Multivariate Nonparametric Tests

Christof Seiler

Stanford University, Spring 2016, Stats 205

Overview

- So far, we have seen only univariate nonparametric tests
- Today, we'll cover multivariate generalizations
- Two-sample tests
  - Data depth-based: Tukey depth function
  - Graph-based: Friedman and Rafsky test

Data Depth-Based Two-Sample Tests

- In univariate nonparametric analysis, we relied heavily on ranks
- Ranks are straightforward in the univariate case
- We just use the natural ordering of observations along the real line
- Moving from univariate to multivariate setting, we need to make some more considerations
- In $\mathbb{R}^d$ there is no natural ordering
- Just a straightforward extension of the median to define a center can fail
- A $\mathbb{R}^d$ coordinate-wise median can lie outside the convex hull of the data

Data Depth-Based Two-Sample Tests

- The usual ranks:
  - We ranked n observations in ascending order
  - From that we constructed test statistics
  - For instance, the median is defined as the order statistics of rank $(n+1)/2$ (when n is odd)
  - The median can be computed in $O(n)$ time
  - The problem is that generalizing this to higher dimension is straightforward
- So we consider a different ranking system
- We rank observations as assigning
  - the most extreme observation depth 1
  - the second smallest and second largest observations depth 2
  - Until we end up with the deepest observation, the median
- This can be extended to higher dimensions more easily

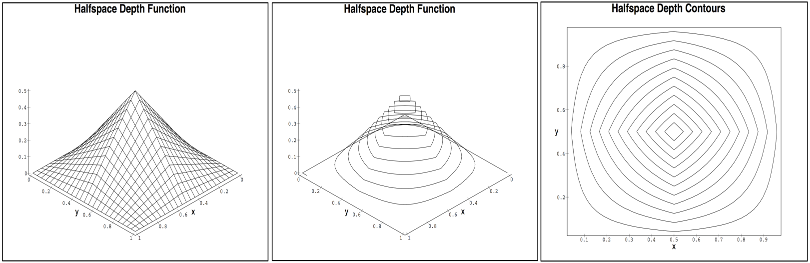
Data Depth-Based Two-Sample Tests

- Tukey propsed the **depth function** to address this issue
- Take a distribution F on $\mathbb{R}^d$
- A depth function $D(x, F)$
- Then, the Half space depth function proposed by Tukey, for $x \in \mathbb{R}^2$ is:

$$D_H(x, F) = \inf\{F(H) : x \in H \text{ closed halfspace}\}$$

Data Depth-Based Two-Sample Tests

- Example: Uniform distribution on the unit square in $\mathbb{R}^2$



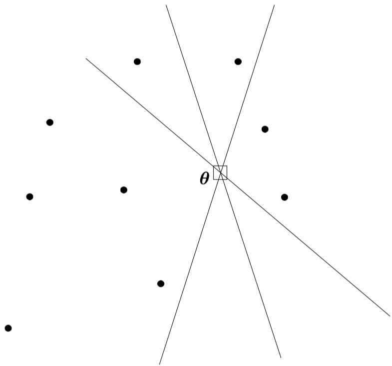
Source: Serfling (2011). (Slides)

- In contrast, density function is constant with no contours of equal density

Data Depth-Based Two-Sample Tests

- The sample halfspace depth of θ is the minimum fraction of data points in any closed halfspace containing θ

$$D_H(\theta, X_1, \ldots, X_n) = \underset{\|u\|=1}{\text{minimize}} \sum_{i=1}^{n} I(u^T X_i \geq u^T \theta)$$

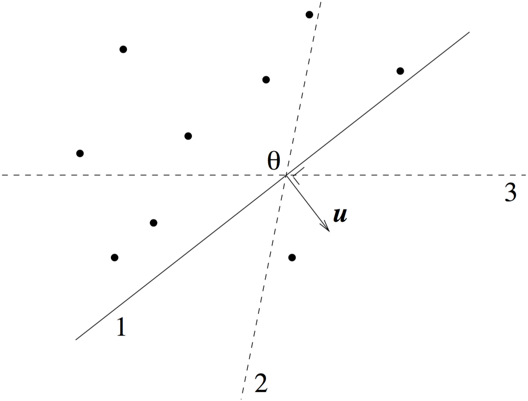


Source: Rousseeuw and Struyf (1998)

Data Depth-Based Two-Sample Tests

- The sample halfspace depth of x is the minimum fraction of data points in any closed halfspace containing θ

$$D_H(\theta, X_1, \ldots, X_n) = \underset{\|u\|=1}{\text{minimize}} \sum_{i=1}^{n} I(u^T X_i \geq u^T \theta)$$



Source: Rousseeuw and Hubert (2015)

Data Depth-Based Two-Sample Tests

- Let $X_1, \ldots, X_{n_1} \sim F$ and $Y_1, \ldots, Y_{n_2} \sim G$
- Null hypothesis $H_0 : F = G$
- Alternative: different location shift and/or a scale
- Liu and Singh (1993) test statistic :

$$Q = \sum_{j=1}^{n_2} \sum_{i=1}^{n_1} I(D(X_i, \{X_1, \ldots, X_{n_1}\}) \leq D(Y_j, \{X_1, \ldots, X_{n_1}\}))$$

- The statistic Q gauges the overall "outlyingness" of the G population with respect to the given F population
- It can detect whether G has a different location and/or has additional dispersion as compared to F

Data Depth-Based Two-Sample Tests

- Special case: Depth function for the univariate Mann-Whitney test

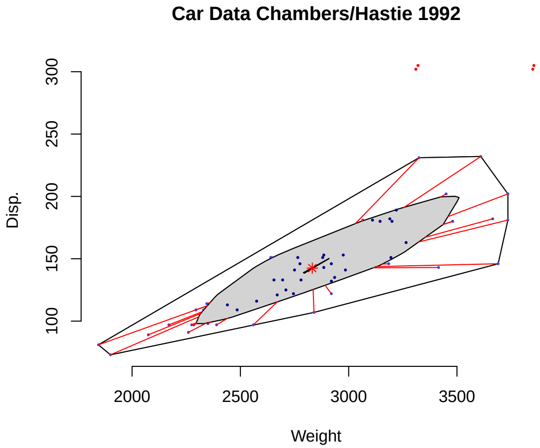
$$T = \sum_{j=1}^{n_2} \sum_{i=1}^{n_1} I(X_i < Y_j)$$

by taking

$$D(x, F) = F(x)$$

- Zuo and He (2006) proved asymptotic normality of this statistic

Data Depth-Based Two-Sample Tests



- The star is the Tukey median
- Bag: The dark area contains 50%
- Fence: Inflating the “bag” by factor 3 relative to Tukey median
- Loop: Convex hull containing all points inside the fence

Data Depth-Based Two-Sample Tests

- Gets increasingly difficult to compute in high dimensions
- Computation time is polynomial in n but exponential in d
- Rousseeuw and Struyf (1998) proposed an approximation
- They compute m random directions out of all $\binom{n}{d}$ directions perpendicular to hyperplanes through d data points

Data Depth-Based Two-Sample Tests

- Set current depth to n
- Repeat m times:
  - Draw a random sample of size d
  - Determine a direction u perpendicular to the d-subset
  - Project all data points on the line L through θ with direction u
  - Compute the univariate depth k of θ on L
  - Set depth to $\min(\text{current depth}, k)$
- This algorithm has time complexity $O(md^3 + mdn)$

Graph-Based Two-Sample Tests

- Alternative multivariate nonparametric tests are based on graphs
- We consider one test based on minimal spanning trees
- A set of n points in $\mathbb{R}^d$ can be computed in $O(dn^2)$ time

Graph-Based Two-Sample Tests

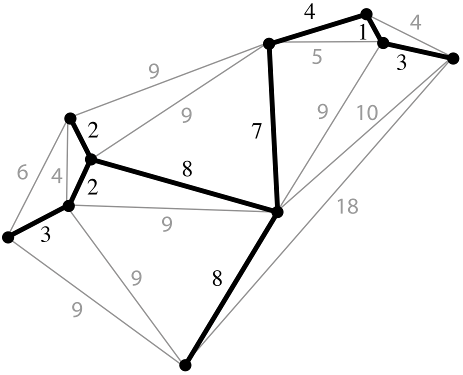
- The Wald-Wolfowitz runs test can be used to evaluate sequences of runs
- For instance to test whether the following sequence is random
  `HHHTTTHHHTHHHTTTT`
- This sequence of coin tosses has 6 runs
  `HHH TTT HHH T HHH TTTT`
- The test statistics is the total number of runs
- Reject H_0 for small and large number of runs
- This has been used to study the hot hand in basketball

Graph-Based Two-Sample Tests

- For univariate continuous observations:
  - Pool the observations
  - Rank the observations
  - Count the number of runs
- Run: sequences of observations that are from the same sample and follow each other
- Test statistics is the total number of runs

Graph-Based Two-Sample Tests

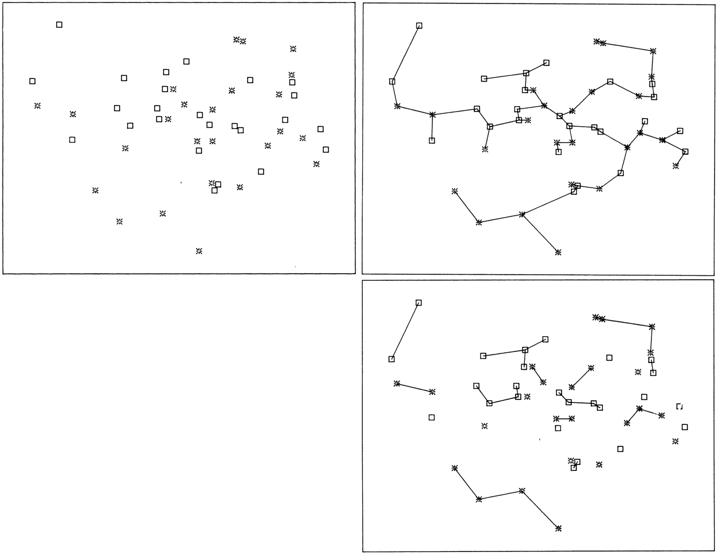
- The Friedman and Rafsky test is a generalization of Wald-Wolfowitz runs test to higher dimensions
- The difficulty is that we need to sort observations
- Friedman and Rafsky purpose to use minimal spanning trees as a multivariate generalization of the univariate sorted list



Graph-Based Two-Sample Tests

- For univariate sample, the edges of the MST are defined by adjacent observations in the sorted list
- The Wald-Wolfowitz runs test can be described in this alternative way:
    1. Construct minimal spanning trees of pooled univariate observations
    2. Remove all edges for which the defining nodes originate from different samples
    3. Define the test statistics as the number of disjoint subtrees that result
- For multivariate samples, just construct minimal spanning tree in step 1 from multivariate observations

Graph-Based Two-Sample Tests



Source: Friedman and Rafsky (1979)

Graph-Based Two-Sample Tests

- Reject H_0 for small and large number of subtrees (runs)
- The null distribution of the test statistics can be computed using permutation tests
  - fix tree
  - permute labels
- Good power in finite samples for multivariate data (against general alternatives: location, spread, and shape)

Graph-Based Two-Sample Tests

- Has been applied to mapping cell populations in flow cytometry data (Hsiao et al. 2016)
  - two cell populations
  - *d* measurements on each cell
  - determine whether the expression of a cellular marker is statistically different
  - suggesting candidates for new cellular phenotypes
  - indicate splitting or merging of cell populations
- Recent development for very high-dimensional data sets (Chen and Friedman 2015)

References

- Tukey (1974). Mathematics and the Picturing of Data
- Friedman and Rafsky (1979). Multivariate Generalizations of the Wolfowitz and Smirnov Two-Sample Tests
- Liu and Singh (1993). A Quality Index Based on Data-Depth and Multivariate Rank Tests
- Holmes (1997). Lecture Notes on Computer Intensive Methods in Statistics
- Rousseeuw and Struyf (1998). Computing Location Depth and Regression Depth in Higher Dimensions
- Zuo and He (2006). On the Limiting Distributions of Multivariate Depth-Based Rank Sum Statistics and Related Tests
- Serfling (2012). Depth (pdf preprint)
- Rousseeuw and Hubert (2015). Statistical Depth Meets Computational Geometry: A Short Survey
- Bhattacharya (2015). Power of Graph-Based Two-Sample Tests
- Chen and Friedman (2015). A New Graph-Based Two-Sample Test for Multivariate and Object Data
- Hsiao, Liu, Stanton, McGee, Qian, and Scheuermann (2016). Mapping Cell Populations in Flow Cytometry Data for Cross-Sample Comparison using The Friedman-Rafsky Test Statistic as a Distance Measure