Distribution Estimation

Joe

Introduction

Session Objective

1. Motivate distribution estimation

2. Use parametric and non-parametric methods to estimate the model for some small data sets

This talk borrows greedily from Ryan & Matt, see their talks for different perspectives.

Why Estimation

Observe via webcam busses at: 6:12, 6:15, 6:20, 6:21, and 6:30. you arrive at 6:35, what's the probability you have to wait > 5 minutes?

A friend has an unfair coin, and he challenges you call tails on a series of 10 flips. He has done this with other friends, and you've seen 7, 6, 8, 6, and 5 heads. What is the probability you'll lose money on even odds?

The Path of the Statistician

Observational Data

Use domain knowledge to select
an appropriate model

Fit observational data to your model,
verify the fit is sufficiently accurate.

Use statistical model to make predictions

Question

What are the distributions we would use to model our sample problems?

Random Variables

Randomly Distributed Variables

A random variable is one that obey's probabilistic statements:

$$P(X > 0) = 0.5$$
$$P(-1 < X < 1) = 0.25$$
$$P(X < 0) = 0$$
$$P(X > 1 \mid X > 0) = 0.5$$

Independent and Identically Distributed

A collection of randomly distributed variables are said to be Identically distributed if they have the same probability distribution.

They are said to be independent if the probability of one variable does not effect another:

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

Examples are sequences of coin flips, roulette wheel spins, ect.

Parametric Methods

Method of Moments (MOM)

Recall: All distributions have moments (i.e. mean, variance, skew, kurtosis).

Method of moments:

1. Assume an underlying distribution
2. Compute the relevant sample moments
3. Plug the sample moments into the PMF/PDF of the assumed distribution

I would roughly describe the MOM as the naive approach to statistical modeling

MOM for Example 1

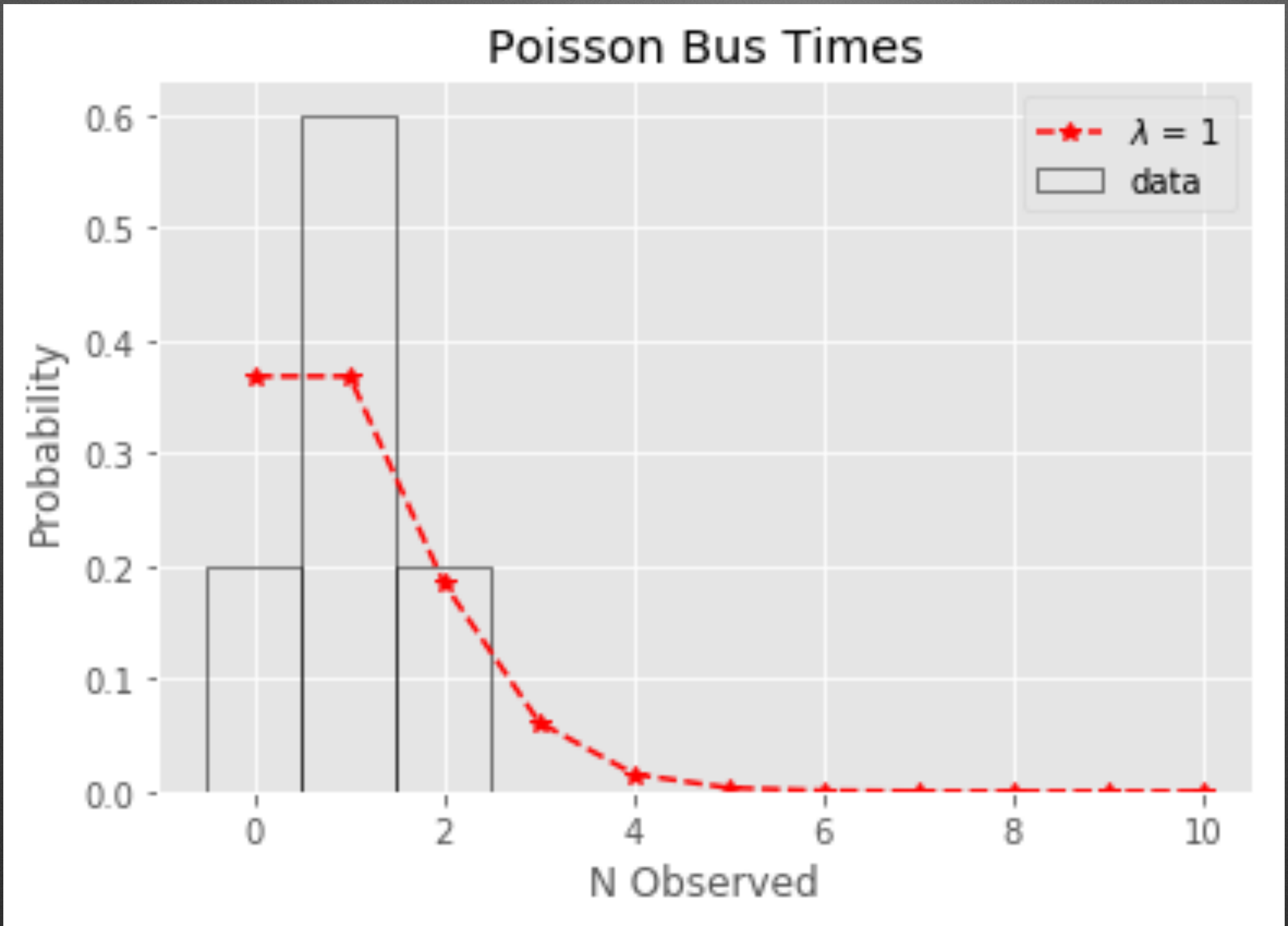
Distribution - Poisson

1 relevant parameter, so we need to do 1 calculation

$$\bar{x} = \frac{1+1+2+0+1}{5} = 1$$

For Poisson, recall $\mu = \lambda$.

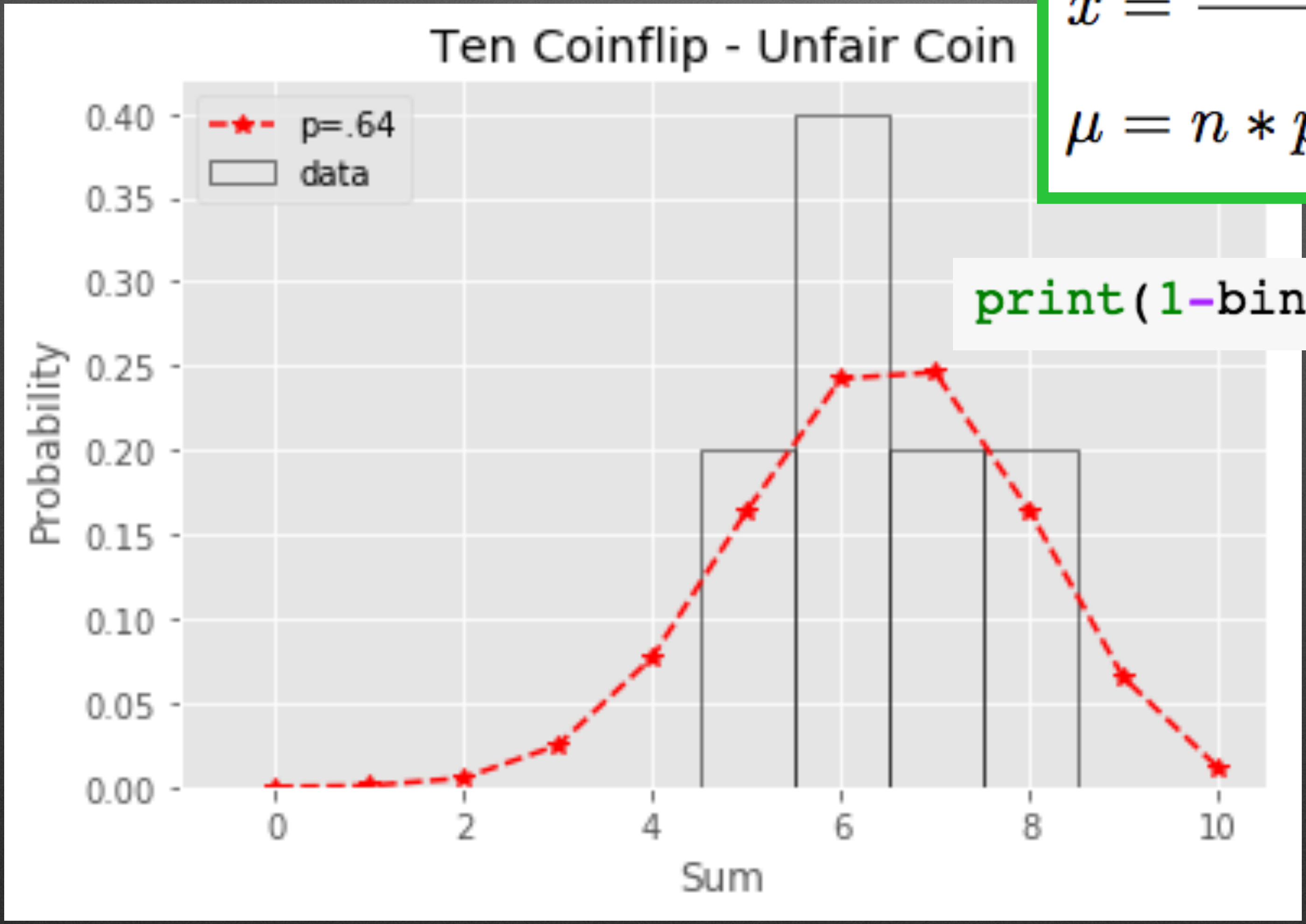
MOM for Example 1 (con't)



P(0) = .368

MOM for Example 2

Distribution - Binomial.



$$\bar{x} = \frac{7+6+5+8+6}{5} = 6.4$$

$$\mu = n * p$$

```
print(1-binom.cdf(5, 10, .64))
```

```
0.729158464263
```

Maximum Likelihood Estimation (MLE)

Recall: Likelihood - a posterior evaluation of probability.

MLE aims to identify the parameters for an assumed probability distributions which maximizes the likelihood of producing the observed dataset.

Maximum Likelihood Estimation:

1. Assume an underlying distribution (like we do in MOM)
2. Define likelihood function
3. Identify parameter set that maximizes the log of the likelihood function

How to find Parameters

$$
\begin{aligned}
\mathcal{L}(\theta|x_1, x_2, \ldots, x_n) &= f(x_1, x_2, \ldots, x_n|\theta) \\
&=^{i.i.d} \prod_{i=1}^{n} f(x_i|\theta) \\
\rightarrow \hat{\theta}_{mle} &= argmax\ log(\mathcal{L}(\theta|x_1, \ldots, x_n)) \\
\rightarrow \frac{\partial log(\mathcal{L})}{\partial \theta_i} &= 0
\end{aligned}
$$

How to find Parameters

$$\mathcal{L}(\theta|x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n|\theta)$$

$$=^{i.i.d} \prod_{i=1}^{n} f(x_i|\theta)$$

Why log?

$$\rightarrow \hat{\theta}_{mle} = argmax\ log(\mathcal{L}(\theta|x_1, \ldots, x_n))$$

$$\rightarrow \frac{\partial log(\mathcal{L})}{\partial \theta_i} = 0$$

Exact solutions are possible

MLE for Example 2

$$X_i \overset{iid}{\sim} Bin(n, p) \qquad i = 1, 2, \ldots, n \qquad f(x_i|p) = \binom{n}{x_i} p^{x_i}(1-p)^{n-x_i}$$

$$\log \mathcal{L}(p) = \sum_{i=1}^{n} \left[\log \binom{n}{x_i} + x_i \log p + (n - x_i) \log(1-p) \right]$$

$$\frac{\partial \log \mathcal{L}(p)}{\partial p} = \sum_{i=1}^{n} \left[\frac{x_i}{p} - \frac{n - x_i}{1-p} \right] = 0$$

$$\hat{p}_{MLE} = \boxed{\frac{\bar{x}}{n}}$$

For the Binomial distribution, MOM and MLE give the same answer!

Check out this paper for the derivations of the parameters via MLE for several distributions (most have the same values). Let's hop in the NB to do a MLE fit 'by hand'

Maximum A Posteriori (MAP)

MLE aims to find $f(x_1, \ldots, x_n|\theta)$.
MAP aims to find θ to maximize $f(\theta|x_1, \ldots, x_n)$.
How you ask? Baye's Theorem!

$$f(\theta|x) = \frac{f(x|\theta) * g(\theta)}{\int_{\theta' \in \Theta} f(x|\theta') * g(\theta') d\theta'}$$

$$\propto f(x|\theta) g(\theta)$$

Maximum A Posteriori (MAP)

MLE aims to find $f(x_1, \ldots, x_n|\theta)$.

MAP aims to find θ to maximize $f(\theta|x_1, \ldots, x_n)$.

How you ask? Baye's Theorem!

$$f(\theta|x) = \frac{f(x|\theta) * \boxed{g(\theta)}}{\int_{\theta' \in \Theta} f(x|\theta') * g(\theta') d\theta'}$$

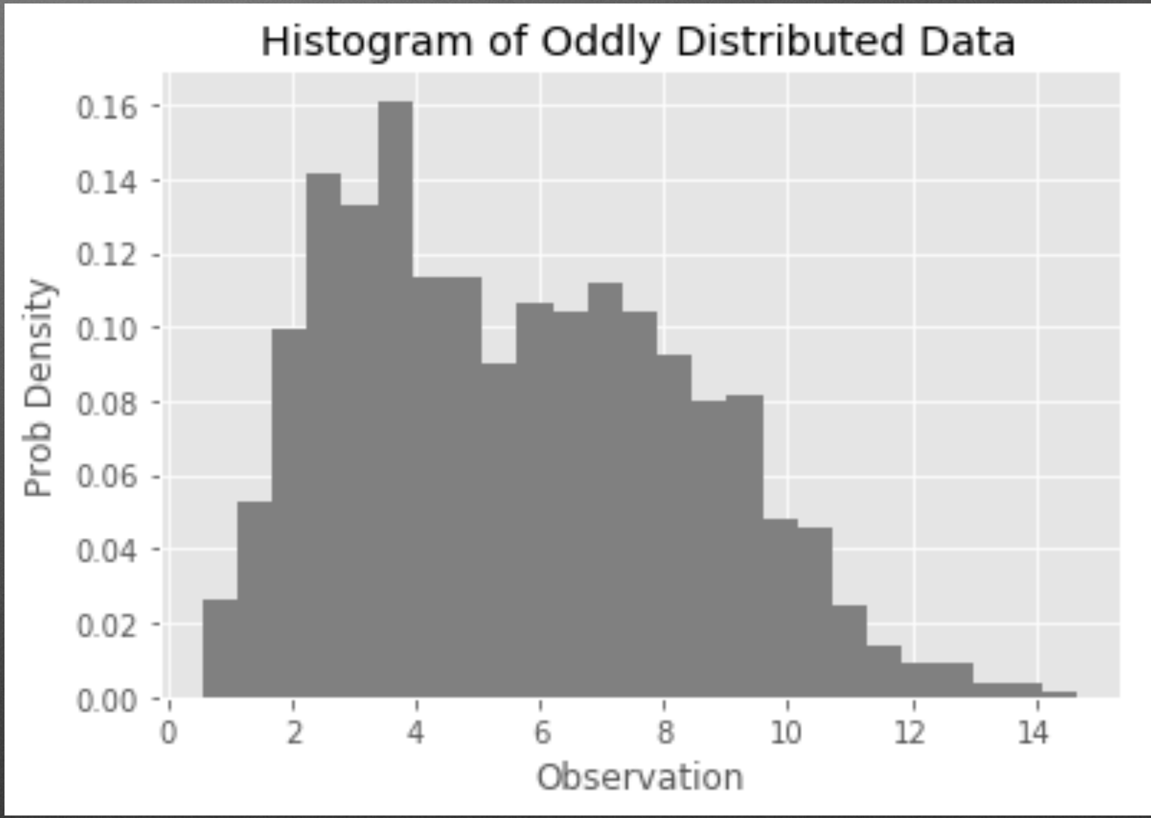
$$\propto f(x|\theta)\boxed{g(\theta)}$$

Bayes' Theorem uses prior beliefs to shape interpreted outcomes.

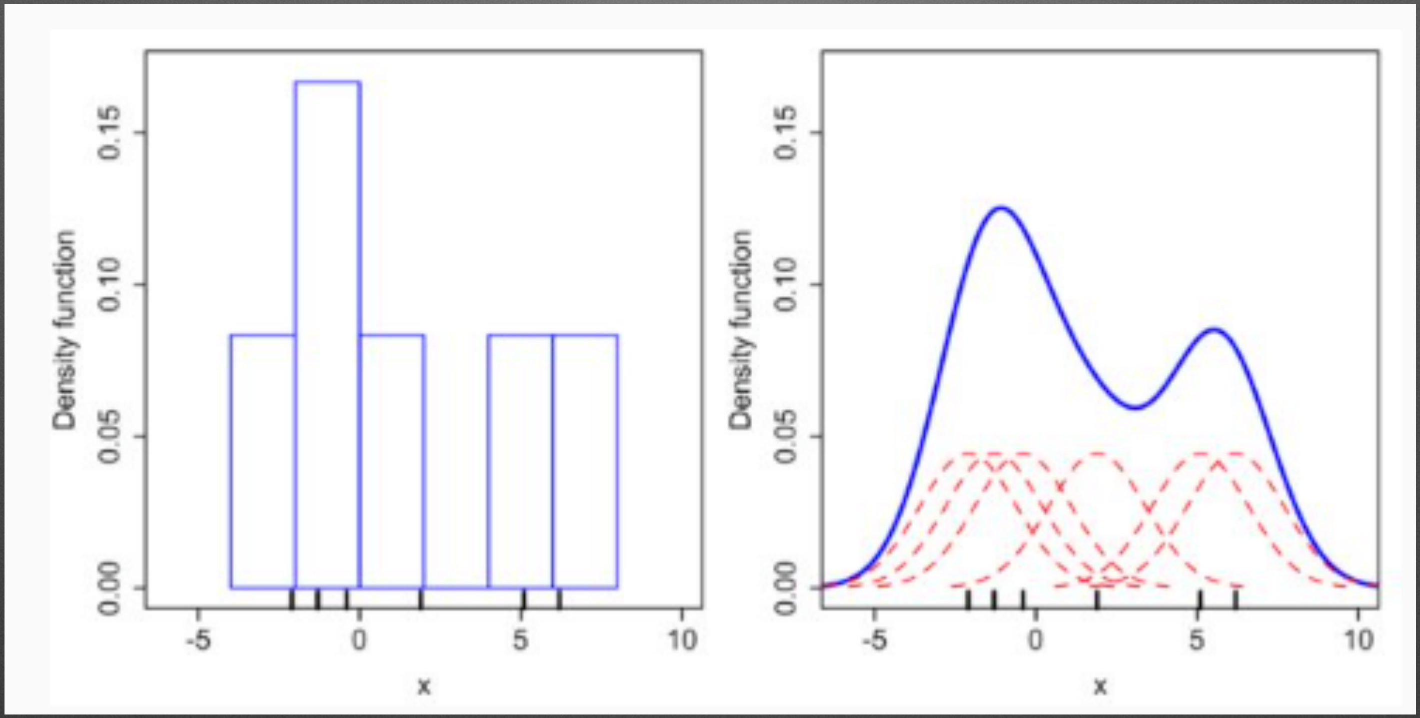
Non-Parametric Methods

Kernel Density Estimation (KDE)

While histograms convey the true shape of the data, it can be problematic to infer what the 'true' shape of the data is due to noise in the data. KDE solves this by applying a smoothing function to observed data.



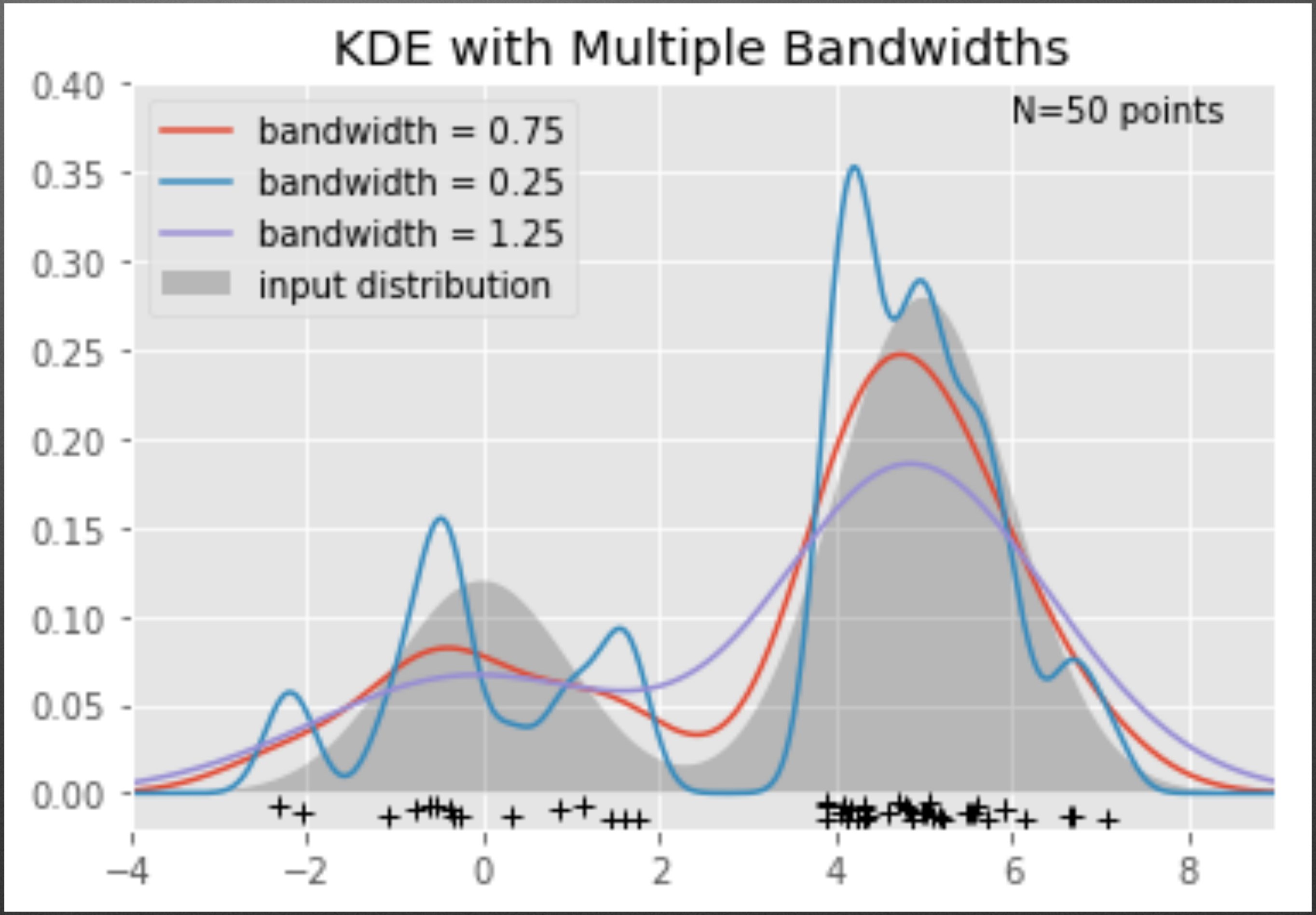
Kernel Functions



Kernel Functions are localized fits to data (for example, gaussians) that are added to form smooth distributions of data.

Kernel Density Estimation (KDE)

KDE controls over/under fitting via a 'bandwidth' parameter.



General Guidlines

Estimating Distributions

Parametric - leverage domain knowledge to select a relevant distribution, and use sample properties to estimate the values of model parameters
Interpretability, Predictive power
Error prone if relevant distribution is unknown

Non-Parametric - don't assume a model, fit a combination of distributions to form an ensemble fit
Can fit oddly shaped data without significant modeling time
Interpretability, Predictive power

Session Objective

1. Motivate distribution estimation

2. Use parametric and non-parametric methods to estimate the model for some small data sets

This talk borrows greedily from Ryan & Matt, see their talks for different perspectives.