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:

Contour Plots

Level Sets:

Contour Plots:

Heat Maps

Visualizing High-D Structure / Projections

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

Example situations when we have many continuous variables:

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

As we will see on the upcoming lab and HW, we can project the data into lower-dim space and visualize the results. Why might this be a bad idea?

Distance = Metric = Distance Metric = Distance Function

Function that defines distance between pairs of observations in a dataset

Properties:

Examples:

Distance Matrices

A **distance matrix** is a data structure that efficiently organizes the pairwise distances between all observations in a dataset.

Pairwise distances are organized into the lower-triangle of a matrix, D

The $(i, j)^{th}$ element of the matrix contains the distance between x_i and x_j:

$D[i, j] = d(x_i, x_j)$

Examples:

Projections: MDS and PCA

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.

Principal Components Analysis: tries to represent large number of correlated continuous variables with a (usually) smaller number of uncorrelated "principal components" (new variables)

Visualizing High Dimensional Structure with Dendrograms

There is no easy way to visualize how far apart observations are in high-dimensional space. One option we do have: **Dendrograms**

Hierarchical Clustering to Obtain Dendrograms

We can get different dendrograms via **hierarchical linkage clustering**.

Single linkage: the distance between two groups is the shortest possible distance between two points, one from each group

Complete linkage: the distance between two groups is the largest possible distance between two points, one from each group