Non-Parametric Bayes:

Bayes for Infinite Dimensional Problems

DS6040 Fall 2024
Teague R. Henry

University of Virginia | School of Data Science

Outline

- What is non-parametric Bayes?
- Infinite Mixture Models via the Chinese Restaurant Process
- Gaussian Process Models

Parametric Bayes

Every model we have talked about this semester is *parametric*

Parametric Model – A model with a fixed number of parameters

- Any given regression model has a fixed number of features (and parameters)
- A finite mixture model has a fixed number of mixtures
- LDA/QDA has a fixed number of classes

Even Bayesian Model Averaging (or other model selection techniques) are parametric

Non-Parametric Bayes

Non-Parametric Models –

- Models that grow/change with increasing amounts of data

Warning: Non-parametric is poorly defined

- It can refer to models that do not assume parametrized distributions
  - Wilcoxian Sign Rank Test

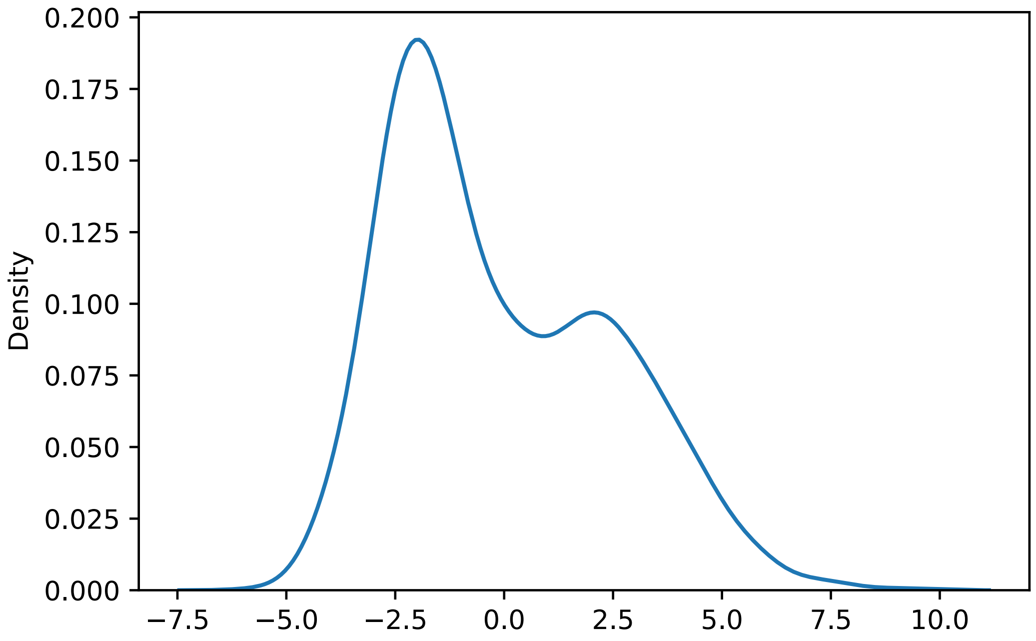
Non-Parametric Bayesian Models are ones where the structure is not specified *a priori.*

Mixture Models and Latent Variables

Data Augmentation – Propose unobserved (latent) variables to explain your data...

Latent variables are defined by the relations you impose between them observed data.

Example – Mixture membership...



Simple Gaussian Mixture Model

Observed data - X_i

Model - $X_i \sim N\left(\mu_{z_i}, \sigma_{z_i}^2\right)$

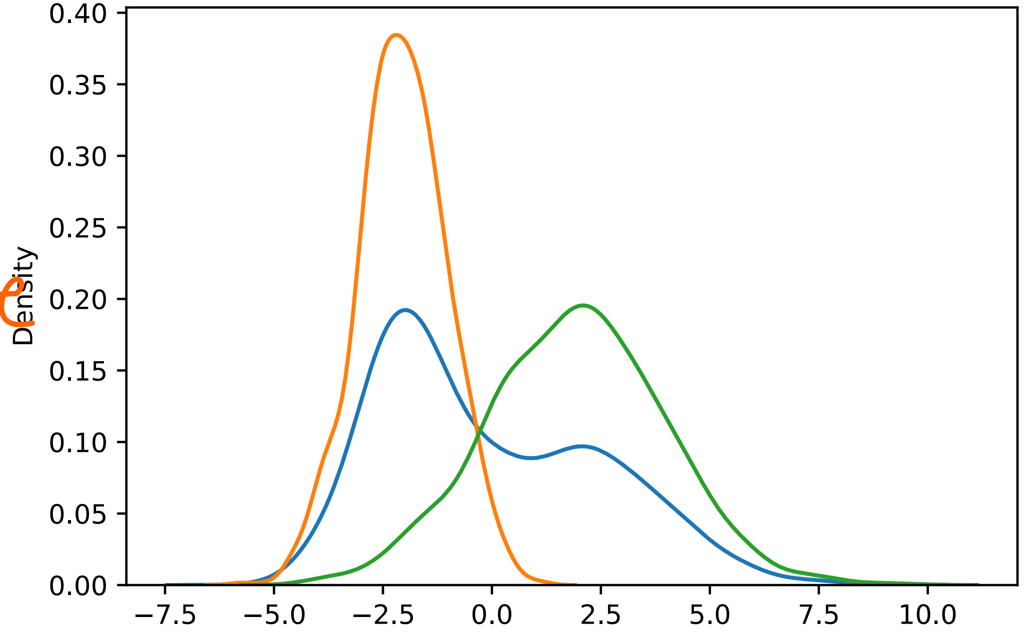
What is z_i?

- Not a parameter…
- Not observed data…

orange & green - mixture models

z_i is a latent variable we made up.

- We are making a strong claim here, that X_i is normal, conditional on z_i.



blue - original density of data

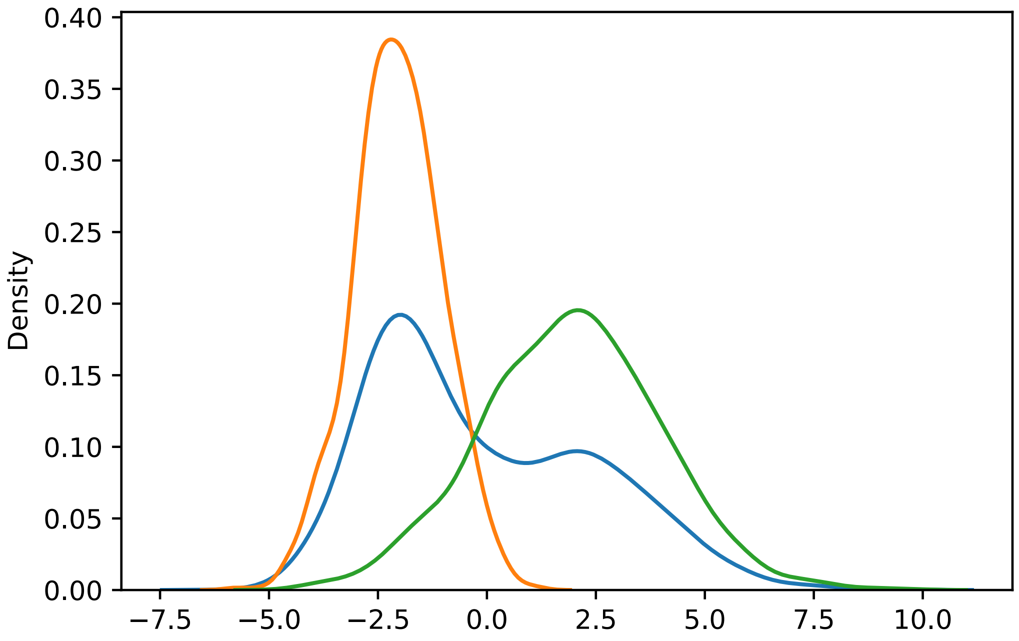
A Word of Warning

Beware clustering, mixture models and unsupervised learning methods…

<u>Reification Fallacy -</u>

Just because you find a cluster / component, doesn't mean it's a real thing…

Sometimes a weird distribution is just that, a weird distribution…



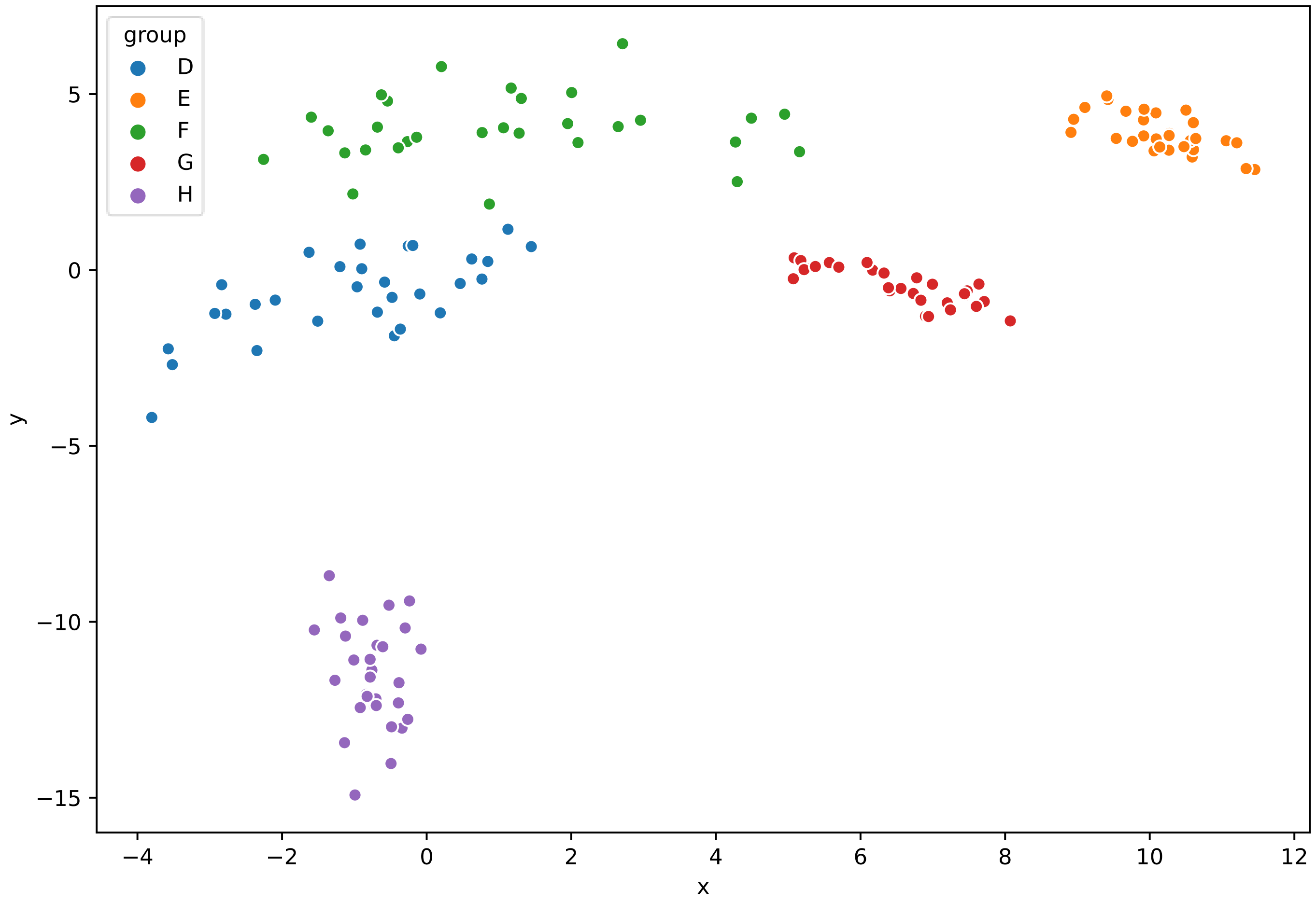
Infinite Mixture Models

Mixture Models –

- Specify K components
- Automatically classify observations into classes/components

How many components?

- Very hard question
- Important to get right



Chinese Restaurant Process

Consider a restaurant with infinite tables and group style seating.

As customers enter, they look around and chose a table:

- An occupied table with probability $\propto$ number of occupants
- The next unoccupied table with probability $\propto \boldsymbol{\alpha}$ "proportional"

This process produces a finite number of occupied tables

- $\boldsymbol{\alpha}$ (concentration parameter) increases number of occupied tables

Stick Breaking Process

$$P(z_i = k) = \sum_{\cancel{i}\,k=1}^{\infty} \beta_k \delta_k(z_i)$$

- δ_k is the Dirac Delta Function (1 when $z_i = k$, 0 otherwise)

How do we ensure an infinite sum sums to 1?

- First, generate $\beta_k' \sim Beta(1, \alpha)$ (note, does not sum to 1)
- Then we can recursively define β_k:

$\beta_1' = .25$	$\overbrace{\beta_2' = .20}^{\text{re-normalized}}$	$\overbrace{\beta_3' = .50}^{\text{re-normalized}}$	$\beta_4' = \ldots$
$\beta_1 = .25$	$\beta_2 = .15$	$\beta_3 = .30$	$\beta_4 = \ldots$

0 ——— 1

Infinite Mixture Model

A mixture model with an infinite number of potential classes…

- With a strong tendency for classes to be empty.

Conceptual difference is in how to model an infinite number of classes?

- Theoretically, class membership is a *Dirichlet Process*
  - *i.e the Chinese Restaurant/Stick Breaking Process*
- Practically, infinity doesn't play nice with computers
- We use truncated infinite mixture models – put an upper limit on occupied classes
- Note that a Dirichlet Process is not the same thing as a Dirichlet Distribution
  - They are related though.

Infinite Mixture Model

$$\boldsymbol{z}|\boldsymbol{w} \sim Categorical_K(\boldsymbol{w})$$
$$\boldsymbol{w} = stickbreak(\boldsymbol{v})$$
$$\boldsymbol{v_k}|\alpha \sim Beta(1, \alpha)$$
$$\alpha \sim Gamma(a, b)$$
$$y_i|z_i, \mu, \sigma \sim Normal(\mu_{z_i}, \sigma_{z_i})$$

Stick breaking prior on the probability of class membership allows the prior class membership to approach 0.

- This means that if you overestimate the number of classes, some classes will be allowed to contain no data. This is a good thing.

Infinite Mixture Model

```
data {
  int<lower=0> K;  // Number of cluster
  int<lower=0> N;  // Number of observations
  real y[N];  // observations
  real<lower=0> alpha_shape;
  real<lower=0> alpha_rate;
  real<lower=0> sigma_shape;
  real<lower=0> sigma_rate;
}
```

high #

hyper parameters

```
parameters {
  real mu[K]; // cluster means
  // real <lower=0,upper=1> v[K - 1];  // stickbreak components
  vector<lower=0,upper=1>[K - 1] v;  // stickbreak components
  real<lower=0> sigma[K];  // error scale
  real<lower=0> alpha;  // hyper prior DP(alpha, base)
}
```

commented out (old version)

controls how many "tables" are filled

```
transformed parameters {
  simplex[K] eta;
  vector<lower=0,upper=1>[K - 1] cumprod_one_minus_v;

  cumprod_one_minus_v = exp(cumulative_sum(log1m(v)));
  eta[1] = v[1];
  eta[2:(K-1)] = v[2:(K-1)] .* cumprod_one_minus_v[1:(K-2)];
  eta[K] = cumprod_one_minus_v[K - 1];
}
```

normalized vector of probabilities

```
model {
  real ps[K];
  // real alpha = 1;

  alpha ~ gamma(alpha_shape, alpha_rate);  // mean = a/b =
shape/rate
  sigma ~ gamma(sigma_shape, sigma_rate);
  mu ~ normal(0, 3);
  v ~ beta(1, alpha);
  for(i in 1:N){
    for(k in 1:K){
      ps[k] = log(eta[k]) + normal_lpdf(y[i] | mu[k], sigma[k]);
    }
    target += log_sum_exp(ps);
  }
}
```

hierarchical prior

stick breaking process

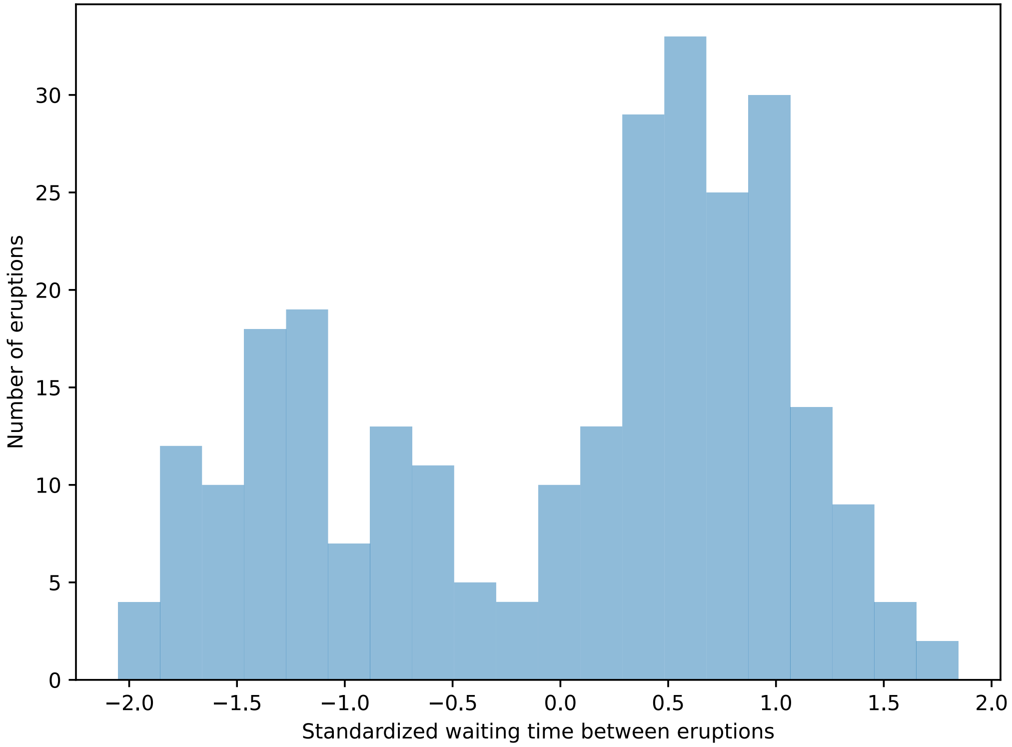
prior probability of class membership

posterior likelihood

Infinite Mixture Model – Old Faithful

Old Faithful – Geyser in Yellowstone

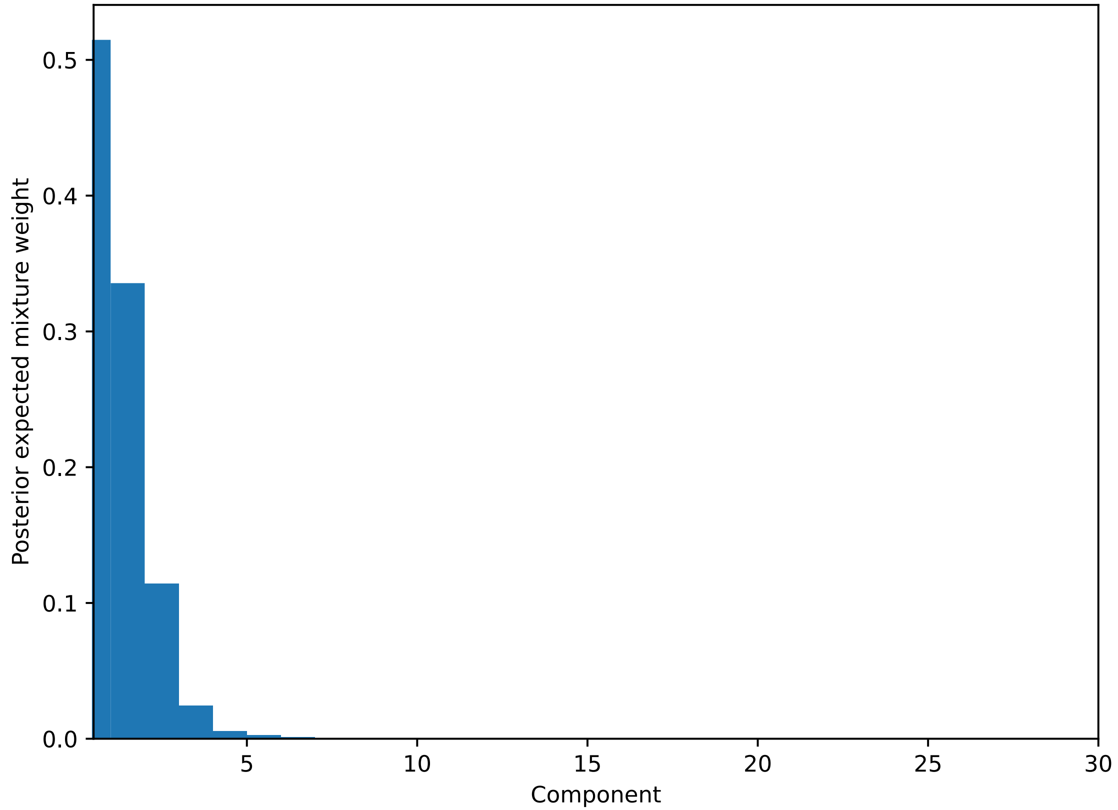
- Model wait times as an infinite mixture



Within cluster, wait times are normally distributed.

- Let’s start with 30 max.

Infinite Mixture Model – Old Faithful



Infinite Mixture Model – Old Faithful

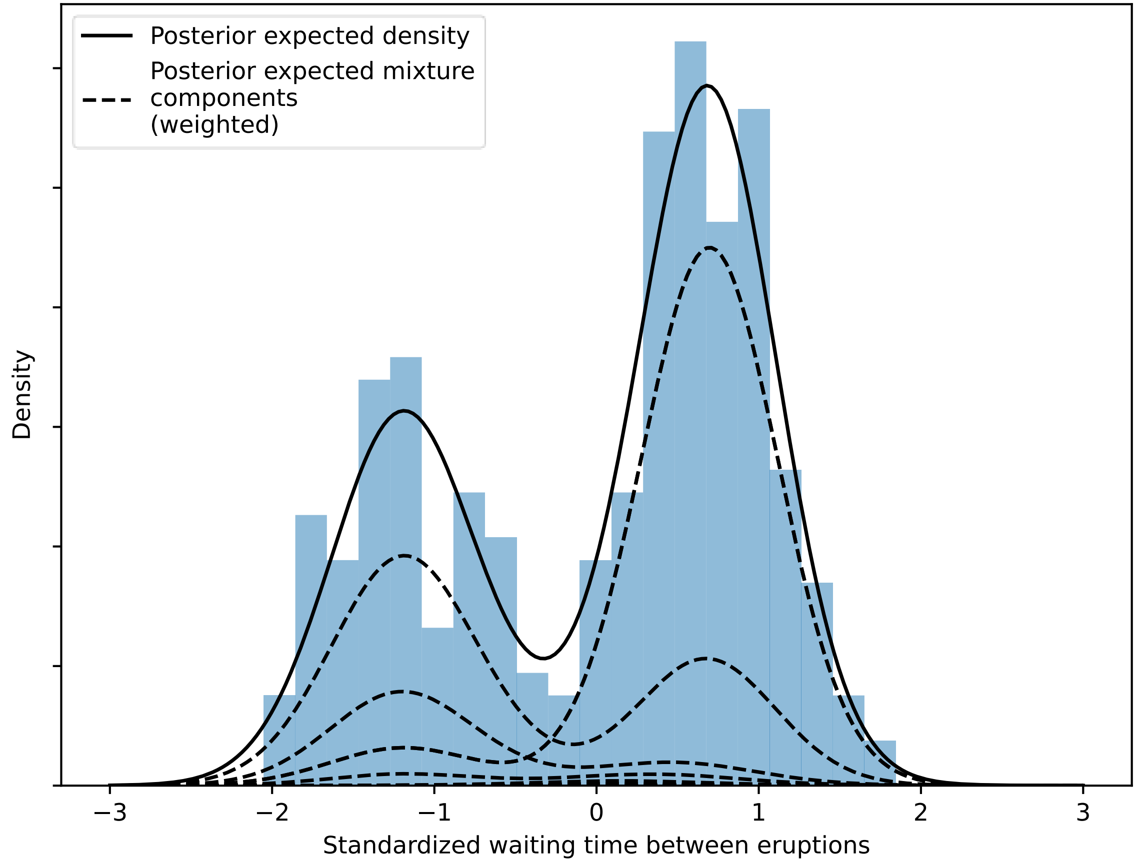
5 components –

- Really 2 components

Ask for infinite components, get infinite components…

- 5 components fit the data best!
- But 2 components are more interpretable…

stick breaking method adds a couple more components to fit our intuitions.
It's important to understand what it means to be a mixture in a mixture model

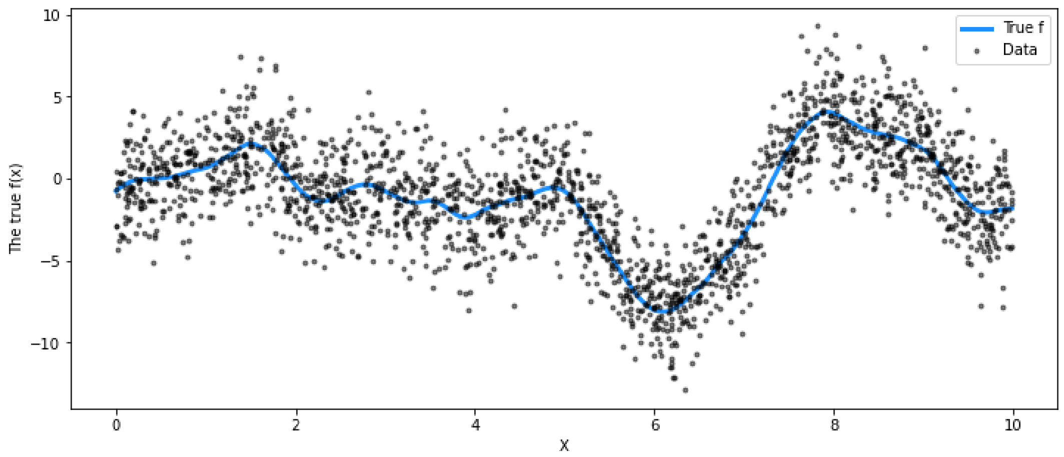


UVA DATA SCIENCE

Gaussian Process Models

Infinite mixture models at a basic level are analogous to finite mixture models. But we can go further…

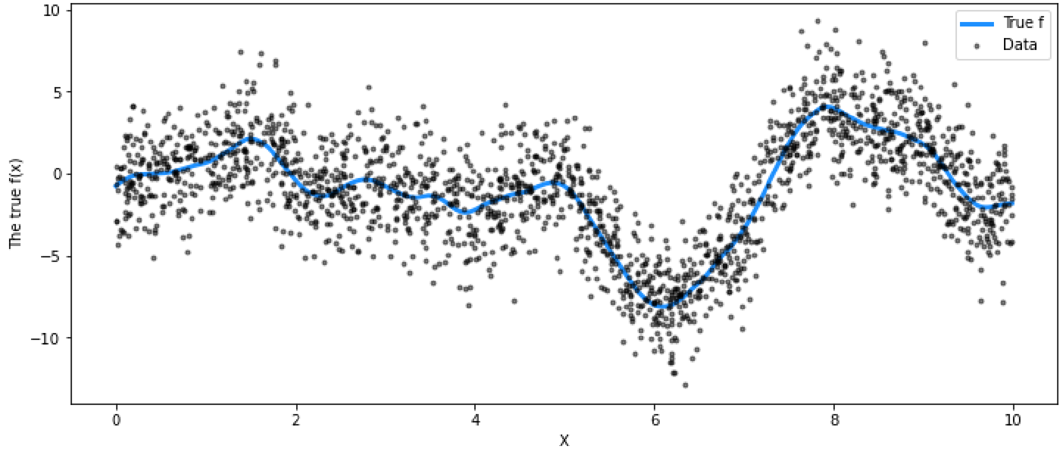
How can we model arbitrary functions of data?



Gaussian Process Models

Gaussian Process –

- For every finite set of indices $\{i\}$ $X_{i_1}, \ldots, X_{i_k}$ is a multivariate normal.
- Equivalent to saying any linear combination of the observed X is univariate Normal.
  - Sort of via the central limit theorem
- <u>**This is not equivalent to saying X_i is normally distributed**</u>
- This Gaussian Process models how X are interrelated (via time or space as the index)



b/c this is a Gaussian process model, not a Gaussian distribution

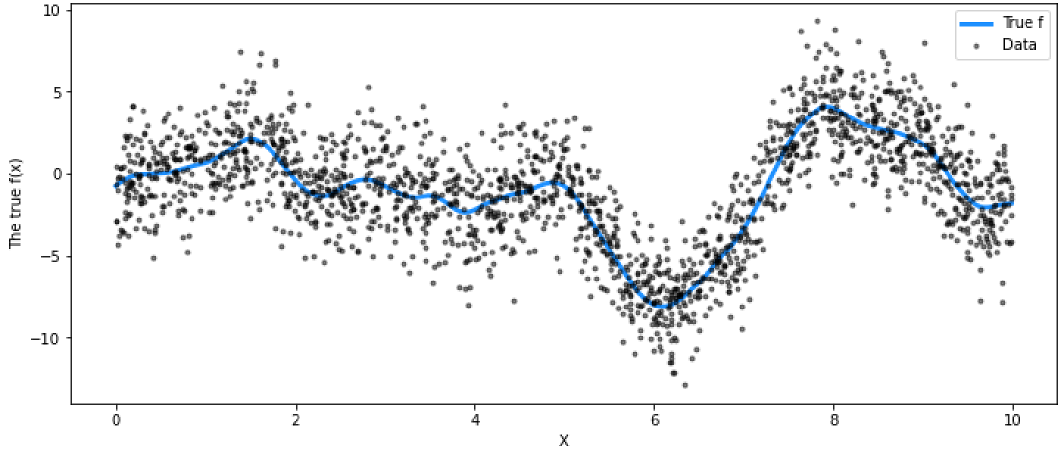
Covariance Kernels

Kernel – A function analogous to an infinite dimensional 2D-matrix.

- $K(x_1, x_2) \rightarrow \mathbb{R}$
- x acts as the index

Covariance Kernel – What is the expected covariance between two observations of X?

- Positive definite kernel – Any finite set of X plugged into the kernel function results in a valid covariance matrix (i.e. positive definite) → a matrix that has a + determinant, among other things. There's also correlations present among points

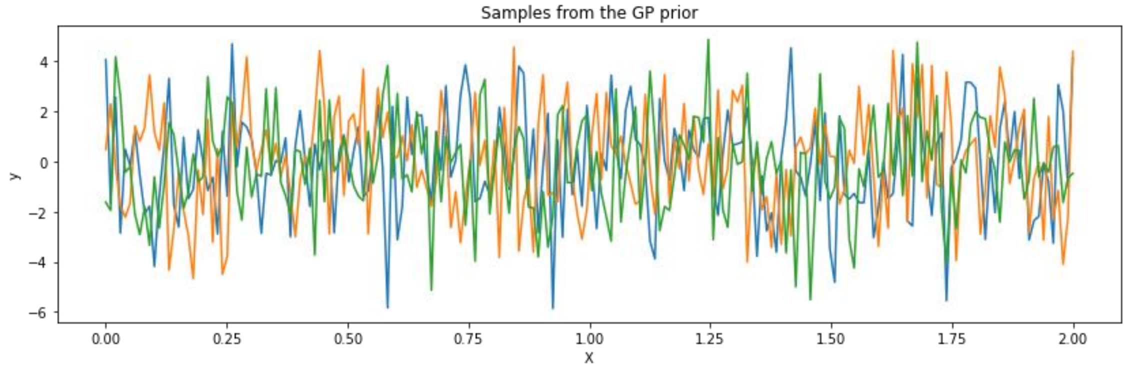


x represents fine points

Covariance Kernels

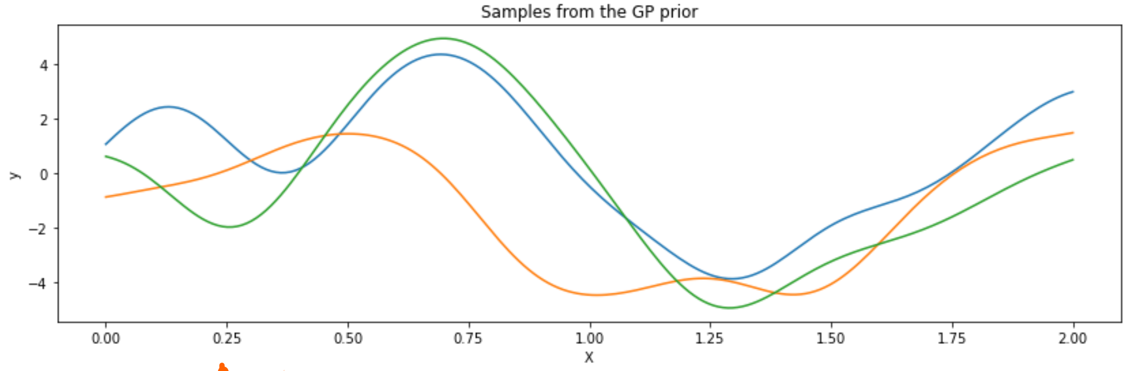
White Noise Kernel –

- $k(x, x') = \sigma^2 \delta_{xx}$
- If $x = x'$ variance is σ^2, else 0.
- No autocorrelation!



Exponentiated Quadratic –

- $k(x, x') = \exp[-\frac{(x-x')^2}{2L^2}]$
- L is a hyperparameter



↳ data based samples

Choosing Covariance Kernels

Covariance kernels define the functional form of your data.

For example:

- Exponential Quadratic kernels allow for smooth trends
- White noise allows for uncorrelated white noise…

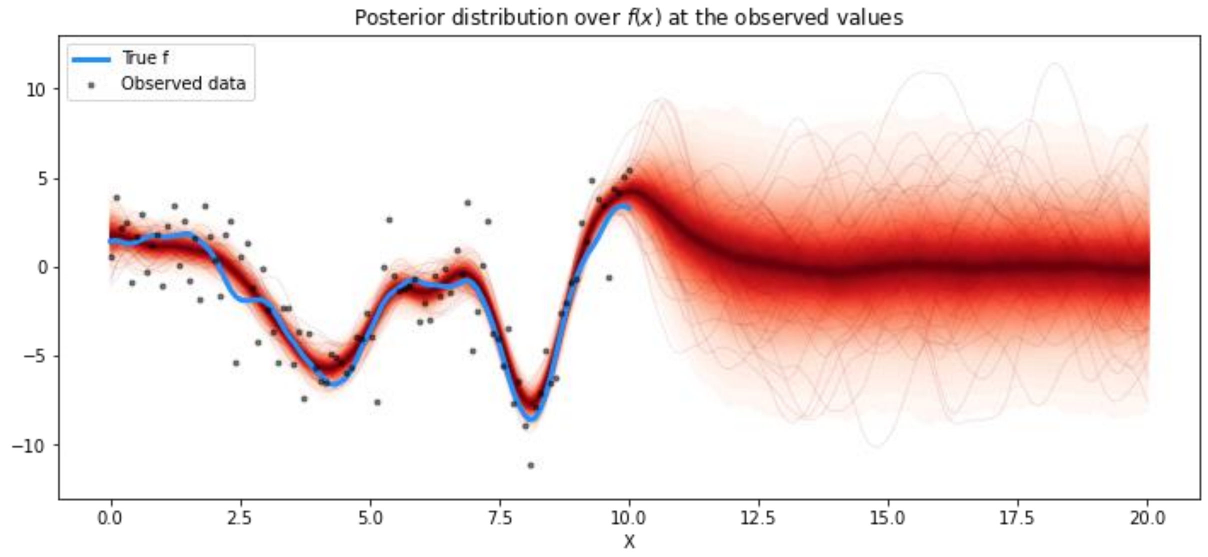
Kernels, because they produce positive definite matrices, can be combined:

- Additively: An exponential quadratic + white noise allows for trend with additional variance around the trend.
- Multiplicatively: Kernels multiplied together result in valid covariance matrices.

What Gaussian Process Models Do

The construction of a GP model is complex, but fundamentally, GP models are approximating complex non-linear functions.

- They can be used for inference, as you can interpret the various kernel estimates, but this is difficult
- They are much better for prediction…



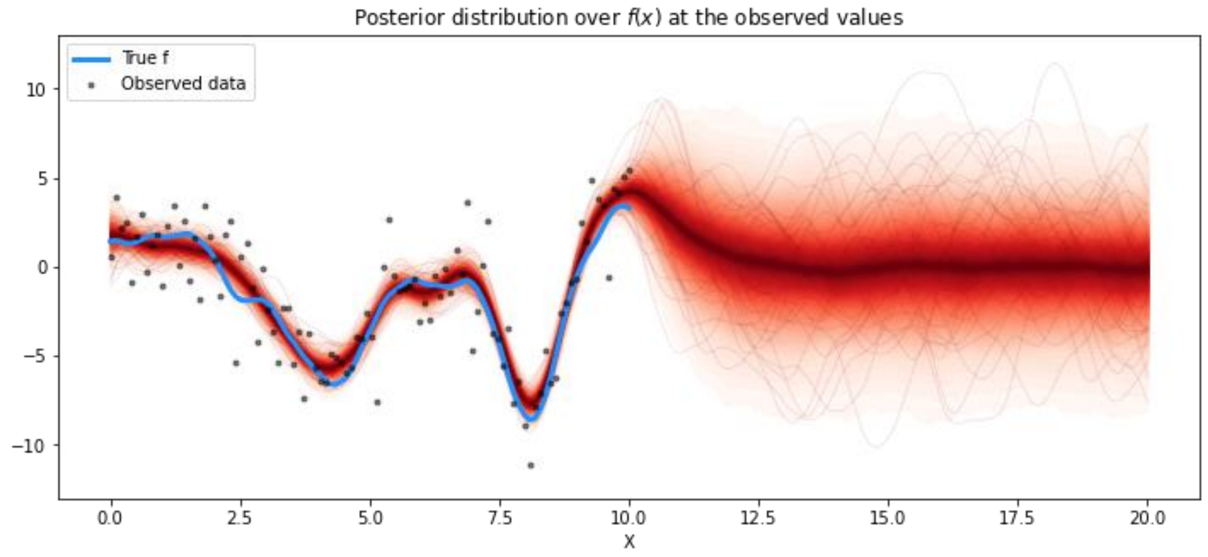
red lines — samples of functions over time

UVA DATA SCIENCE

Gaussian Process Models in Stan

Stan has dedicated GP modules/functions.

- Read the documentation carefully, and work through the math yourself.
- You can specify GP models for non-normal outcomes, Stan has guides for doing this.
- Check carefully for overfit!



we often want more generalizable models w/ decent variance

Summary

Non-Parametric Bayes –

- When the structure of the model needs to be flexible and grow with increasing data.
- Infinite mixture models allow one to determine the number of components rather than setting it a priori
- Gaussian Process Models allow for the modeling and prediction of arbitrary non-linear models
- Inference (interpreting what each parameter estimate means) is difficult with these models
- But prediction is much simpler!