Độ đo tương quan

NHẬP MÔN THỊ GIÁC MÁY TÍNH

Trình bày: TS Trần Thái Sơn; Email: ttson@fit.hcmus.edu.vn

Độ đo khoảng cách

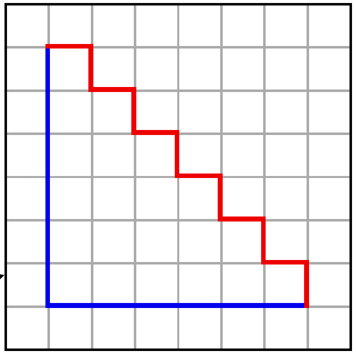
- Heuristic
  - Minkowski-form
  - Weighted-Mean-Variance (WMV)
- Thống kê phi tham số
  - χ^2
  - Kolmogorov-Smirnov (KS)
  - Cramer/von Mises (CvM)
- Lý thuyết thông tin
  - Kullback-Liebler (KL)
  - Jeffrey-divergence (JD)
- Các độ đo khoảng cách nền:
  - Histogram intersection
  - Quadratic form (QF)
  - Earth Movers Distance (EMD)
- Khoảng cách Mahalanobis
- Khoảng cách Hausdoff

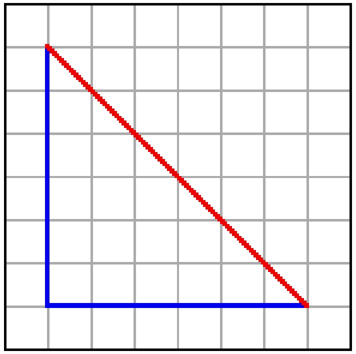
Khoảng cách Heuristic

- Công thức khoảng cách Minkowski L_p

$$D(I,J) = \left(\sum_i \left| f(i,I) - f(i,J) \right|^p \right)^{1/p}$$

- Trường hợp đặc biệt:
  - L_1: giá trị tuyệt đối, khoảng cách cityblock hay Manhattan.
  - L_2: khoảng cách Euclidian.





Khoảng cách Heuristic

- Độ lệch trung bình trọng số (Weighted-Mean-Variance):
  - Chỉ chứa một phần rất ít thông tin về phân phối (distribution).

$$D^r(I,J) = \frac{\left|\mu_r(I) - \mu_r(J)\right|}{\left|\sigma(\mu_r)\right|} + \frac{\left|\sigma_r(I) - \sigma_r(J)\right|}{\left|\sigma(\sigma_r)\right|}$$

Thống kê phi tham số

- Nonparametric Test Statistics:
- Phương pháp χ^2
- Độ đo dựa trên độ tương đồng của 2 mẫu:

$$D(I,J)=\sum_i \frac{\left(f(i;I)-\hat{f}(i)\right)^2}{\hat{f}(i)},$$

$$\text{where } \hat{f}(i)=\left[f(i;I)+f(i;J)\right]/2$$

Thống kê phi tham số

- Khoảng cách Kolmogorov-Smirnov:
  - Độ đo dựa trên độ tương đồng của 2 mẫu:

$$D^r(X,Y) = \max_i |F^r(i;X) - F^r(i;Y)|$$

- Kramer/von Mises

$$D^r(X,Y) = \sum_i (F^r(i;X) - F^r(i;Y))^2.$$

Lý thuyết thông tin

- Kullback-Liebler:
  – Chi phí để mã hóa một phân phối thành phân phối khác.

$$D(X,Y)=\sum_i f(i;X)\log\frac{f(i;X)}{f(i;Y)},$$

- Jeffrey divergence:
  – Như KL nhưng tính toán số học nhiều hơn.

$$D(X,Y)=\sum_i f(i;X)\log\frac{f(i;X)}{\hat{f}(i)}+f(i;Y)\log\frac{f(i;Y)}{\hat{f}(i)}.$$

Khoảng cách nền

- Ground Distance:
- Phép giao 2 histogram:
  - Tốt cho so khớp từng phần:

$$d_{\cap}(H,K) = 1 - \frac{\sum_{\mathbf{i}} \min(h_{\mathbf{i}}, k_{\mathbf{i}})}{\sum_{\mathbf{i}} k_{\mathbf{i}}}$$

- Dạng bình phương:
  - Ký hiệu A là ma trận dương ($A \geq 0$) là ma trận tương quan giữ I và J. fI và fJ là vector của I và J.

$$D(I,J) = \sqrt{(\mathrm{f}_I - \mathrm{f}_J)^t \, \mathrm{A} \, (\mathrm{f}_I - \mathrm{f}_J)}$$

Khoảng cách nền

- Khoảng cách Earth Mover

$$\text{Let } P = \{(p_1, w_1), \ldots, (p_n, w_n)\}$$

$$Q = \{(q_1, w_1), \ldots, (q_m, w_m)\}$$

$$D(i, j) = dist(p_i, q_j)$$

$$\min_{f_{ij}} \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij}$$

$$\text{s.t} \quad f_{ij} \geq 0$$

$$\sum_{j=1}^{n} f_{ij} \leq w_{p_i}$$

$$\sum_{i=1}^{m} f_{ij} \leq w_{q_j}$$

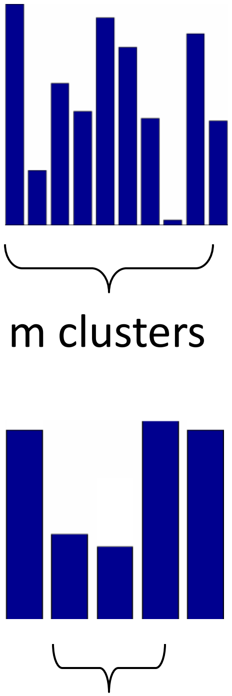
$$\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} = \min\left(\sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{n} w_{p_j}\right)$$

$$\Rightarrow \quad EMD(P, Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}}$$

Khoảng cách nền

- Earth Mover Distance

P



m clusters

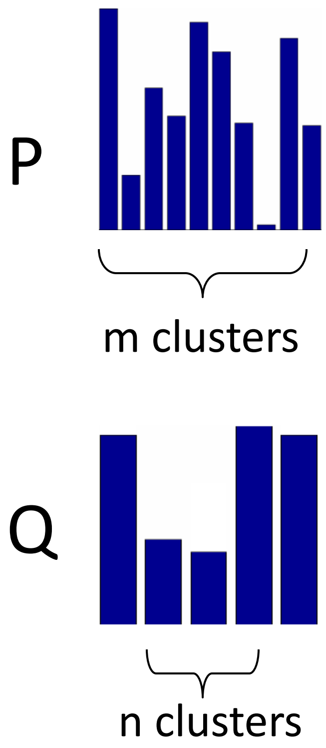
Q

n clusters

$$\sum_{\text{All movements}} (\text{distance moved}) * (\text{amount moved})$$

Khoảng cách nền

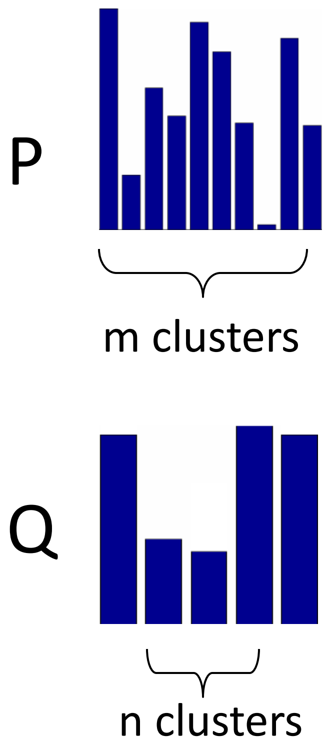
- Earth Mover Distance



$$\sum_{i=1}^{m}\sum_{j=1}^{n} \text{(distance moved) * (amount moved)}$$

Khoảng cách nền

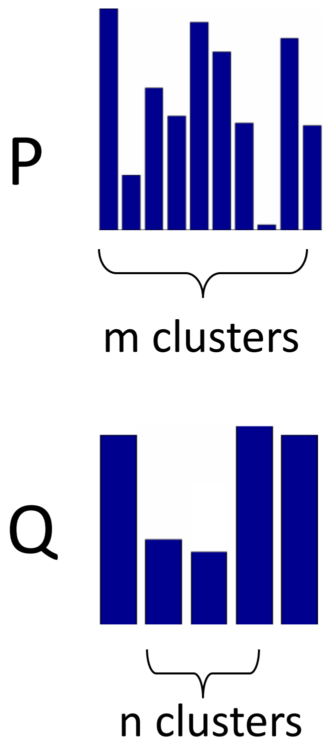
- Earth Mover Distance



$$\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} \ast \text{(amount moved)}$$

Khoảng cách nền

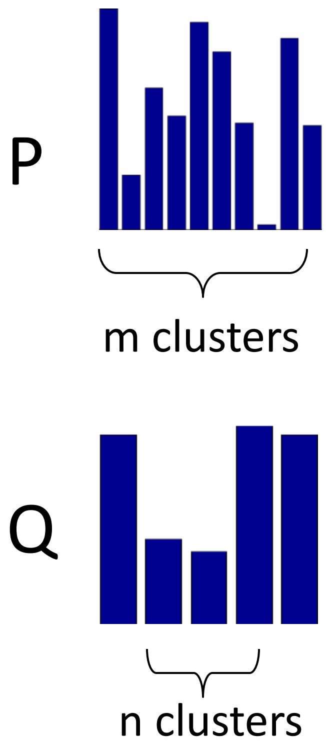
- Earth Mover Distance – Linear programming.



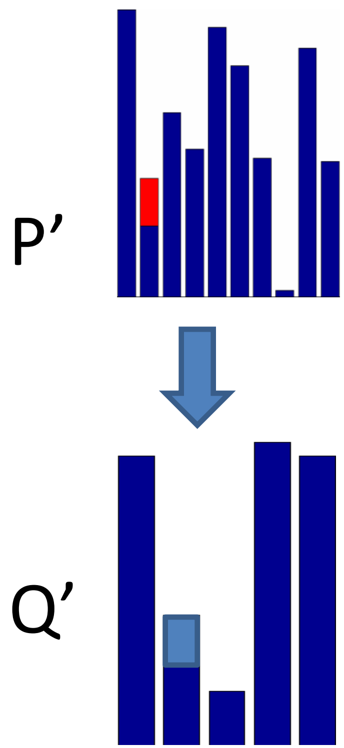
$$\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij}\, f_{ij} = \mathrm{WORK}$$

Khoảng cách nền

- Earth Mover Distance – Điều kiện.



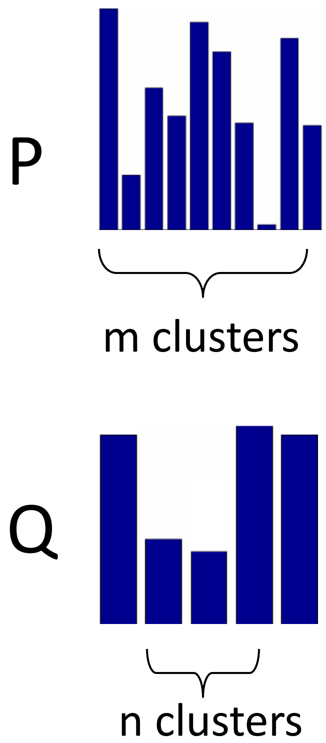
1. Chỉ có thể di chuyển "earth" từ P sang Q



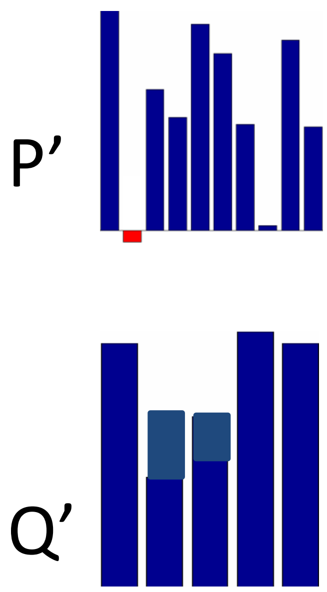
$$f_{ij} \geq 0$$

Khoảng cách nền

- Earth Mover Distance – Điều kiện.



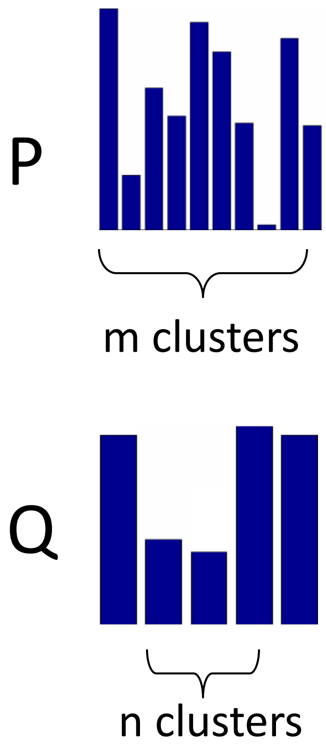
2. Không thể chuyển nhiều "earth" hơn vốn có.



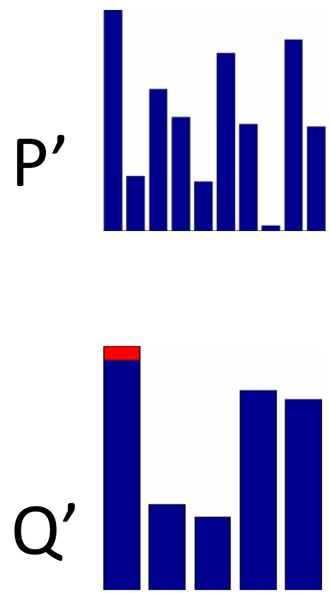
$$\sum_{j=1}^{n} f_{ij} \leq w_{p_i}$$

Khoảng cách nền

- Earth Mover Distance – Điều kiện.



3. Q không thể nhận nhiều "earth" hơn nó có thể giữ.

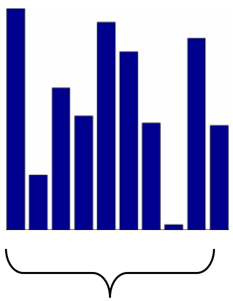


$$\sum_{i=1}^{m} f_{ij} \leq w_{q_j}$$

Khoảng cách nền

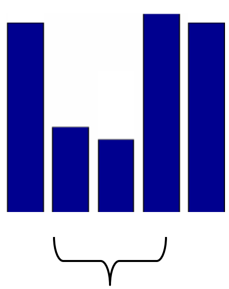
- Earth Mover Distance – Điều kiện.

P



m clusters

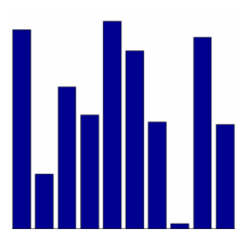
Q



n clusters

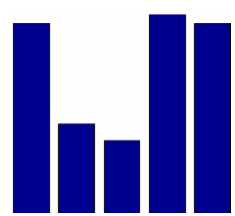
4.Phải di chuyển nhiều "earth" nhất có thể

P'



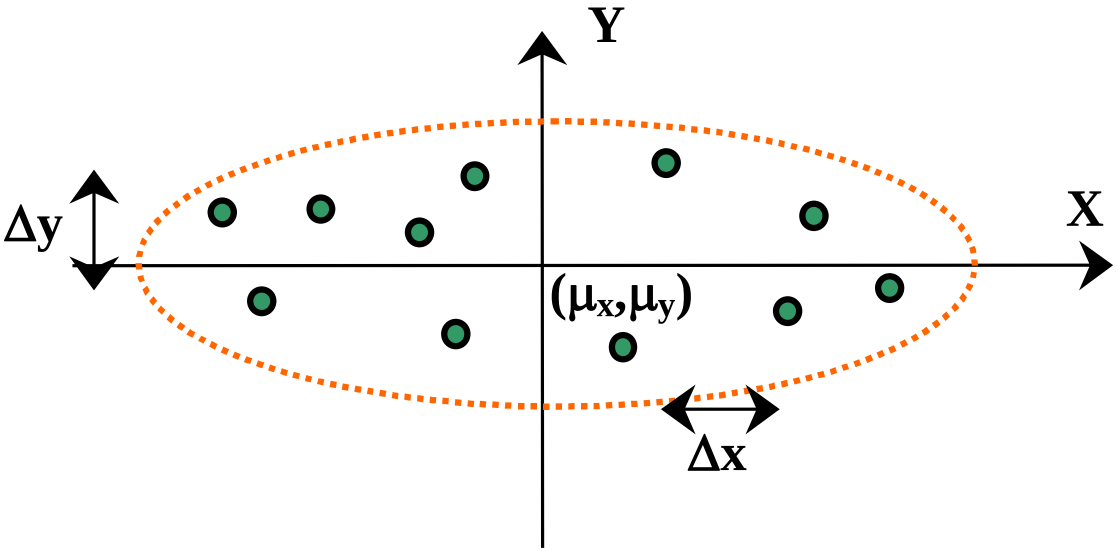
$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \min\left(\sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{n} w_{q_j}\right)$$

Q'



Khoảng cách Mahalanobis

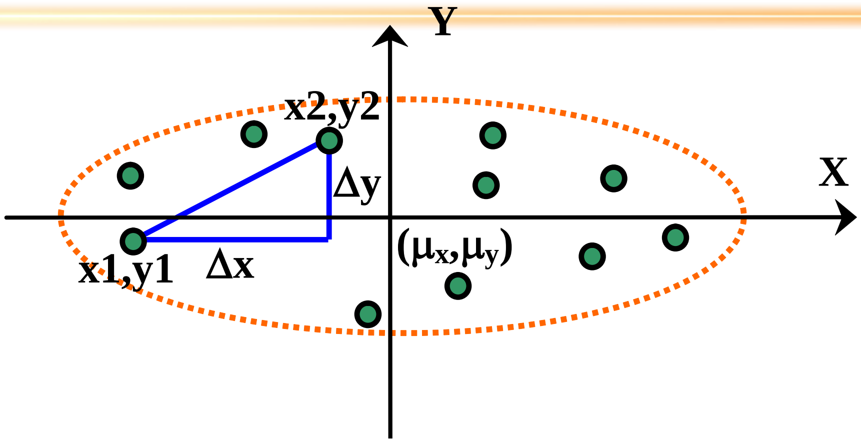
- Khoảng cách Euclidian có trọng số trên tất cả các chiều như nhau, tuy nhiên trong thống kê không phải lúc nào cũng vậy.



Euclidian distance $\Delta x = \Delta y$

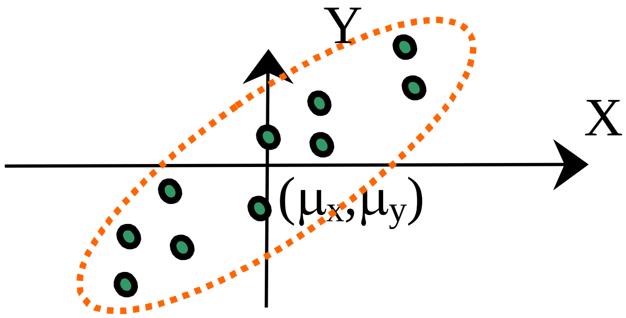
However statistically $\Delta x < \Delta y$

Khoảng cách Mahalanobis



- Với hiệp phương sai nhỏ hay = 0 ta có thể chuẩn hóa khoảng cách bằng cách chi cho phương sai.
- $\Delta' x = (x_2 - x_1) / \text{sqrt}(\sigma_{xx})$ $\Delta' y = (y_2 - y_1) / \text{sqrt}(\sigma_{yy})$
- Dist = $\text{sqrt}(\Delta' x^2 + \Delta' y^2)$
- Ma trận hiệp phương sai là ma trận đường chéo và khoảng cách giữa 2 điểm có thể viết lại:
- $\text{sqrt}[(x_2 - x_1, y_2 - y_1)^T \Sigma^{-1} (x_2 - x_1, y_2 - y_1)]$ = khoảng cách Mahalanobis.

Khoảng cách Mahalanobis



- Khoảng cách Mahalanobis = sqrt$[(x_2-x_1,y_2-y_1)^T \Sigma^{-1} (x_2-x_1,y_2-y_1)]$ cũng tốt cho trường hợp hiệp phương sai lớn.
- Giá trị độ đo thay đổi như cách hiệp phương sai tăng.
- Σ là ma trận hiệp phương sai.

Khoảng cách Mahalanobis

- Mahalanobis là phiên bản đa chiều của z-score. Độ đo của khoảng cách từ trọng tâm (giá trị trung bình đa chiều) của phân phối, với hiệp phương sai (phương sai đa chiều) của phân phối.
- Trường hợp điểm ngoài lề (outlier) nếu xác suất của nó < 0.0001 hay bé hơn. Giá trị này tuân theo phân phối chi-square với độ tự do của biến bằng số lượng biến trong phép tính.
- Mahalanobis yêu cầu các biến là metric, ví dụ biến khoảng và biến thứ tự được xem là metric.

Tham khảo

[1] Chapter 2, Shape Analysis and Classification: Theory and Practice, L.D.F. Costa, R.M. Cesar Jr , CRC. Press, 2000.