1

Today: More on 2-D and 3-D Continuous Data

Sam Ventura

36-315

Today: Defining Contour Plots and Heat Maps

Visualizing High-D Structure

Department of Statistics

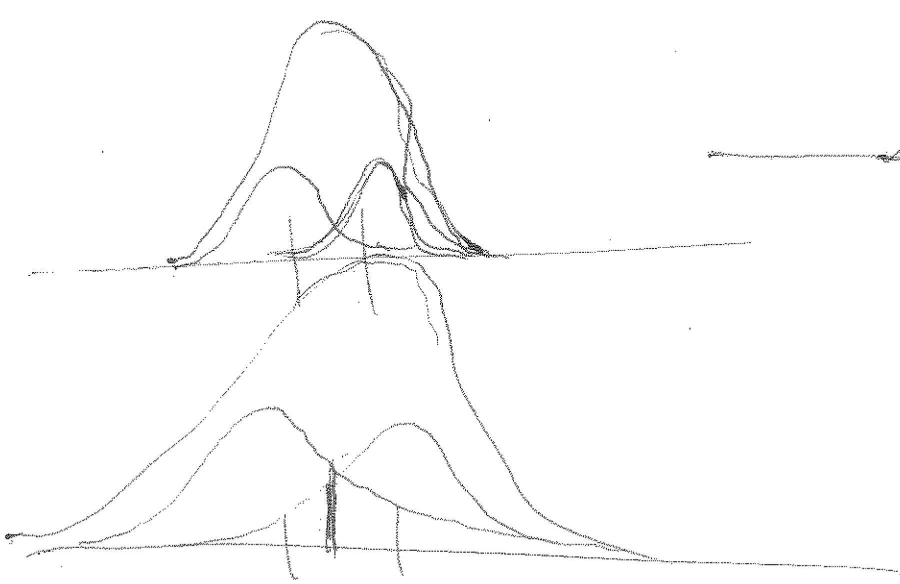
Carnegie Mellon University

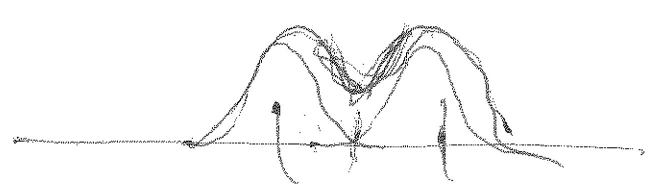
March 1, 2017

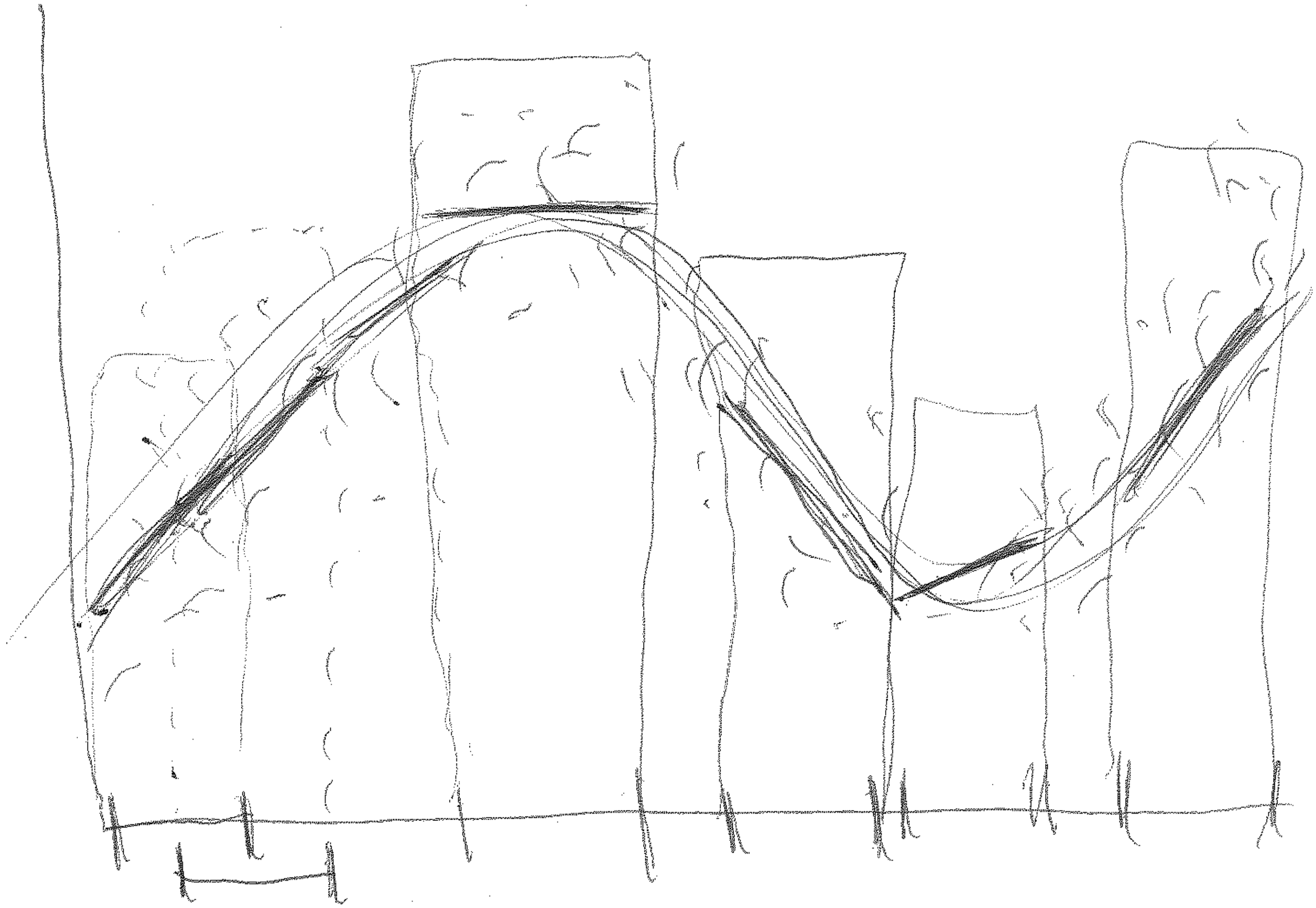
Lab Exam, 1-D KDE, Writing about 2-D Continuous Data

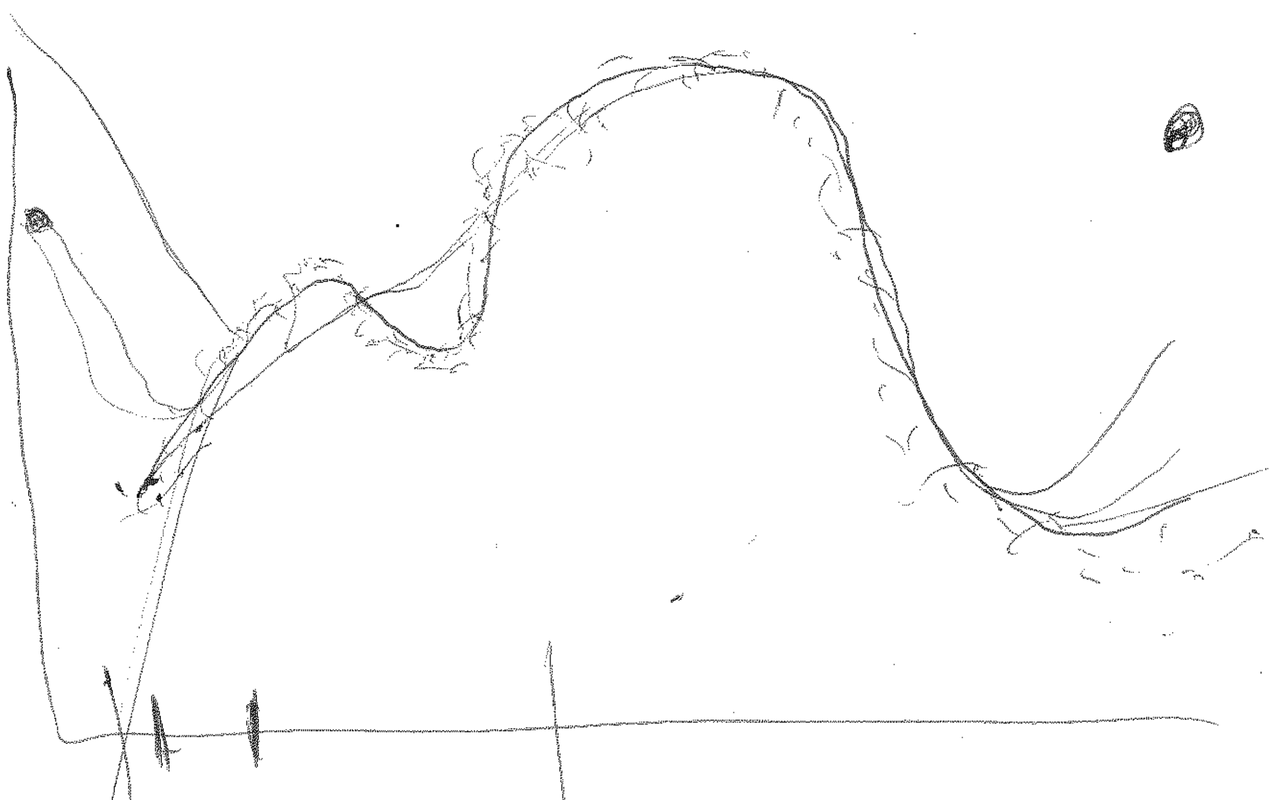
Lab exam timeline:

- Friday: you get the data* and the questions**









$$\hat{f}_X(x) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right) \rightarrow \text{1D KDE}$$

2-D Kernel Density Estimation $\quad X = (X_1, X_2)$

Goal: Estimate the joint distribution of X_1, X_2:

Assuming X_1 and X_2 are independent:

$$\hat{f}(x) = \hat{f}(x_1, x_2) = \frac{1}{n \cdot h_1 \cdot h_2} \sum_{j=1}^{n} K_1\left(\frac{x_1 - X_{j1}}{h_1}\right) K_2\left(\frac{x_2 - X_{j2}}{h_2}\right)$$

$\hookrightarrow$ bw for X_1 $\quad \hookrightarrow$ bandwidth for X_2

Assuming X_1 and X_2 are dependent:

$$\hat{f}(x) = \hat{f}(x_1, x_2) = \frac{1}{n \cdot |H|} \sum_{j=1}^{n} K\left(H^{-1}(x - X_j)\right)$$

$$H = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \rightarrow \text{bandwidth matrix}$$

$\approx$ covariance matrix

Contour Plots

Level Sets:

Contour Plots:

Heat Maps

Visualizing High-D Structure / Projections

What do we do when we have **many** continuous variables?

Projections: Sometimes we want to project the high dimensional data into a smaller subspace without losing "important structure".

Multi-dimensional scaling: looks for a configuration in a k-dimensional subspace such that the distances between observations in the subspace best match the distances in the original p-dimensional space.

Today: 2-D KDE, Contour Plots, Heat Maps, Distance Matrices, Dendrograms

Sam Ventura
36-315

Wednesday: Colors (guest speaker)

Department of Statistics
Carnegie Mellon University

March 20, 2017

2-D Kernel Density Estimation

Goal: Estimate the joint distribution of X_1, X_2:

Assuming X_1 and X_2 are independent:

Assuming X_1 and X_2 are dependent:

$\rightarrow X = (X_1, X_2)$

Contour Plots

$f_X(x) \rightarrow$ probability density function

Level Sets:

$$L(\lambda; f_X(x)) = \{x : f_X(x) > \lambda\}$$

threshold

↳ all areas of our <u>feature space</u> that have density $> \lambda$

↳ where the observations live

"possible values" of our variables X_1, X_2

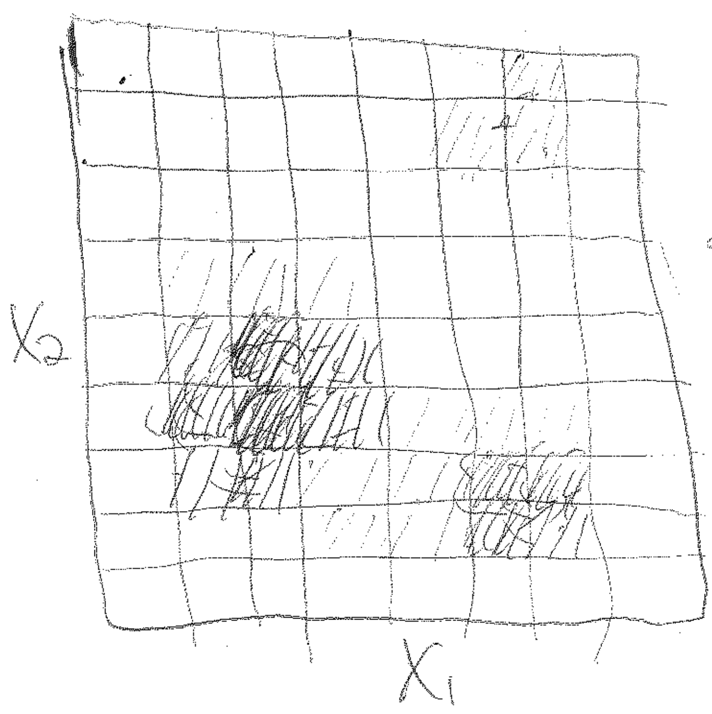
Contour Plots:

$\lambda_0 < \lambda_1 < \lambda_2 < \lambda_3 < \lambda_4$

In ggplot(): geom_tile()

Heat Maps

↳ "birds eye view" of 2D density estimate



- Divide the feature space into a (usually 2-D) grid
- for each <u>tile</u> in the grid, color that tile by the average density in that space (from 2D KDE)
- Use a logical color scale/gradient to represent high/medium/low density areas