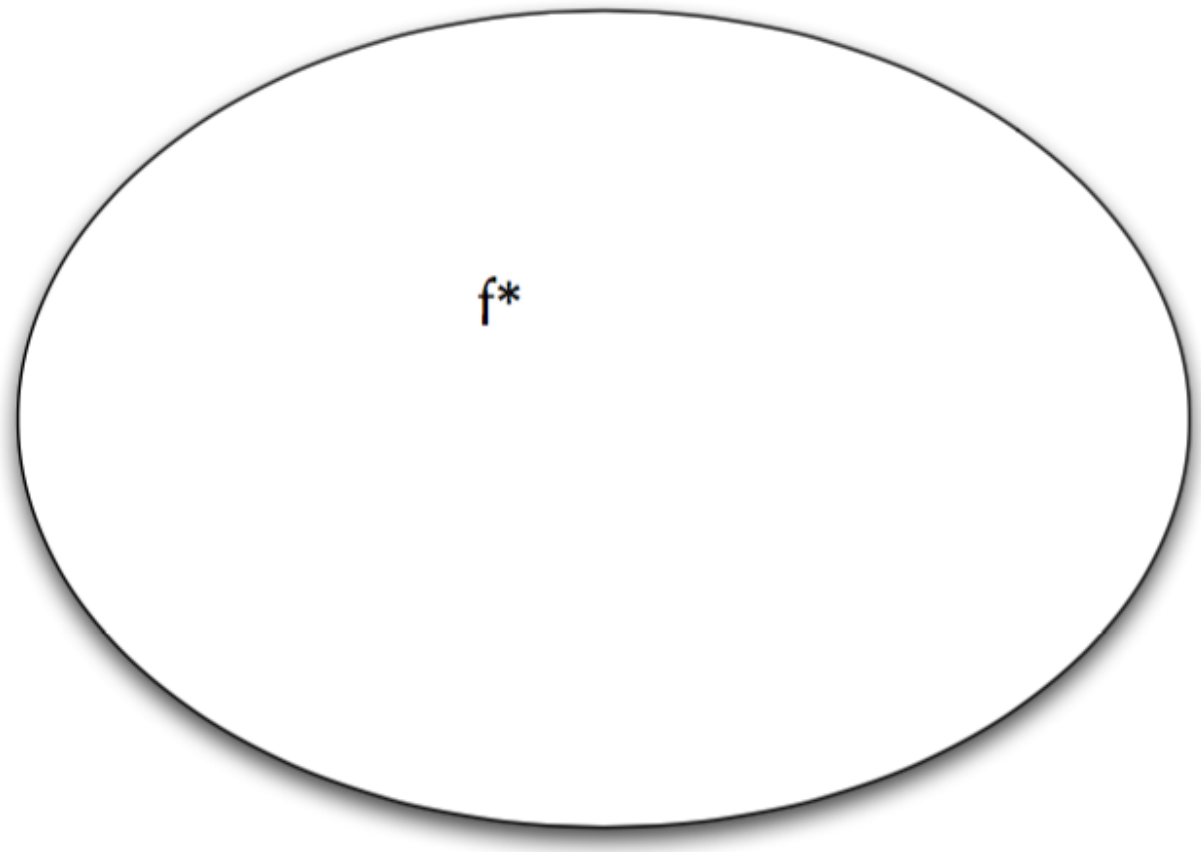
$$X \mid \mu \sim N(\mu, \sigma^2) \tag{1}$$

$$\mu \sim N(\mu_o, \tau^2) \tag{2}$$

Why is this parametric?

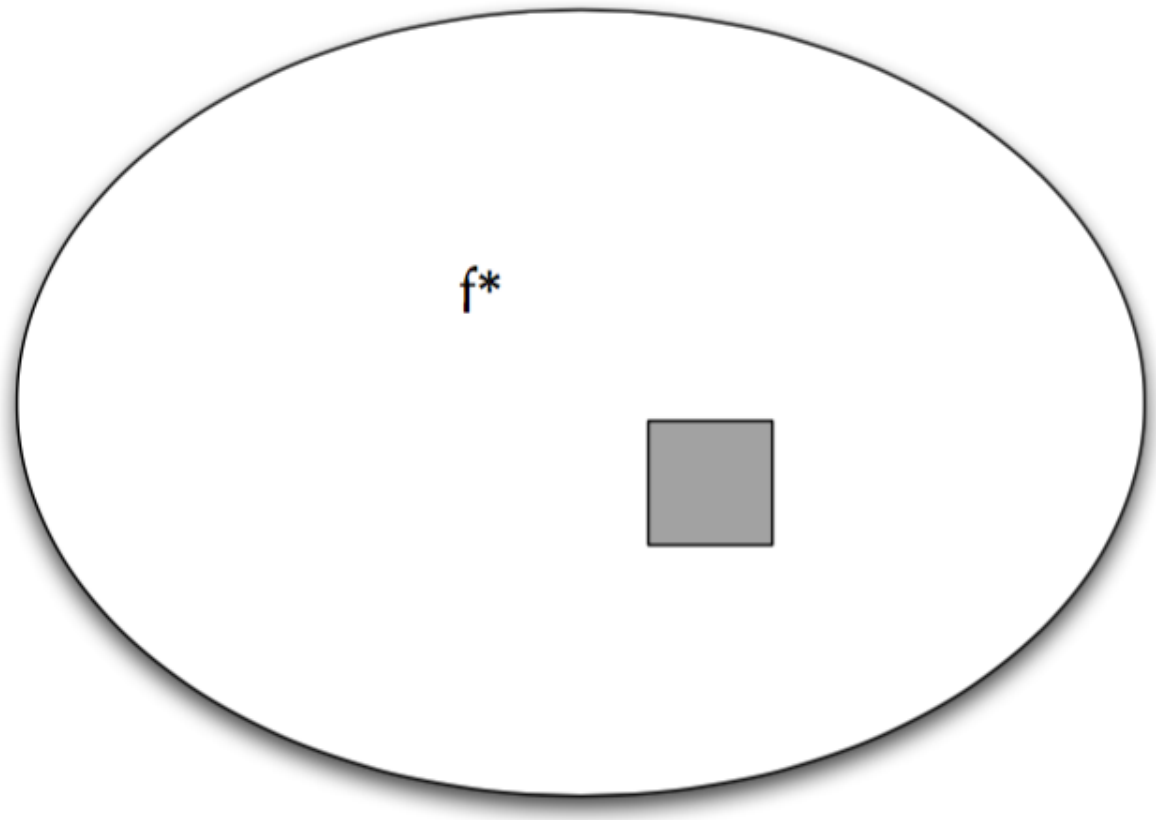
Why is it NOT an NP model?

Bayesian Nonparametrics



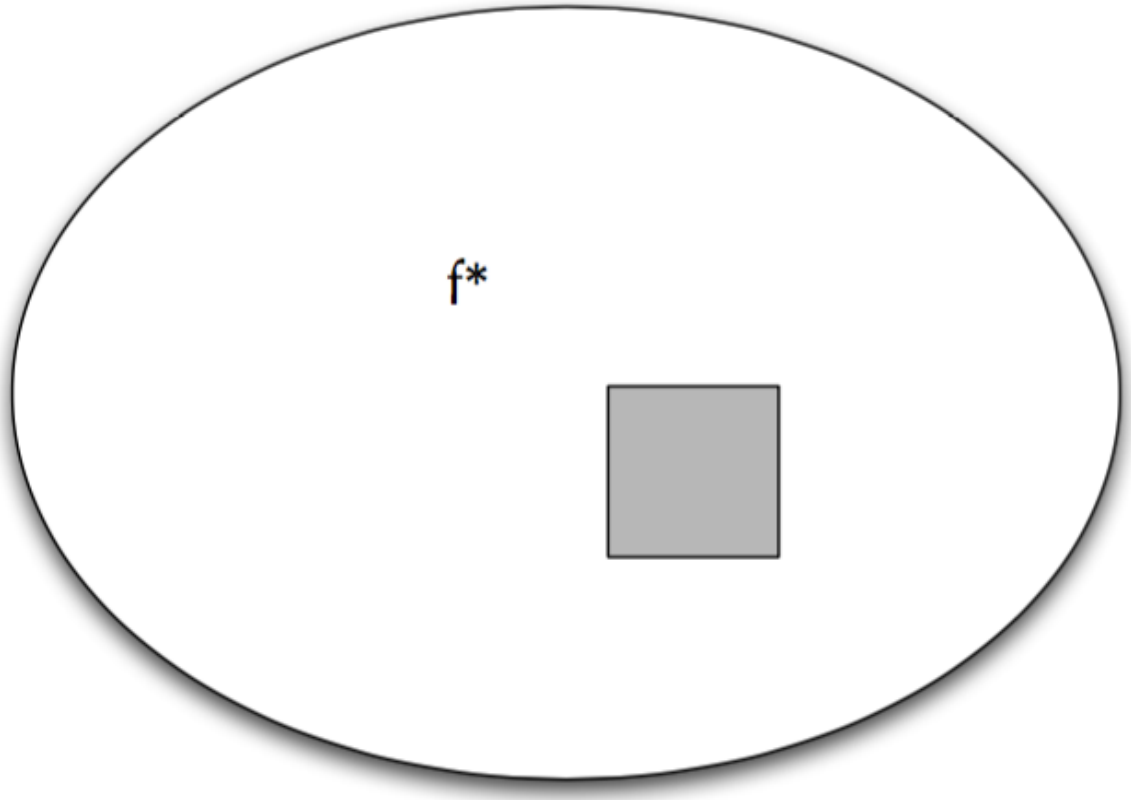
Let's do classification. Have inputs and outputs (binary). What is the true output?

Bayesian Nonparametrics



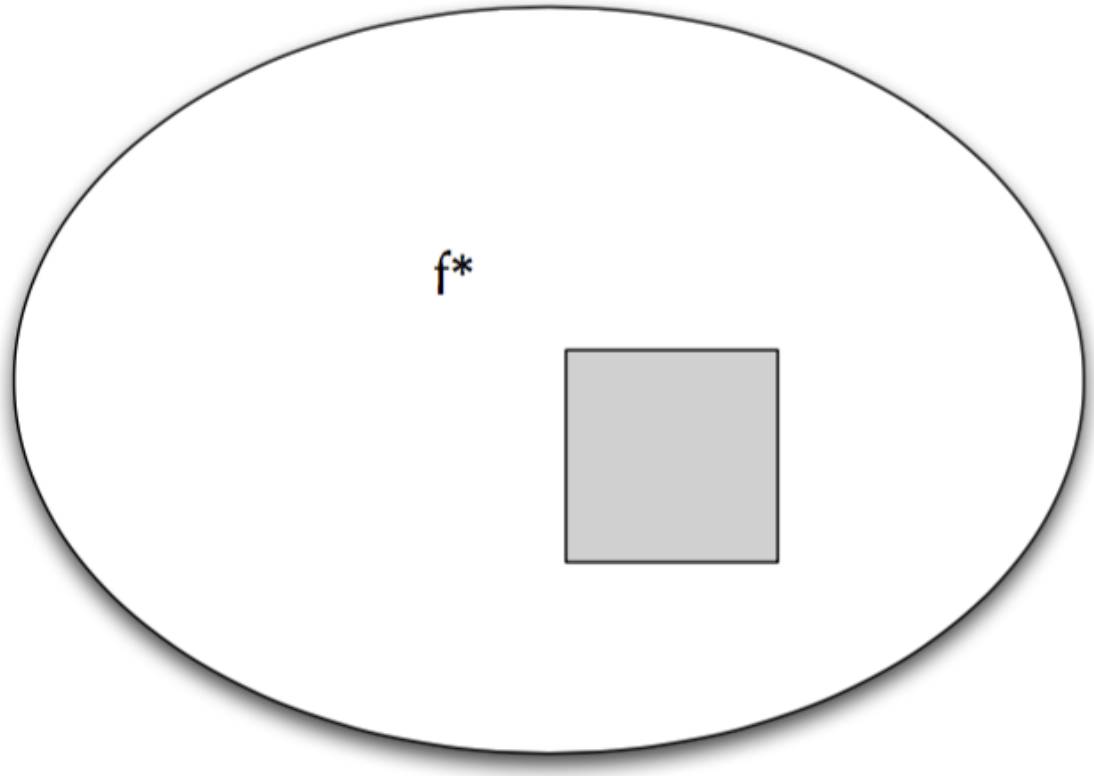
A parametric model is a small part of this space (the grey box)!

Bayesian Nonparametrics



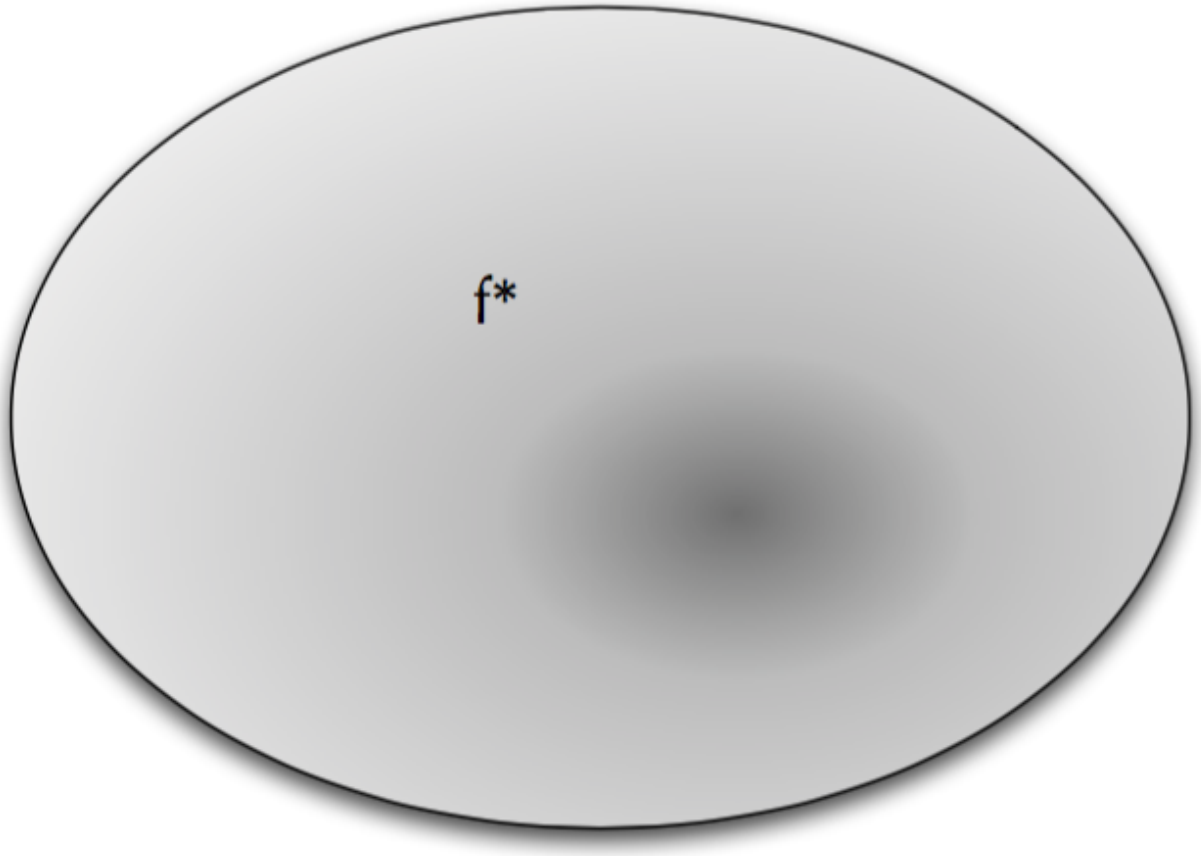
Let's increase the complexity of the model. Unlikely that true function will ever be in the parametric space.

Bayesian Nonparametrics



Let's increase the complexity of the model. Unlikely that true function will ever be in the parametric space.

Bayesian Nonparametrics



Instead, we look at a very reach space for efficient learning. Place higher probability for models that are likely and vice versa.

What is the benefit of BNP?

1. They work well for model selection and averaging. (We won't cover this).
2. They are good at fitting large functional spaces.
3. They are good for structural learning.

Are Nonparametric Models Nonparametric

Nonparametric just means not parametric: cannot be described by a fixed set of parameters.

Nonparametric models still have parameters, they just have an infinite (very large) number of them.

Nonparametric models still make modelling assumptions, they are just less constrained than the typical parametric models.

Issues with Bayesian Nonparametrics

1. Developing classes of nonparametric priors suitable for modelling data.
2. Developing algorithms that can efficiently compute the posterior is important.
3. Developing theory of asymptotics in nonparametric models.

Goal: assign data points $x_1, \ldots x_n$ into clusters $z_1, \ldots, z_K$.

What would be a good application or an example of our goal?

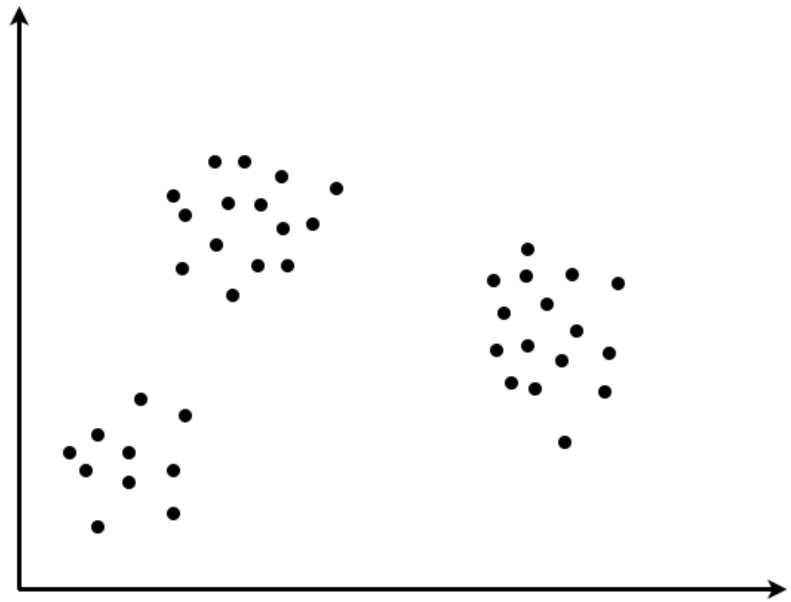
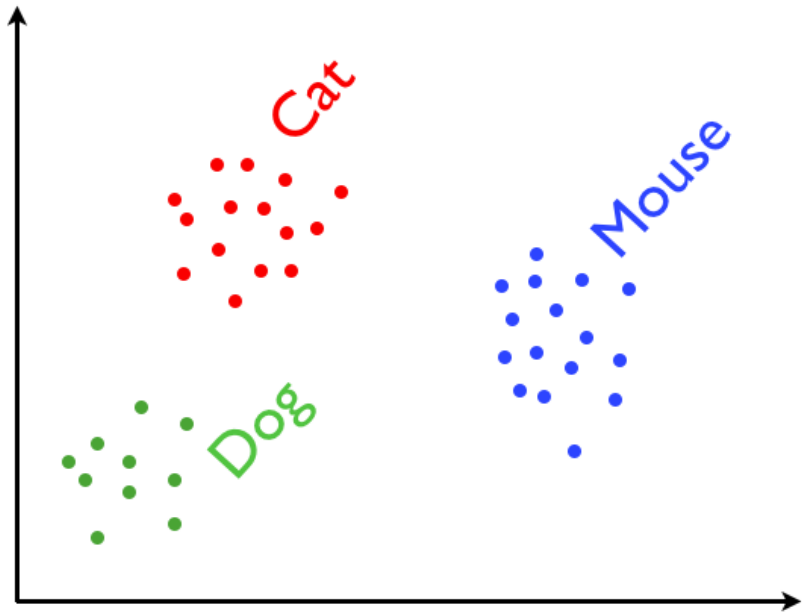


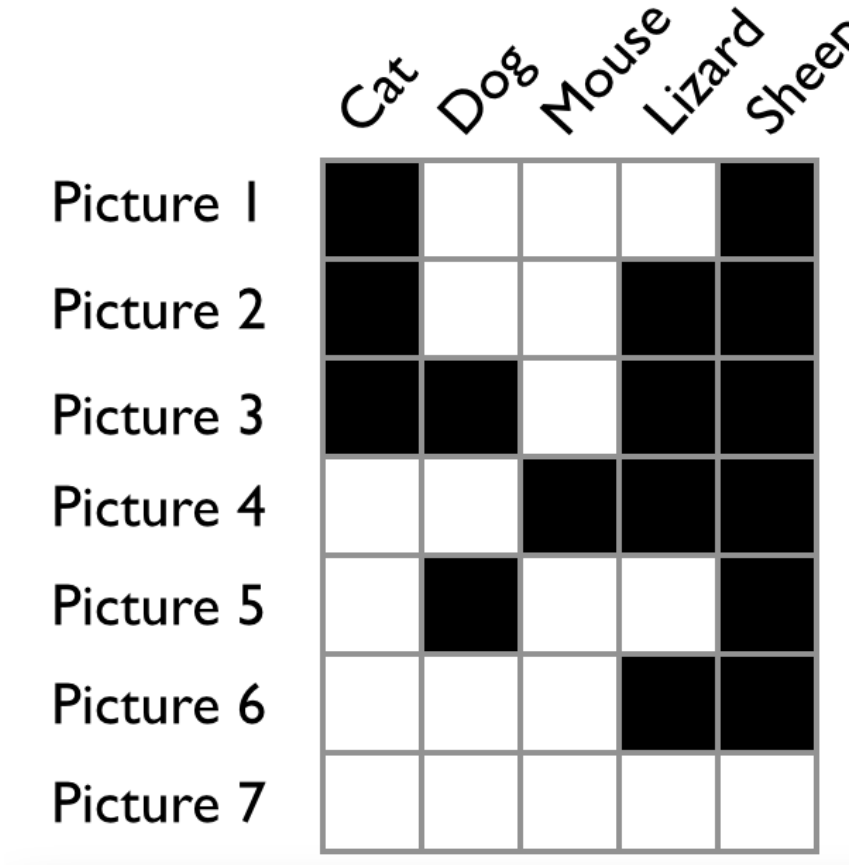
Figure 1: Suppose we have some data points that we want to cluster into groups.

The data points could be classifying data points (pictures of animals) into 3 groups that are dog, cat, and mouse.



For another application, we might cluster students in a class into their majors: math, engineering, statistics, political science, environment, and economics.

Example of what the cluster assignment of pictures (data points) to cat, dog, etc might look like.



Notation

- $x_i, i = 1, \ldots, n$: data points
- $z_i, k = 1 \ldots, K$: cluster assignments
- Note that K and n are fixed and known.

Finite Mixture Models

For each data point i

$$z_i \mid \pi \sim \text{Multinomial}(\pi) \tag{3}$$

$$x_i \mid z_i, \theta_k* \sim F(\theta_{z_i}*) \tag{4}$$

Mixing proportions

$$\pi = (\pi_1, \ldots \pi_k) \mid \alpha \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$$

Cluster k:

$$\theta_k \mid H \sim H$$

Remark: α is fixed and known. How to choose?

Finite Mixture Models

Indicator z_i denotes which component (data point) x_i belongs to.

$$z_i \mid \pi \sim \text{Multinomial}(\pi), k = 1, \ldots, K \tag{5}$$

$$x_i \mid z_i = k, \mu, \Sigma \sim N(\mu_k, \Sigma_k), i = 1, \ldots, n \tag{6}$$

Now let's introduce conjugate priors for the parameters:

$$\pi \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \tag{7}$$

$$\mu_k, \Sigma_k \sim H = \text{N-IW}(0, s, d, \phi), k = 1, \ldots, K \tag{8}$$

[Rasmussen (2000)]

Gibbs Sampling for Bayesian Mixture Models

All conditional distributions are simple to compute.

Let $n_1(z)$ be the number of components (data points) assigned to cluster 1, etc.

$$p(z_i = k \mid -) \propto \pi_k N(x_i; \mu_k, \Sigma_k), \tag{9}$$

$$\pi \mid z \sim \text{Dirichlet}(\alpha/K + n_1(z), \ldots, \alpha/K + n_K(z)), \tag{10}$$

$$\mu_k, \Sigma_k \mid - \ \sim \text{N-IW}(\nu', s', d', \phi') \tag{11}$$

Derivations are left as an exercise.

A computational issue: This Gibbs sampler is not efficient.

Collapsed Gibbs Sampling

The main idea is that we integrate out any variables that we don't need.

We continue with a Gibbs sampling scheme and this speeds up computation. Let's integrate out π, μ, Σ. (Exercise).
Then

$$p(z_i = k \mid -) \propto \frac{\alpha/K + n_k(z_{-i})}{\alpha + n - 1} \times p(x_i \mid x^{(-i)}, n_k(z_{-i})) \quad (12)$$

Interpretation of the three pieces above:

1. α/K: A pseudo count.
2. $n_k(z_{-i})$: Count up the number of times cluster k was chosen among all the others except the ith.
3. $\alpha + n - 1$: The normalizing constant!

Infinite Mixture Models

1. We will take $K \to \infty$.
2. There are at most $n < K$ occupied components, so most components are empty. We can lump these empty components together.

Occupied components:

$$p(z_i = k \mid -) \propto \frac{\alpha/K + n_k(z_{-i})}{\alpha + n - 1} \times p(x_i \mid x_k^{-i}) \tag{13}$$

For clusters that are occupied ($n_k > 0$), things are well defined.

Let K* = the empty clusters and $\{\}$ denotes an empty component.
Empty components:

$$p(z_i = k_{empty} \mid z^{-i}) \propto \frac{\alpha \times \frac{(K-K*)}{K}}{\alpha + n - 1} \times p(x_i \mid \{\}) \tag{14}$$

Infinite Mixture Models

Now let $K \to \infty$.

Occupied components:

$$p(z_i = k \mid -) \propto n_k(z_{-i}) \times p(x_i \mid x_k^{-i}) \tag{15}$$

Empty components:

$$p(z_i = k_{empty} \mid z^{-i}) \propto \alpha \times p(x_i \mid \{\}) \tag{16}$$

But wait – this makes zero sense in general!

- The actual infinite limit of finite mixture models does not make sense: any particular component will get a mixing component of zero.
- In the Gibbs sampler, we got around this by lumping empty clusters together.
- Other ways of making this limit precise:
  - Looking at the prior clustering structure induced by the Dirichlet prior over mixing proportion—Chinese restaurant process.
  - Re-order components so that those with larger mixing proportions tend to occur first, before taking the limit–stick-breaking construction.
- Both are different views of the Dirichlet process (DP).
- DPs: can be thought of infinite dimensional Dirichlet distributions.
- The $K \to \infty$ Gibbs sampler is for DP mixture models.

A Tiny Bit of Measure Theory

In order to formally define some upcoming terms, we need to understand a measure.

Informally, a measure is like a ruler!

On the real line, the measure is the length of a subset of the real line.

We typically can't work with the entire real line and must work with subsets (or families of subsets). [This is known as a σ-algebra]

Then due to this, we will have some properties so that we can apply our measure or ruler.

A Tiny Bit of Measure Theory

- A σ-algebra Σ is a family of subset of a set Θ such that
  - Σ is not empty (we can measure something).
  - If $A \in \Sigma$ measurable set, the $\Theta \setminus A \in \Sigma$
  - If $A_1, A_2, \ldots \in \Sigma$ then $\cup_{i=1}^{\infty} A_i \in \Sigma$.
- (Θ, Σ) is a measure space and $A \in \Sigma$ are the measurable sets.
- A measure μ over (Θ, Σ) is a function $\mu : \Sigma \to [0, \infty]$ such that
  - $\mu(\varnothing) = 0$.
  - If $A_1, A_2, \ldots \in \Sigma$ are disjoint then $\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$
  - Everything that we consider will be measurable
  - A probability measure is one where $\mu(\Theta) = 1$.

A Tiny Bit of Measure Theoretic Probability Theory

If p is a probability measure on (Θ, Σ), then a random variable X taking values in α is simply a measurable function $X : \Theta \to \alpha$.

- Think of the probability space (Θ, Σ, ρ) as a black-box random number generator
- and X as a function taking random samples in Θ and producing random samples in α.

What's really going on here?

A random variable isn't actually random. It's just a measurable function (fixed).

How would you implement a function to generate random draws from a Normal(10,1)? (The only thing at your disposal is a random number generator).

Stochastic processes

A stochastic process is a collection of random variables $\{X_i\}_{i \in I}$ over the same measurable space (Θ, Σ), where I is an index set.

- Stochastic processes are different from other models as that they can be infinite (even uncountably so).
- This raises issues of how do you even define them and how to do you ensure they exist (mathematically speaking).

Stochastic processes form the core of many Bayesian nonparametric models, so they will be useful.

Dirichlet Distribution

A Dirichlet distribution is a distribution of the K-dimensional probability simplex

$$\triangle_K = \{(\pi_1, \ldots, \pi_k) : \pi_k \geq 0, \sum_k \pi_k = 1\}$$

We say that $(\pi_1, \ldots, \pi_k)$ is Dirichlet distributed:

$$(\pi_1, \ldots, \pi_k) \sim \mathsf{Dir}(\alpha_1, \ldots, \alpha_k)$$

if

$$p(\pi_1, \ldots, \pi_k) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_{k-1}}$$

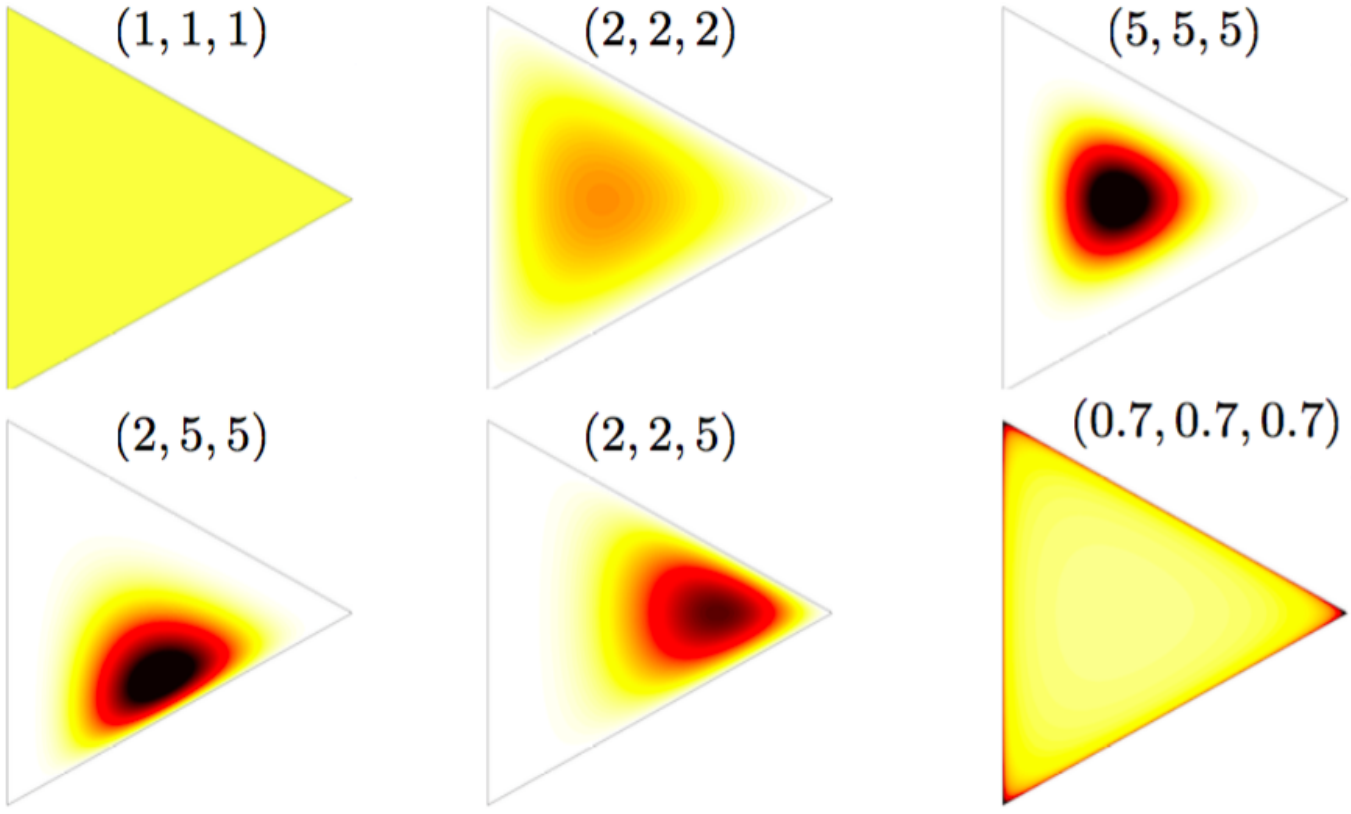


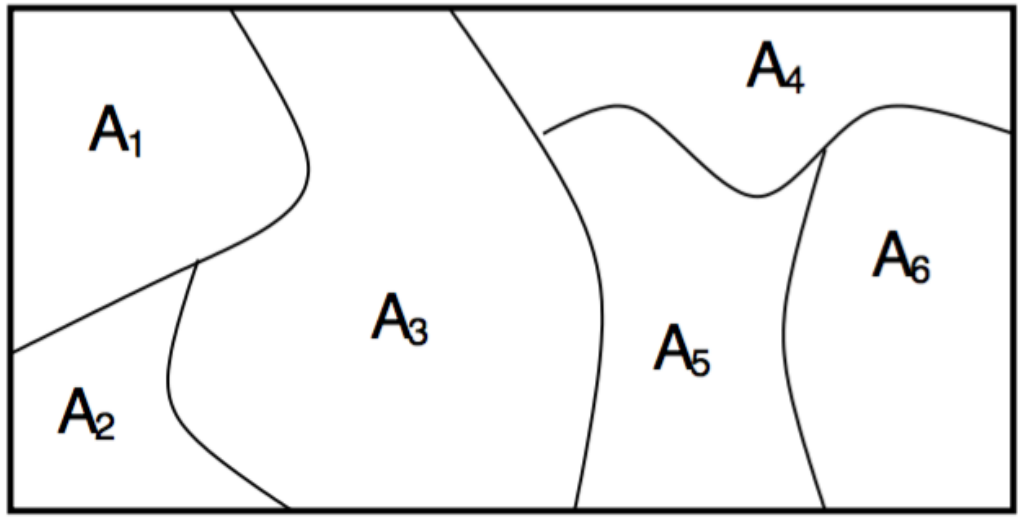
Figure 2: Far left: We get a uniform prior on the simplex. Moving to the right we get things unimodal. On the bottom, we get distributions that are multimodal at the corners.

Dirichlet Process

A Dirichlet Process (DP) is a randomly probability measure G over (Θ, Σ) such that for any finite set of measurable partitions $A_1 \cup \cdots \cup A_k = \Theta$, we have

$$(G(A_1)), \ldots, G(A_k)) \sim \text{Dir}(\alpha H(A_1), \ldots, \alpha H(A_k)) \tag{17}$$

where α is the concentration parameter and H is the base measure.



[Ferguson, 1973], Formal definition, Not constructive.

Dirichlet Process

A Dirichlet Process is a distribution over distributions. Suppose

$$G \sim DP(\alpha, H)$$

$$x_i \mid G \sim G, i = 1, \ldots, N$$

(iid given G)

The Dirichlet-Multinomial conjugacy carries over to the DP:

$$G \mid x_1, \ldots, x_{n-1} \sim DP(\alpha + n, \frac{\alpha H + \sum_i \delta_{x_i}}{\alpha + n - 1})$$

Exercise: Verify that conjugacy holds.

Polya Urn Scheme

$$G \sim DP(\alpha, H) \tag{18}$$
$$x_i \mid G \sim G \quad \text{for all } n \tag{19}$$

Marginalizing out G we get

$$x_n \mid x_1, \ldots, x_{n-1} \sim \frac{\alpha H + \sum_i \delta_{x_i}}{\alpha + n - 1}$$

This is equivalent to

$$x_n \mid x_1, \ldots, x_{n-1} = \begin{cases} x_i^* & \text{with prob } \dfrac{c_n(x_i)}{\alpha + n - 1} \\ \text{new draw from H} & \text{with prob } \dfrac{\alpha}{\alpha + n - 1} \end{cases}$$

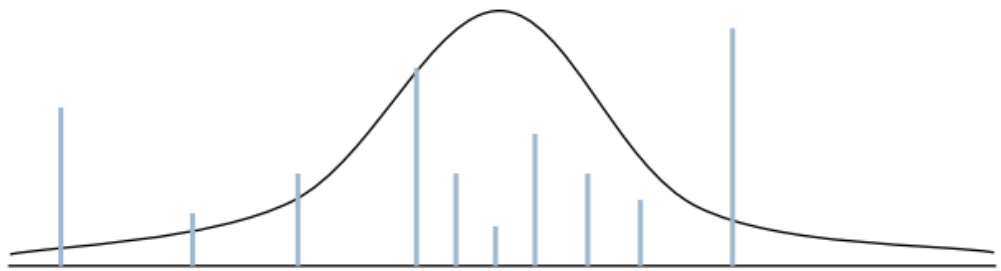
where $c_n(x_i)$: number of customer at table i when customer n makes a move.

Marginalizing out G we get

$$x_n \mid x_1, \ldots, x_{n-1} \sim \frac{\alpha H + \sum_i \delta_{x_i}}{\alpha + n - 1}$$

is called the Polya urn scheme.

Assume that G is a distribution over colors and each X_n represents the color of a single ball placed in the urn.



Marginalizing out G we get

$$x_n \mid x_1, \ldots, x_{n-1} \sim \frac{\alpha H + \sum_i \delta_{x_i}}{\alpha + n - 1}$$

Start with an empty urn. On step i:

1. With probability proportional to α, draw $X_n \sim H$ and add a ball of that color to the urn.
2. With probability proportional to $n - 1$ (number of balls currently in the urn), pick a ball at random from the urn.
   2.1 Record it's color as X_n and return its ball into the urn, along with a new one of the same color.

Repeat for $i = 1, \ldots N$

[Blackwell & MacQueen 1973, Hoppe 1984]

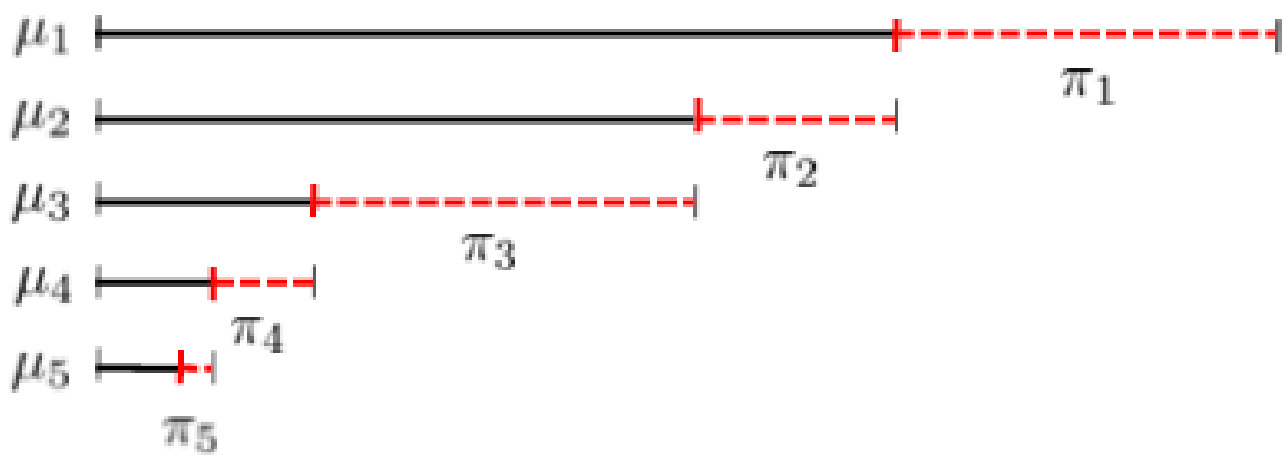
Stick Breaking Construction of the DP

In 1994, Sethuraman developed a constructive way of G, called the "stick breaking" construction.

First assume we have a stick of length 1 (or unit length).

At each iteration, we break part of the stick of with probability π_i.

The proportion remaining at each iteration is μ_i.



V_i is proportion to cut at iteration i. The remaining length is then

$$\prod_{j=1}^{i-1}(1-V_j)$$

And the broken part is

$$\pi_i(\boldsymbol{v}) = V_i \prod_{j=1}^{i-1}(1-V_j).$$

Let's assume that

$$V_i \sim \mathrm{Beta}(1,\alpha)$$

$$G = \sum_{i=1}^{\infty} \pi_i(\boldsymbol{v}) \delta_{X_i} \tag{20}$$

$$V_i, \sim \text{Beta}(1, \alpha) \tag{21}$$

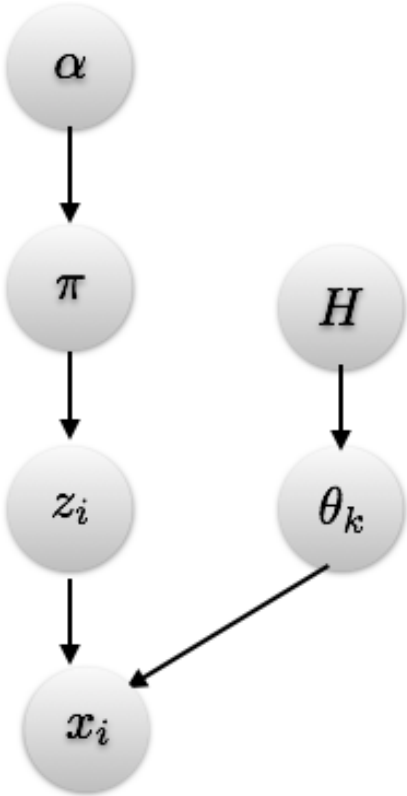
$$X_i \sim H \tag{22}$$

$$f(V_i = v_i \mid \alpha) = \alpha(1 - v_i)^{\alpha - 1} \tag{23}$$

$$\pi_i(\boldsymbol{v}) = V_i \prod_{j=1}^{i-1} (1 - V_j) \tag{24}$$

DP Mixtures

This ties back into Finite Mixture Models and Infinite Mixture Models

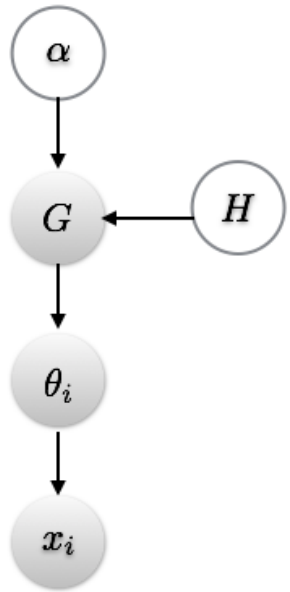


$$G \mid \alpha, H \sim DP(\alpha, H) \tag{25}$$

$$\theta_i \mid G \sim G \tag{26}$$

$$x_i \mid \theta_i \sim F(\theta_i) \tag{27}$$

- θ_i are clustered according to a Polya urn scheme
- Other representation is via by the stick breaking representation.



- Formalize representations for the CRP and Stick breaking.
- There are many other extensions of a DP.
- In lab and in homework you will go through the CRP.
- Remember that the CRP and Stick breaking are special cases of the DP.
- We did not have time to cover inference, but I'm happy to point you to references, etc if you want to read more.