Nonparametric Tests for Multivariate Data

Gauri Gupta

Multivariate Statistics

Parametric Tests (A Brief Recap)

- Make assumptions about the population distribution (eg : Normality of the data, homoscedasticity, independence) and estimate various hypothesis on the parameters of the distribution
- Even if we do not make a direct assumption about the distribution, in some cases we may be able to approximate it to a Normal distribution (possible due to the Central Limit Theorem)
- We are able to establish confidence intervals for the parameters of the population using various sample statistics, compare population parameters, etc.
- Common Examples are the student's t-test, ANOVA, linear regression (after we make normality, independence and homoscedastic assumption for the residuals), etc.

Recap t-test Contd.

- We assume that the population distribution is Normal
- We assume that samples are independently drawn
- Population Standard Deviation is unknown
- Using the t-test statistic :
- $$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

  where $\bar{x}$ = sample mean, μ = population mean, s = sample std dev.

- We are able to establish confidence intervals for the sample mean and hence test for the null hypothesis for a fixed value of the population mean

Some other parametric tests

Distribution of X_i	Sample size n	Variance σ^2	Statistic	$1-\alpha$ confidence interval
$X_i \sim \mathcal{N}(\mu, \sigma)$	any	known	$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$	$\left[\overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$
$X_i \sim$ any distribution	large	known	$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$	$\left[\overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$
$X_i \sim$ any distribution	large	unknown	$\frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}} \sim \mathcal{N}(0,1)$	$\left[\overline{X} - z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right]$
$X_i \sim \mathcal{N}(\mu, \sigma)$	small	unknown	$\frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$	$\left[\overline{X} - t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \overline{X} + t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right]$

α = the probability a confidence interval will not include the population parameter

Non-Parametric Tests

- Non-Parametric tests do not make assumptions about the underlying distribution of the sample and in general are distribution free (although sometimes making assumptions about the population distribution, but keeping all the parameters of the distribution unknown is also included within the paradigm of Non-Parametric Tests, example : we sometimes assume that the data is from a continouos distribution)[1]
- Non-Parametric Methods do not make assumptions on the sample sizes (in general)
- Non-Parametric Methods tend to be comparatively less extensive mathematically
- Non-Parametric Tests essentially allow us to make statistical conclusions without making too many assumptions on the data

1. *Kendall's Advanced Theory of Statistics: Volume 2A—Classical Inference and the Linear Model*, sixth edition

Why are non-parametric tests relevant

- The assumptions made by Parametric methods are not always appropriate assumptions for Real world data.
- In addition, Parametric Methods require quantitative data. However Data may also be ordinal or categorical :
- The marks of students may be available only as grades A/A-/B/B-/ etc, but not as marks
- We may know the ranking between the marks of the students so we know who ranked highest, who ranked second highest and so on, but we do not know the exact marks for each student
- Population Size can be too small to reliably make the Normality Assumption.

One-Sample Location Problem

- ‘Location’ generally refers to the central measure of the population. In Parametric settings, the central tendency was usually taken to be the mean, while in non-Parametric settings it is usually taken to be the median, but can also be other measures of central tendency depending on the test being used.
- The one-sample location test compares the location parameters of 1 sample to a given a constant.
- Example : Suppose we have the blood pressure distribution for a population. Then the one-sample location test would be a comparison of the ‘location’ (median, or mean in some cases) of the distribution to a given reference value.

Contd.

- **Two Sample Location Test** : In a similar manner, the two-sample location test compares the location parameter of 2 samples to each other.

Example : Suppose we have 2 treatments for a disease and we are trying to find out which treatment is better, or if one is working significantly better than the other. The 2 populations here correspond to the different groups of people who are administered the different treatments (one group may also be a control group given a placebo). Then we can use the test to compare the 'location' parameters of the effectiveness of both populations to test which treatment was more effective.

Univariate Sign Test

- The Sign Test tests the hypothesis that given a random pair of measurements (xi, yi), xi and yi are equally likely to be greater than each other.
- Mathematically, let p = Pr. (X > Y), then
- Null Hypothesis Ho : p = 0.5 (i.e. xi and yi are equally likely to be greater than each other)
- Method of testing the null hypothesis :
- Let W be the number of pairs for which yi – xi > 0 (for a one-sided test) or |yi – xi| > 0 (for a 2-sided test), and let the size of the sample be m, i.e. we have m pairs of the form (xi, yi).
- If we assume the Null hypothesis to be true, then W should follow a Binomial Distribution with parameters m and probability of success = 0.5

Example

Deer	Hind Leg Length (cm)	Foreleg Length (cm)	Difference
1	142	138	+
2	140	136	+
3	144	147	−
4	144	139	+
5	142	143	−
6	146	141	+
7	149	143	+
8	150	145	+
9	142	136	+
10	148	146	+

Example Contd.

- There are m = 10 deer, and the sample shows 8 positive differences and 2 negative differences.
- Assuming the Null hypothesis to be true, the expected positive differences would be 5. We find the probability of observing a result as extreme as or more extreme than the observed result, i.e. the probability of observing 8,9,10 positive differences. For a 2-tailed test, we can also include the probability of observing 0,1,2 positive differences.
- The probability of observing 'i' positive differences is :

$$\binom{10}{i}(1/2)^i(1/2)^{10-i}$$

Example Contd.

- The total probability of observing a result as extreme as 8 positive differences (or more extreme) is then = 2 * 0.00098 + 2 * 0.00977 + 2 * 0.04395 = 0.109
- The null hypothesis (i.e. that the probability of the hind leg being longer than the foreleg is the same) is not rejected at a significance level of 0.05
- In fact, if there were 9 positive differences in the sample, then we would be able to reject the null hypothesis using the sign test at a significance level of 0.05

Multivariate Sign Test

- First let us first formally write the One-sample location problem
- Let X1, X2, …. Xn be i.i.d samples from any continuous p-dimensional distribution "located" or having its central measure at the vector $\theta = (\theta_1, \theta_2, \dots \dots, \theta_p)$
- Then our Null Hypothesis that we wish to test is :
- $H_0 : \theta = 0 \quad vs\ H_a : \theta \neq 0$
- We will see that later in the test we describe that replacing the population variable x with x - θ_0 does not make a difference, hence $\theta = 0$ is taken without loss of generality

Multivariate sign test

- What is the first issue :
- No Notion of a single ‘sign’ in multiple variables
- For univariate data, sign represents either +1 or -1 on the real line
- For multivariate data (say in p-dimensions), a notion of sign could be to look at the ‘direction’ of the vector in p-dimensions.
- Define the sign (called the spatial sign function) in p dimensions to be :

$$\mathbf{S}(\mathbf{x}) = \begin{cases} \|\mathbf{x}\|^{-1}\mathbf{x}, & \mathbf{x} \neq \mathbf{0}, \\ \mathbf{0}, & \mathbf{x} = \mathbf{0}, \end{cases}$$

||x|| : L2 norm of data point x, which is a p-dimensional vector

Test Statistic

- To create an Affine-Invariant sign test, the spatial sign function is not applied directly to the data points, but to a transformed version of the data.
- The spatial signs are then defined as :

- Here A_X is the data driven transformation matrix (called Tyler's transformation)

$$\mathbf{S}_i = \mathbf{S}(A_x \mathbf{X}_i) \quad \text{for } i = 1, \ldots, n,$$

- A_X is constructed in a manner to ensure that the the covariance between the spatial signs S_i are as close to 0 as possible. In other words :

$$pS_i S_i^T = I_p$$

Why Affine Invariance (or equivariance)

- Formally, given a data matrix X, an affine equivariant estimator of location and scatter (m(X), S(X)) is one for which the conditions:
- $m(AX) = A\, m(X)$
- $S(X) = A\, S(X)\, A^T$ hold for any pxp non-singular matrix A
- Consider a situation where we try to use m(X) and S(X) to compute the statistical distance between a point x and m(X), the central location of X (assuming S(X) is invertible)
- $d\big(x, m(X), S(X)\big) = \sqrt{(x - m(X))^T S^{-1}(X)(x - m(X))}$
- Then for an affine equivariant estimator, this measure of the distance (the measure can also be interpreted as the outlyingness of x with respect to X) does not change with scale and orientation of the columns of X

Affine invariance (proof)

$$
\begin{aligned}
d\big(Ax, m(AX), S(AX)\big) &= \sqrt{(Ax - m(AX))^T S^{-1}(AX)(Ax - m(AX))} \\
&= \sqrt{(A(x - m(X)))^T S^{-1}(AX)(A\big(x - m(X)\big))} \\
&= \sqrt{(x - m(X))^T A^T (A^T)^{-1} S^{-1}(X) A^{-1}(A)\big(x - m(X)\big)} \\
&= \sqrt{(x - m(X))^T S^{-1}(X)\big(x - m(X)\big)} \\
&= d\big(x, m(X), S(X)\big)
\end{aligned}
$$

Why Affine Invariance is desirable

- Notionally, scaling the data and even rotating the data should not change the relative signs between the data points even in the pth dimension.
- Hence, we would want our sign test to give the same results for a data matrix X or for a transformed data matrix AX, as the notion of a 'sign' between 2 data points should remain the same
- Practically, if we do not have an affine invariant estimator, then we would have to re-run the test multiple times to check if the test has sensitivity to scale or location
- Example : PCA analysis is not scale invariant, so practically it is advised that when performing PCA, the PCA analysis is run on various rescaling of the data to assess the sensitivity of the results found from the original scaling of the data
- Another Example is of Deep Neural Networks, where DNN's are not rotation invariant so in practice, it is common to use rotated copies of the input (say images) to generalise the model and make it less sensitive to the orientation of the training data

Relevance of Affine Invariance in Multivariate Data

- Firstly, some univariate non-parametric tests are also scale invariant.
- In real world multivariate data, it is very likely that the data collected is not all in the same scale.
- For example, Height measurements may be done in inches while the waist measurement may be done in cm, or macroeconomic variables like GDP or Net Exports might be measured in different currencies across different countries
- In such cases, we are highly likely to first apply standardisation transformations to the data (if X = [GDP in India in Rupees , GDP in dollars in USA], then we would have to standardise the data to [GDP in India in Rupees, 72 * GDP in dollars in USA], i.e. we would have to scale the second column to have reasonable results.

Test Statistic

The final test-statistic for the multivariate sign test is :

The Null Hypothesis is rejected for large values of Q^2

$$Q^2 = np\bar{S}^T\bar{S} = np||\bar{S}||^2$$

Geometric Interpretations of the Sign Test

- Note that A_X is a matrix such that $pS_iS_i^T = I_p$ condition is satisfied. Thus the Tyler's transformation makes the sign covariance matrix close to $\frac{1}{p} I_p$, which would represent the variance-covariance matrix of a vector that is uniformly distributed in p-dimensions.
- This would mean that the aim of Tyler Transformation is to make the signs (directions) of the transformed data points $A_X X_i$ appear as though they are uniformly distributed on the p-dimensional unit sphere.
- Finally, look at

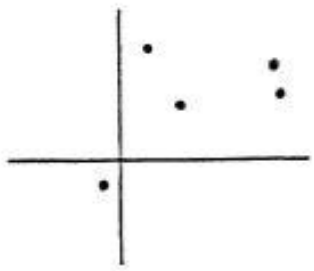
$$Q^2 = np\bar{S}^T\bar{S} = np||\bar{S}||^2$$

- Q^2 is np times the square of the length of the average direction vector of the transformed data points, so it is proportional to the distance between the average direction vector and the assumed central measure (=0).

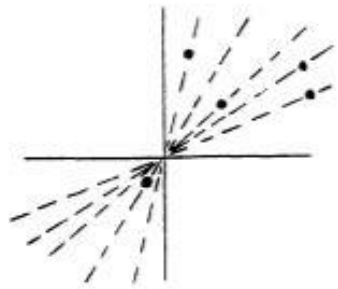
How to get A

- A_X is the upper triangular matrix in the Cholesky decomposition of inverse of Tyler's shape matrix V_X, where V_X is a positive definite symmetric pxp dimensional matrix with Trace(V_X) = p. To find V, it is done iteratively :
- Step 1 : Initialise $V^{(0)} = I_n$
- Step 2 : $$V^{(i+1)} = p(V^{(i)})^{1/2} S_i \bar{S}_i^T (V^{(i)})^{1/2}$$
- Step 3 : Stop when $||pS_i\bar{S}_i^T - I_p||$ is sufficiently small.
- Step 4 : Set $V_X = [p/Trace(V)]V$ and find Ax by the Cholesky decomposition of V_X^{-1}

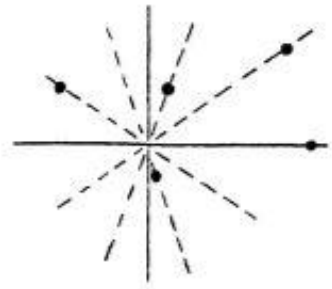
Visualization of the process (Bivariate)



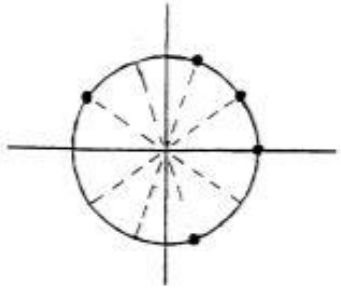
1. Data



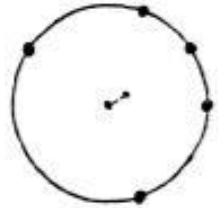
2. Form Axes



3. Space Axes Equally



4. Points on Unit Circle



5. Statistic is Squared Distance from Mean to Origin

K-Sample Location Problem

- Suppose we have multiple different groups of samples from k different populations
- We wish to test whether all the k-populations have an identical distribution, or more precisely, to test if the groups of samples drawn are drawn from an identical distribution.
- Formally, suppose we have k different samples and N total observations where the size of the j^{th} sample is n_j
- Then we have $n_1 + n_2 + \ldots n_k$ = N (fixed N)
- The samples are assumed to be independently taken both between different groups and within a group

K-Sample Location ProbleM (contd)

- Suppose k'th population has CDF $F_k(x)$. Then the most general Null hypothesis is :
- $H_0 : F_1(x) = F_2(x) = \ldots\ldots = F_k(x) \; \forall \; x$ Vs.
- H_a: All the populations are not identically distributed
- For testing the k-sample location problem, we have a weaker hypothesis we test, that the distributions are identical under the null, but that may they differ only by a 'shift' under the alternate hypothesis.
- Given $F_1(x - \theta_1) = F_2(x - \theta_2) = \ldots\ldots = F_k(x - \theta_k) \; \forall \; x$
- We test Null Hypothesis $H_0: \theta_1 = \theta_2 = \ldots = \theta_k$ Against
- $H_a: \theta_i \neq \theta_j$ for some $i \neq j \in \{1,2, \ldots . k\}$

Kruskal-Wallis Test

- Under the Null hypothesis, all the N samples (across all groups) would be drawn from an identical distribution.
- Suppose we order the N samples and assign a rank to each observation. Then since all the N samples are drawn from the same distribution, all the samples are equally likely to get any particular rank i $(where\ i \in \{1,2, \dots, N\}$. Hence the Rank Random Variable (Rank of an observation) is Uniformly distributed over {1,2,…N} for all observations.
- The expected Rank under for each observation under the Null hypothesis is hence

$$= \sum_{i=1}^{N} i * P(rank = i)$$ where P(rank = i) = /N for all i

= (N+1)/2 (Notice it is the same for all observations)

Kruskal-wallis H test

- Then expected sum of ranks for the j'th group would just be :

$$= \sum_{i=1}^{n_j} \text{Expected Rank of ith observation in j'th class}$$
$$= n_j(N+1)/2$$

- Suppose R_j denotes the sum of ranks of the samples from the j'th group
- Then under the Null hypothesis, the sum of squares of the deviation of R_j from the Expected Values should not be too high

Test-Statistic

- Hence an appropriate test statistic could be, where $\bar{R_i} = R_i/n_i$

$$S = \sum_{i=1}^{k} (n_i(\bar{R}_i - (N+1)/2))^2$$

- A high value of S indicates that the sample behavior is very different from what is expected under the Null hypothesis, hence we may reject the Null Hypothesis.
- Under the assumption that the samples are large, let us try to see if there is an approximation possible for S

Test Statistic (contd)

- $\bar{R}_i$ will be approximately normally distributed under the large samples assumption. This is because the average rank is the sum of n_i Random Variables which are independent and identically sampled (The population is the same for all observations within a particular group), so by CLT we can approximate $\bar{R}_i$ to be Normally distributed when there are large samples.
- Note that the sum of $n_i\bar{R}_i$ is nothing but the overall sum of ranks i.e. N(N+1)/2 :

$$\sum_{i=1}^{k} n_i\bar{R}_i = R_{obs1} + R_{obs2} + \ldots R_{\text{obs n}} = 1 + 2 + \ldots N = N(N+1)/2$$

- $\bar{R}_i$ - (N+1)/2 is a zero-mean approximately Normal Random Variable (large samples)

Test Statistic (contd)

- The Variance of $\bar{R}_i$ would be related to the variance of the available ranks and inversely related to $\sqrt{n_i}$ .The standard deviation of a uniform distribution over the available ranks i $\sqrt{(N^2-1)/12}$
- So $$Z_i = \frac{\sqrt{n_i}(\bar{R}_i - (N+1)/2)}{\sqrt{(N^2-1)/12}}$$ are approximately standard-normally distributed.

Test Statistic

- Accounting for the fact that not all Z_i values are independent (since N is fixed, if we know $n_1, n_2, \ldots. n_{k-1}$ then n_k is determined)
- $$H = \frac{N-1}{N} \sum_{i=1}^{k} Z_i^2$$ should be chi-squared with k-1 degrees of Freedom

$$H = \frac{N-1}{N} \sum_{i=1}^{k} \frac{n_i [\bar{R}_i - (N+1)/2]^2}{(N^2-1)/12}$$

Test Statistic

$$H = \frac{12}{N(N+1)}\left(\sum_{i=1}^{k} n_i \bar{R}_i^2\right) - 3(N+1)$$

- H can be approximated by a chi-squared approximation with k-1 degrees of freedom to make critical values of rejection for the Null hypothesis when the **number of samples is large.** In practice, it is generally acceptable to apply the chi-square approximation when the number of observations in a sample are > 5
- When the number of samples is low, we use the table of critical values for the Kruskal-Willis test.

Small Samples Sizes

K=3

	Right-tail probability for H				
n_1, n_2, n_3	0.100	0.050	0.020	0.010	0.001
5, 2, 1	4.200	5.000	—	—	—
5, 2, 2	4.373	5.160	6.000	6.533	—
5, 3, 1	4.018	4.960	6.044	—	—
5, 3, 2	4.651	5.251	6.124	6.909	—
5, 3, 3	4.533	5.648	6.533	7.079	8.727
5, 4, 1	3.987	4.985	6.431	6.955	—
5, 4, 2	4.541	5.273	6.505	7.205	8.591
5, 4, 3	4.549	5.656	6.676	7.445	8.795
5, 4, 4	4.668	5.657	6.953	7.760	9.168
5, 5, 1	4.109	5.127	6.145	7.309	—
5, 5, 2	4.623	5.338	6.446	7.338	8.938
5, 5, 3	4.545	5.705	6.866	7.578	9.284
5, 5, 4	4.523	5.666	7.000	7.823	9.606
5, 5, 5	4.560	5.780	7.220	8.000	9.920

Example

Suppose original observation Table is :

Group 1	Group 2	Group 3
27	20	34
2	8	31
4	14	3
18	36	23
7	21	30
9	22	6

Rank Table

Convert Each entry in the table to its rank

Group 1	Group 2	Group 3
14	10	17
1	6	16
3	8	2
9	18	13
5	11	15
7	12	4

Rank Table

Find Ri for each group i

	Group 1	Group 2	Group 3
	14	10	17
	1	6	16
	3	8	2
	9	18	13
	5	11	15
	7	12	4
Ri	**39**	**65**	**67**
ni	**6**	**6**	**6**

Test Statistic

$$H = \frac{12}{N(N+1)}\left(\sum_{i=1}^{k} n_i \bar{R}_i^2\right) - 3(N+1)$$

$$H = \frac{12}{18(18+1)}\left(\frac{39^2}{6} + \frac{65^2}{6} + \frac{67^2}{6}\right) - 3(18+1) = 2.854$$

Using the Chi-square table, for alpha = 0.05 (5%) and degrees of freedom = 2, we would reject the Test if H > 5.99147. Since H =2.854, we cant reject the null hypothesis at 5% significance in this case

Two sample Test

- Data: Two samples from two distributions.

  $X_1 \ldots\ldots X_n$ from distribution function $F(.)$ and $Y_1 \ldots Y_m$ from distribution function G (.)

- Assumptions:

  X and Y are mutually independent.

  $F(.)$ and $G(.)$ are continuous.

- Hypothesis

  $H_0 : F = G$ i.e. both samples are drawn from same distributions

  versus $H_A : F \neq G$

Overview

- So far, we have seen mostly 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 R^d there is no natural ordering
- Just a straightforward extension of the median to define a center can fail
- A 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 not 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

Depth Function

- Let $X \in R^d$; $d>1$, be a d variate random variable having distribution $F(\cdot)$.
- Centrality or outlyingness of a given observation with respect to the given distribution function $F(\cdot)$ can be measured using the notion of data depth function $D(\cdot, F)$.
- A typical depth function will be a bounded nonnegative function of the form $D : R^d \times F \rightarrow R$
- If $x \in R^d$ then depth of point x will measure how deep/central the point x is with respect to distribution $F(\cdot)$.
- Suppose $x_i \in R^d$; $i = 1, 2, 3, \ldots, n$ be the n observations on $F(\cdot)$ then $D(x_i, F)$ provides a centre outward ranking of these n observations.
- Larger is the depth value, deeper/central is the corresponding observation with respect to the distribution $F(\cdot)$.

Properties of Depth Function

- Affine invariance: If $D(AX + b, F_{AX+b}) = D(X, F_X)$ for any random vector $X \in \mathrm{R}^d$, any $d \times d$ nonsingular matrix A and $b \in \mathrm{R}^d$ then D is to be called as affine invariant.
- Maximality at the centre: For any distribution function F having centre μ; $D(\mu, F) = \sup_{x \in \mathrm{R}d} D(x, F)$.
- Monotonicity relative to the deepest point: For any distribution function F having centre μ, $D(x; F) < D(\mu + \alpha(x - \mu); F)$ holds for $0 \leq \alpha < 1$.
- Vanishing at infinity: $D(x; F) \to 0$ as $||x|| \to \infty$, for every F.

Mahalanobis Depth Function

- The Mahalanobis depth is perhaps the most common depth function associated with multivariate normal theory.
- Mahalanobis norm (Mahalanobis (1936)) $||\cdot||_M$ as:

$$||\mathbf{x}||_M = \sqrt{\mathbf{x}' M^{-1} \mathbf{x}}, \quad \text{for all } \mathbf{x} \in \mathbb{R}^d$$

- F is the given distribution and μ(F) and Σ(F) are corresponding location and covariance measures,
- The Mahalanobis depth is defined as

$$MD(\mathbf{x}, F) = \frac{1}{1 + ||\mathbf{x} - \mu(F)||^2_{\Sigma(F)}}$$

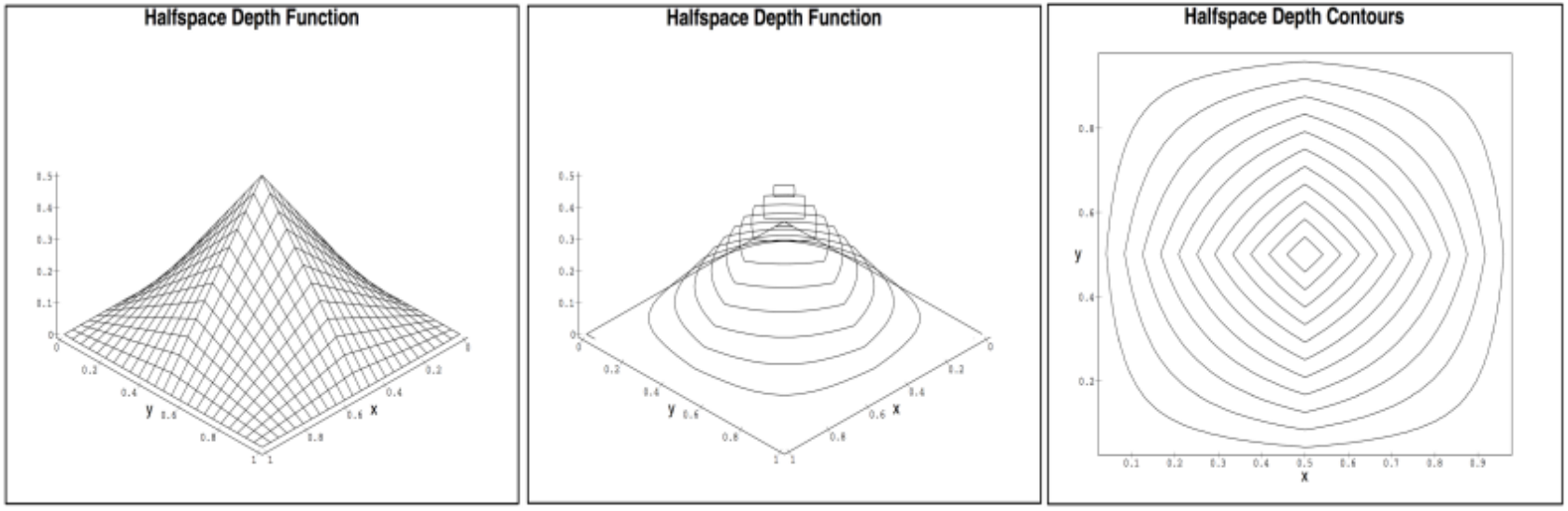
Tukey Depth Function

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

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

Tukey Depth Function

- Example: Uniform distribution on the unit square in R^2



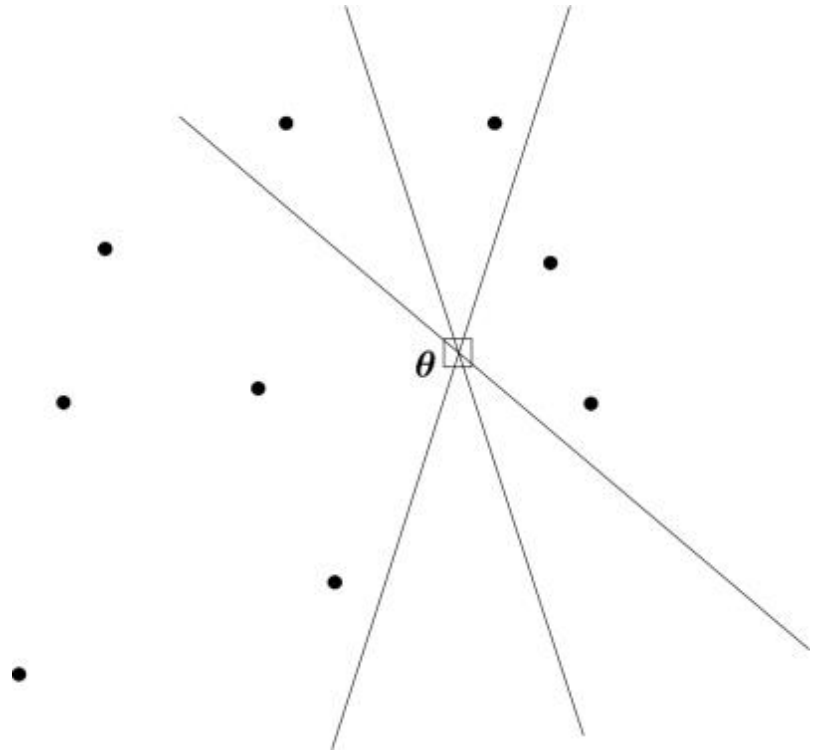
Source: Serfling (2011).

- 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

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

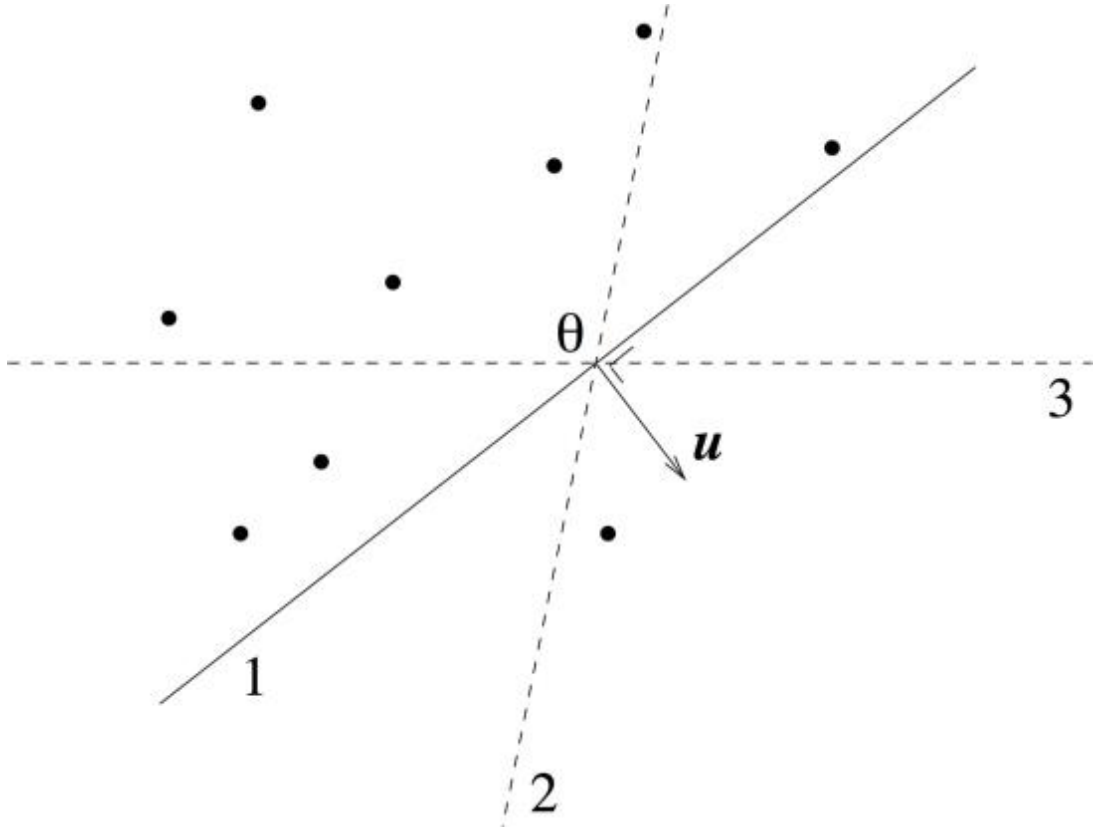


Source: Rousseeuw and Struyf (1998)

Data Depth-Based Two-Sample Tests

- The sample halfspace depth of 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

- 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
- The 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 determines 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

Graph-Based Two-Sample Tests

- Alternative multivariate nonparametric tests are based on graphs
- Sequencing of points in sample using minimal spanning trees

Graph-Based Two-Sample Tests

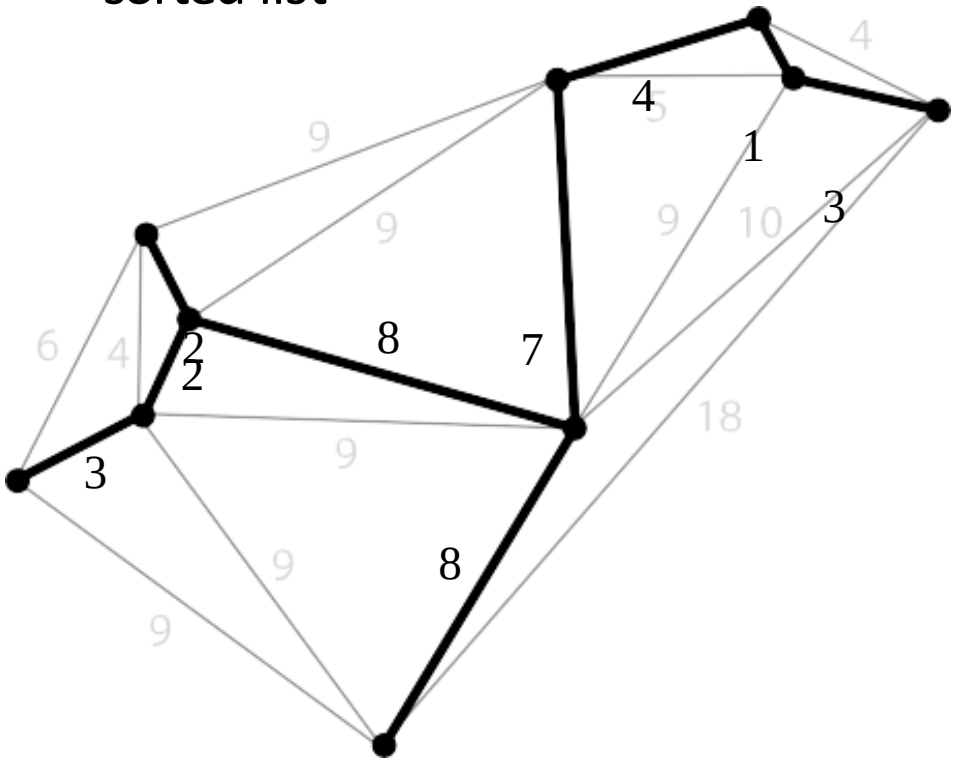
- The Wald-Wolfowitz runs test can be used to evaluate sequences of runs
- For instance to test whether the following sequence is random HHHTTTHHHTHHHTTTT
- This sequence of coin tosses has 6 runs
  HHH TTT HHH T HHH TTTT
- The test statistics is the total number of runs
- Reject H_0 for small and large 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
  - Sort or Rank the observations in ascending order
  - Count the number of runs
- Run: A run of a sequence is a maximal non-empty segment of the sequence consisting of adjacent equal elements
- Test statistics is the total number of runs
- Difficulty in extending this to multivariate observations is that the notion of sorted list cannot be immediately generalised

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



Properties of MST suitable for defining a sorted list

- Minimal spanning trees have two important properties that make them appropriate for application to the two-sample problem
  - Connects all of the nodes with N-1 edges
  - Node pairs defining the edges represent points that tend to be close together i.e. small distance or dissimilarity

MST in Graph Based test

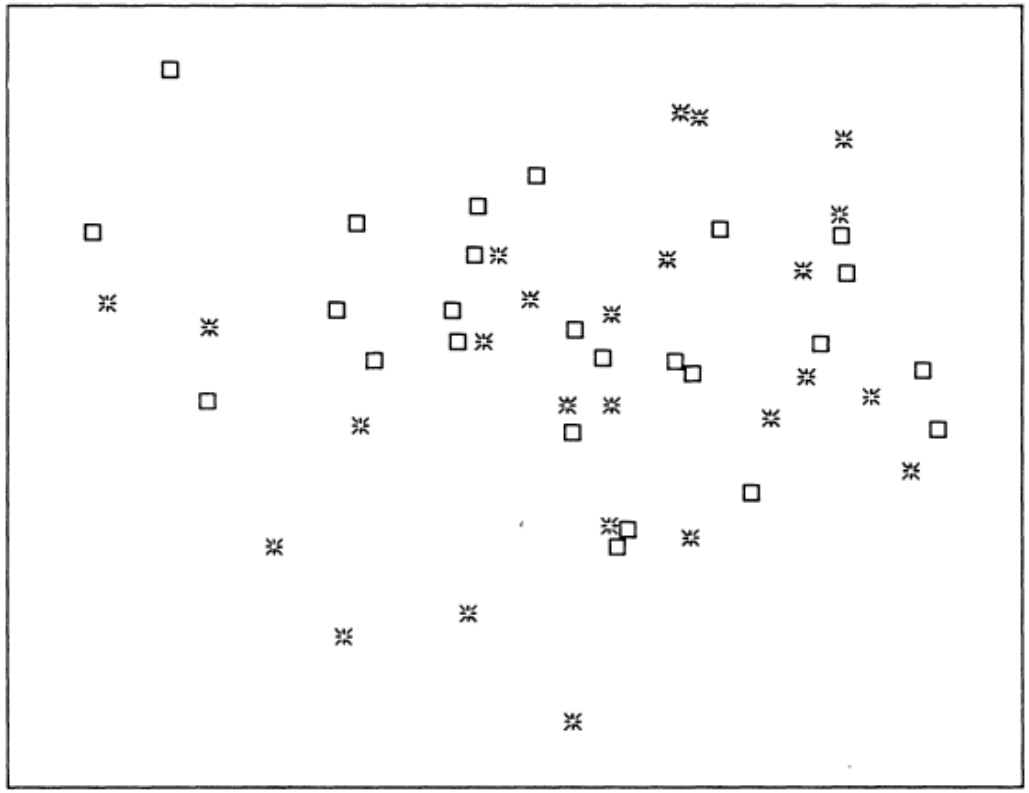
- A minimal spanning tree of an edge weight graph is a spanning tree for which the sum of edge weights is minimum.
- Construct the graph in the following way:
  1. N pooled sample data points in R^d as nodes
  2. Edges linking all pairs i.e. a complete graph with N(N-1)/2 edges
  3. Weight associated with each edge is the Euclidean distance or any other distance metric metric
- MST of this graph is thus the subgraph of minimum total distance (dissimilarity) that provides a path between every two nodes.

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

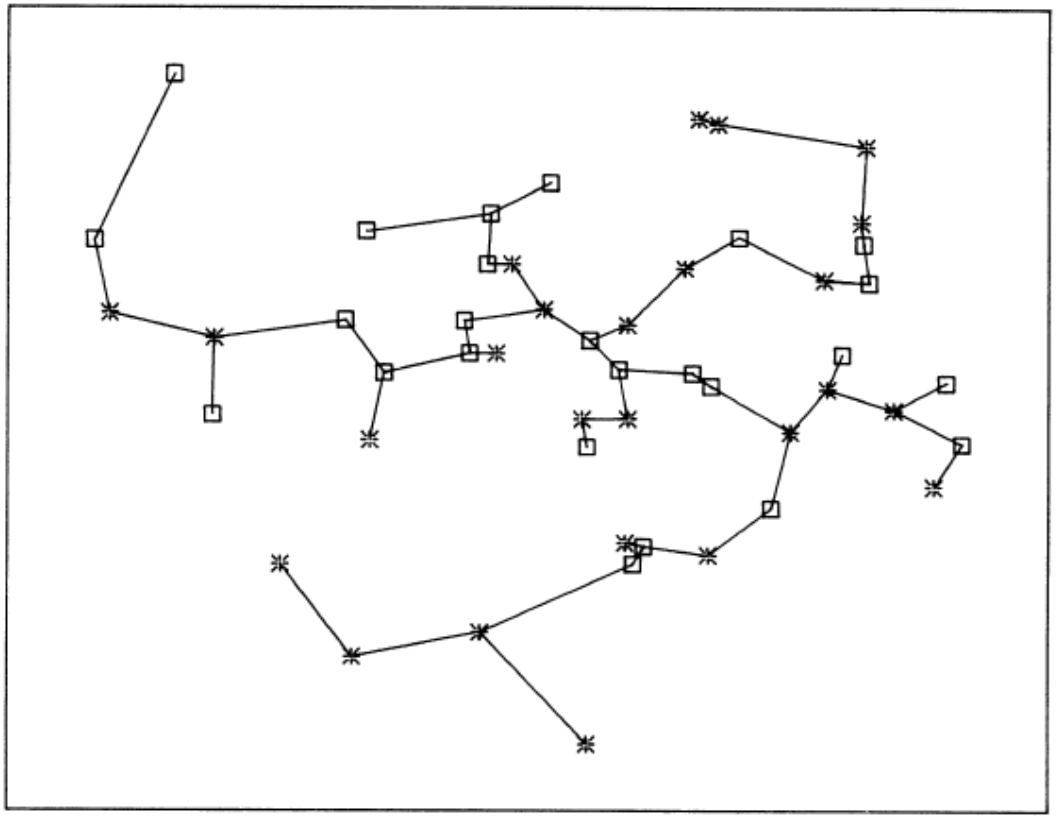
Example

The data shown in Figure are two samples of 25 points each drawn from a standard bivariate normal distribution



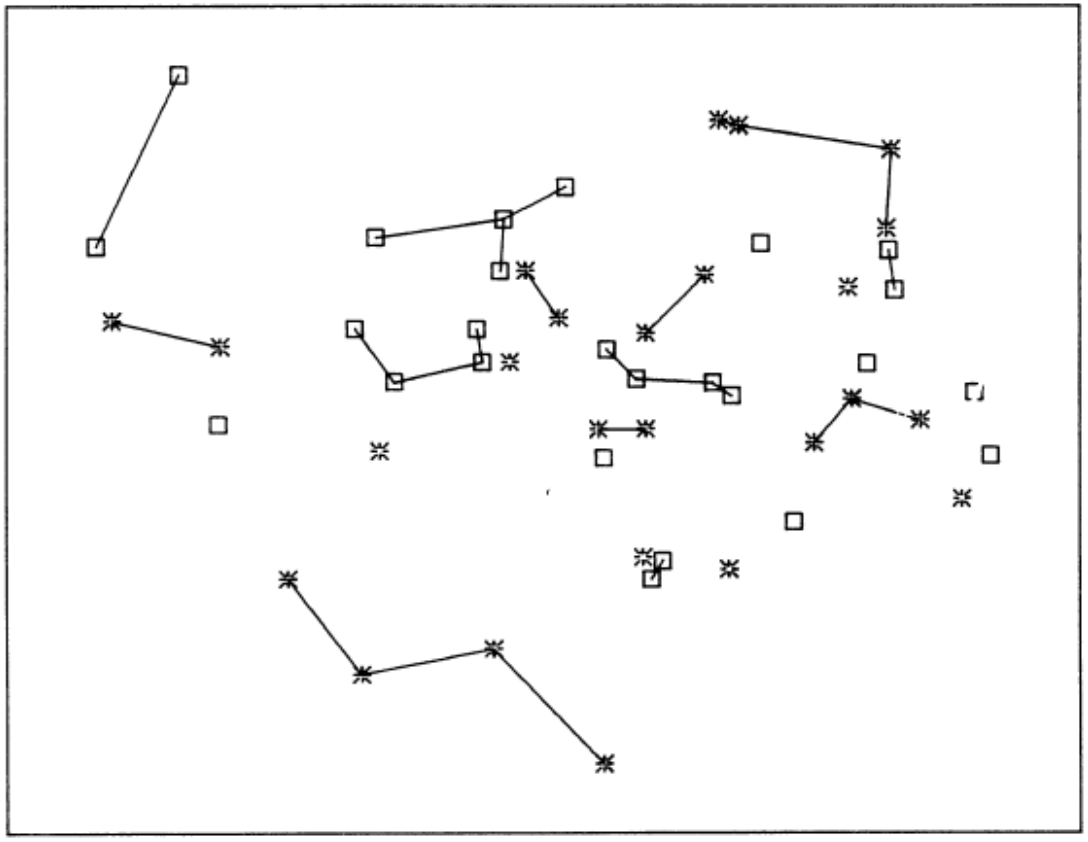
Example

Superimpose the MST of pooled samples

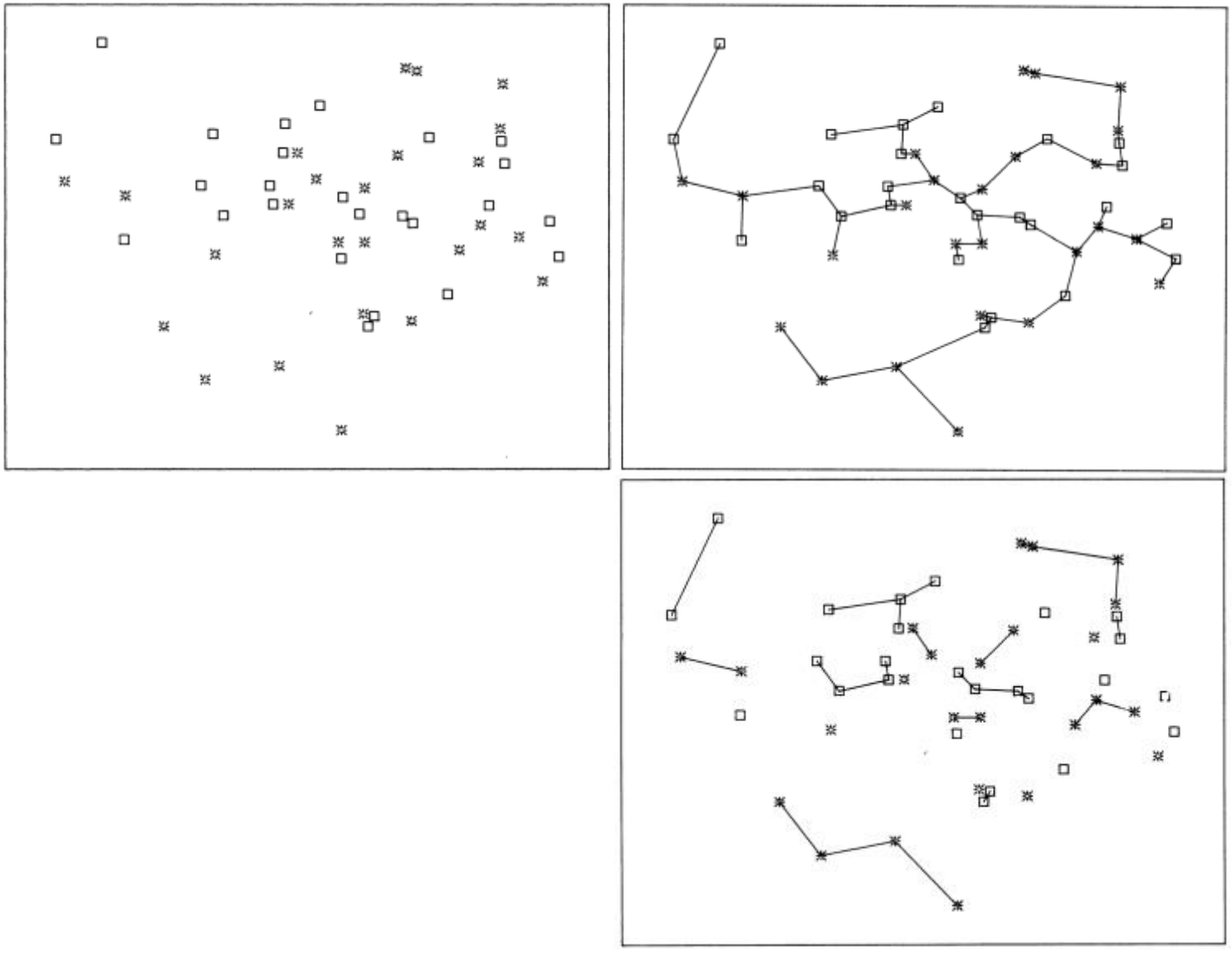


Example

Delete the edges linking nodes from different samples



Example:



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

References

- 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
- M. S. Barale & D. T. Shirke. A test based on data depth for testing location scale of the two multivariate populations
- Shojaeddin Chenouri & Christopher G. Small. A nonparametric multivariate multisample test based on data depth
- Sakineh Dehghan & Mohammad Reza Faridrohani. Depth based signed rank tests for bivariate central symmetry