Fall 2018: Introduction to Data Science

GIRI NARASIMHAN, SCIS, FIU

High-Dimensional Space

Points in High-Dimensional Space

APPLICATIONS

- Information Processing
- Search
- Data Mining
- Machine Learning (ML)

Points in d-Dimensional Space

- Assume that d is large
- What is the volume of a unit ball in d-space?
- Are the points well spread out? Where do most of the points lie?
- Assume that we generate n points at random in d-dimensional ball of radius 1
- What can we say about:
  - Distance between any pair of points
  - Angle between any two vectors from origin to that point

Points in d-Dimensional Space

- Let x and y be points in d-space with coordinates from unit-variance Gaussians.
- How far is x from origin
  - Approx distance squared = d
- No prob mass close to O, although prob density has max at O
- Unit ball has zero volume; integral of prob density over unit ball = 0
- $|x - y|^2 = 2d$
  - Thus vectors x and y are approximately orthogonal
- If x is the North Pole, then most of the points lie near the equator

Properties of d-ball and Gaussians

- Area
- Volume
- Denominator > Numerator
- Most of volume is near equator
- Thus any two vectors are nearly orthogonal
- d-dimensional Gaussian
  - Most points are in annulus

$$A(d) = \frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} \quad and \quad V(d) = \frac{2}{d}\frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)}.$$

With probability $1 - O(1/n)$

1. $|\mathbf{x_i}| \geq 1 - \frac{2\ln n}{d}$ *for all* i, *and*
2. $|\mathbf{x_i} \cdot \mathbf{x_j}| \leq \frac{\sqrt{6\ln n}}{\sqrt{d-1}}$ *for all* $i \neq j$.

$$\sqrt{d} - \beta \leq |\mathbf{x}| \leq \sqrt{d} + \beta$$

Nearest Neighbor Search

- Assume you have database of n entries in d-space
- Answer queries of the form
  - Given x, find the nearest neighbor in the database
- Goal: preprocess database so that queries are answered quickly
- Database does not change much, but large number of queries
- Perform expensive preprocessing, but speed up queries
- Time complexity depends on n and d
- Hence the need for **dimensionality reduction**

Projections: R^d to R^k

- Projection of a set of points/vectors along a new vector v is given by:

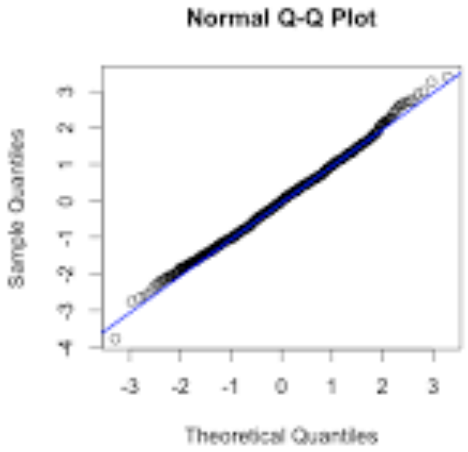
$$f(\mathbf{v}) = (\mathbf{u_1} \cdot \mathbf{v}, \mathbf{u_2} \cdot \mathbf{v}, \ldots, \mathbf{u_k} \cdot \mathbf{v}).$$

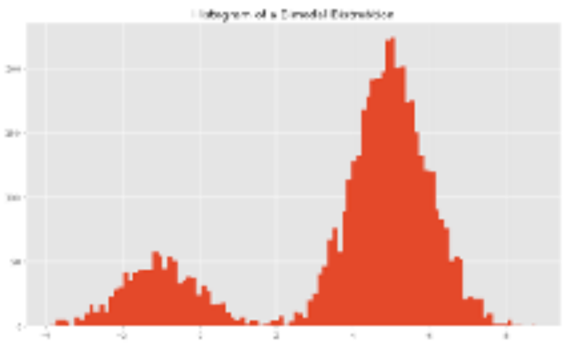
- What happens to distances after projections?
  - Johnson-Lindenstrauss Theorem applies to all pairs
  - 0 < e < 1, any n, k >= (3 ln n)/(ce^2), any vi, vj, with pr 1-1.5/n:

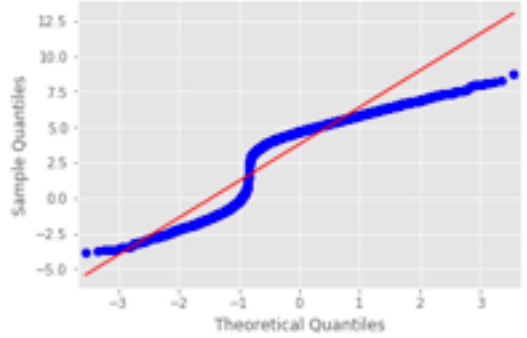
$$(1-\varepsilon)\sqrt{k}\,|\mathbf{v_i} - \mathbf{v_j}| \leq |f(\mathbf{v_i}) - f(\mathbf{v_j})| \leq (1+\varepsilon)\sqrt{k}\,|\mathbf{v_i} - \mathbf{v_j}|.$$

Test for Normality

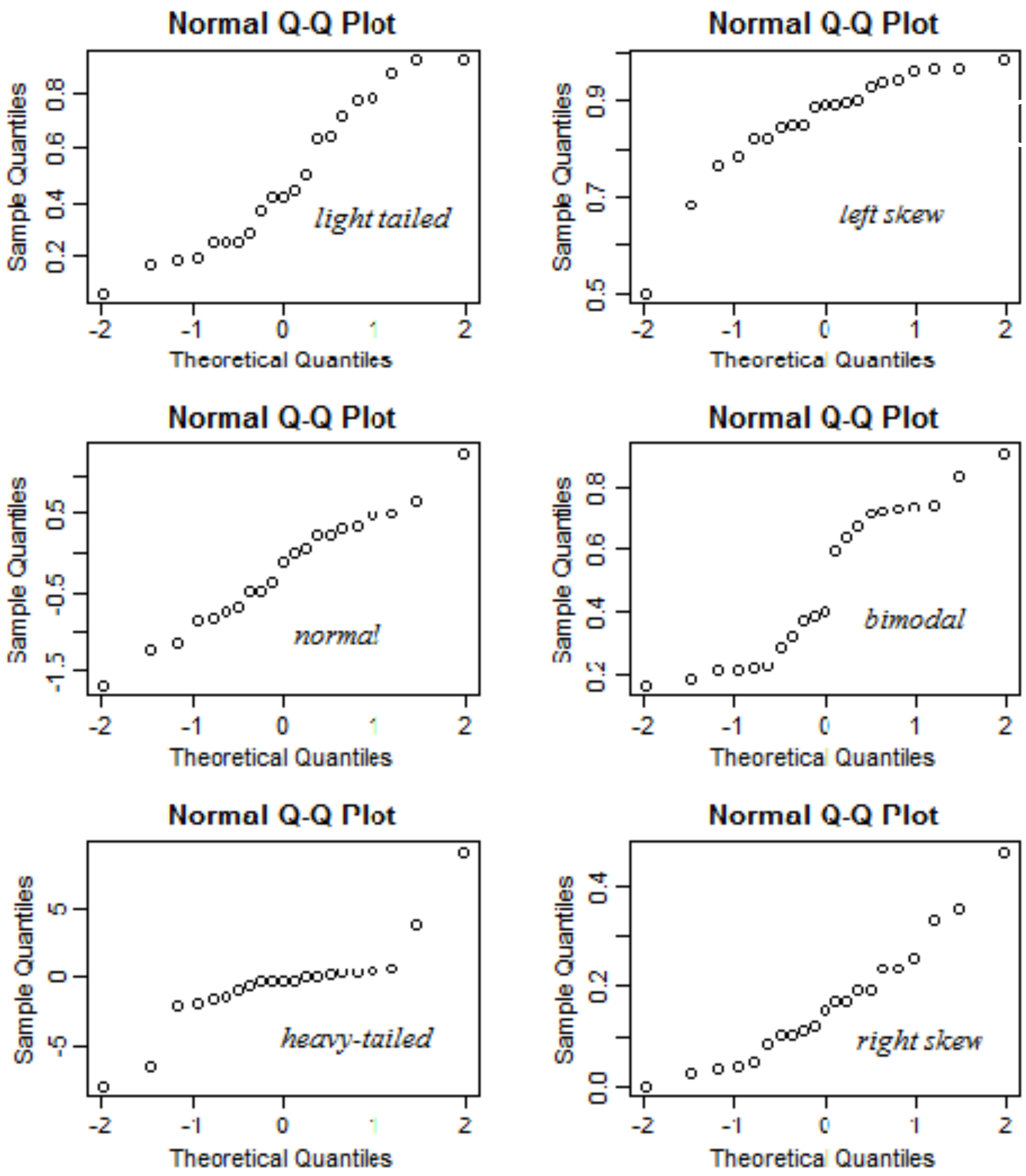
- Informal: plot a histogram and see if it is bell-shaped
- Graphical approach: Do a quantile-quantile (QQ) plot







QQ Plots & Interpretations



Tests for Normality … 2

- D'Agostino's K-squared test,
- Jarque–Bera test,
- Anderson–Darling test,
- Cramér–von Mises criterion,
- Lilliefors test,
- Kolmogorov–Smirnov test,
- Shapiro–Wilk test, and
- Pearson's chi-squared test.

- Some Bayesian approaches exist as well