Similarity Measurement

Distance Measures

- Heuristic
  - Minkowski-form
  - Weighted-Mean-Variance (WMV)
- Nonparametric test statistics
  - χ^2
  - Kolmogorov-Smirnov (KS)
  - Cramer/von Mises (CvM)
- Information-theory divergences
  - Kullback-Liebler (KL)
  - Jeffrey-divergence (JD)
- Ground distance measures
  - Histogram intersection
  - Quadratic form (QF)
  - Earth Movers Distance (EMD)
- Mahalanobis distance

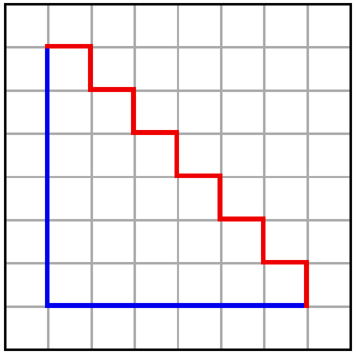
- Hausdoff distance

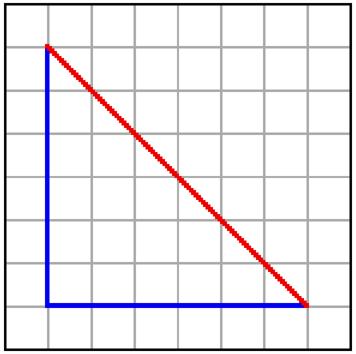
Heuristic Distances

- Minkowski-form distance L_p

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

- Special cases:
  - L_1: absolute, cityblock, or Manhattan distance
  - L_2: Euclidian distance





Heuristic Distances

- Weighted-Mean-Variance
  - Only includes minimal information about the distribution

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

Nonparametric Test Statistics

- χ^2
  - Measures the underlying similarity of two samples

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

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

Nonparametric Test Statistics

- Kolmogorov-Smirnov distance
  - Measures the underlying similarity of two samples

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

- Kramer/von Mises

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

Information Theory

- Kullback-Liebler
  - Cost of encoding one distribution as another

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

Information Theory

- Jeffrey divergence
  - Just like KL, but more numerically

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

Ground Distance

- Histogram intersection
  - Good for partial matches

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

Ground Distance

- Quadratic form
  - Let A denote a positive matrix (A $\geq$ 0) such as the correlation matrix of *I* and *J.*
  - f_I and f_J are two vectors of *I* and *J.*

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

Ground Distance

- Earth Mover Distance

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

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

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

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

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

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

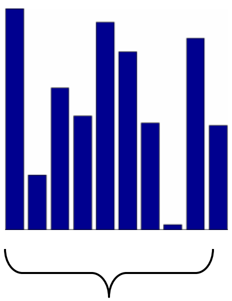
$$\sum_{i=1}^{m} f_{ij} \leq w_{q_j}$$

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

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

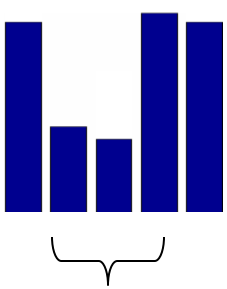
Ground Distance - Earth Mover Distance

P



m clusters

Q

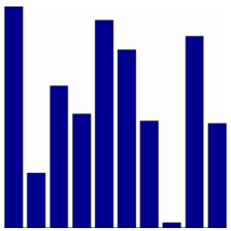


n clusters

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

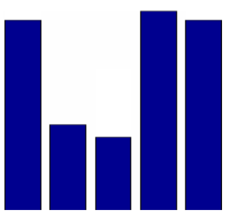
Ground Distance - Earth Mover Distance

P



m clusters

Q

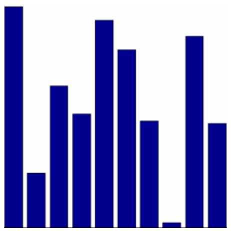


n clusters

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

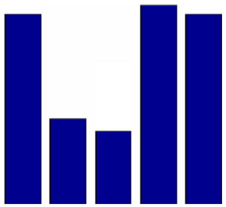
Earth Mover Distance

P



m clusters

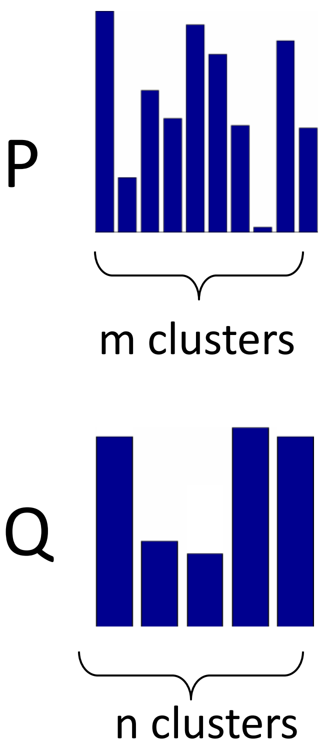
Q



n clusters

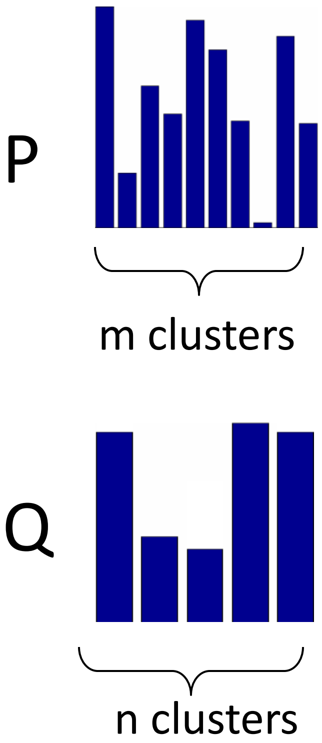
$$\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} \text{ * (amount moved)}$$

Linear programming

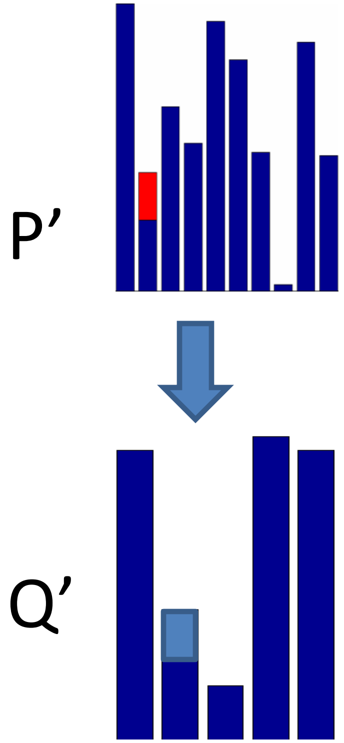


$$\sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} \, f_{ij} = \mathrm{WORK}$$

Constraints



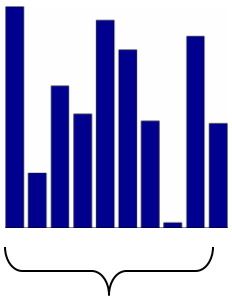
1. Move “earth” only from P to Q



$$f_{ij} \geq 0$$

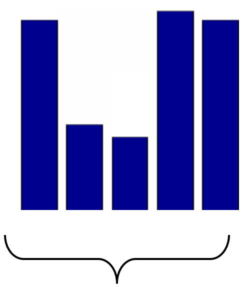
Constraints

P



m clusters

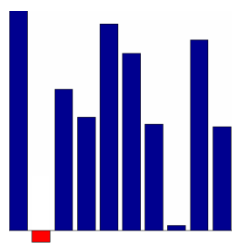
Q



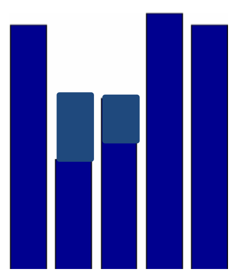
n clusters

2. Cannot send more "earth" than there is

P'



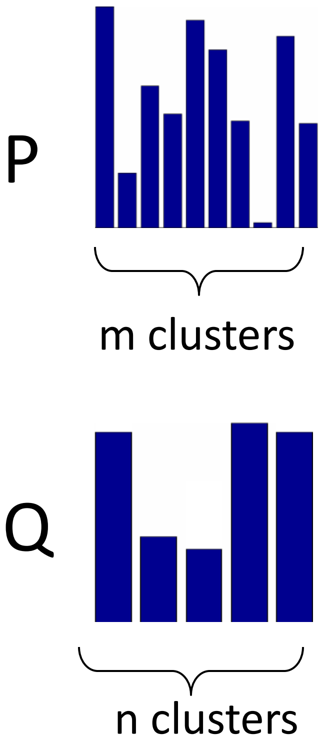
Q'

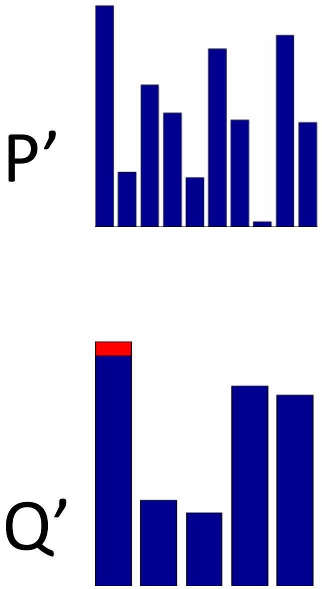


$$\sum_{j=1}^{n} f_{ij} \quad \leq \quad w_{p_i}$$

Constraints

3. Q cannot receive more "earth" than it can hold

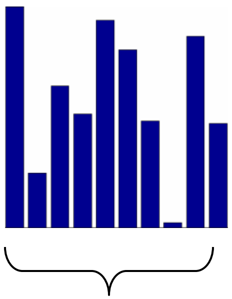




$$\sum_{i=1}^{m} f_{ij} \quad \leq \quad w_{q_j}$$

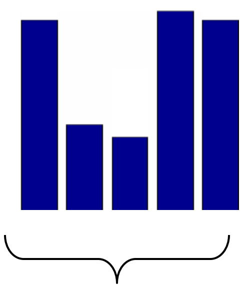
Constraints

P



m clusters

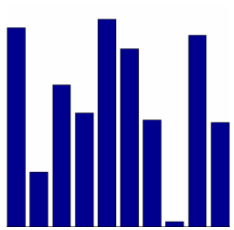
Q



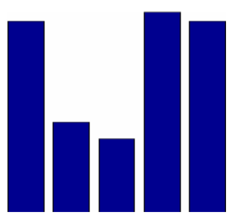
n clusters

4. As much “earth” as possible must be moved

P’



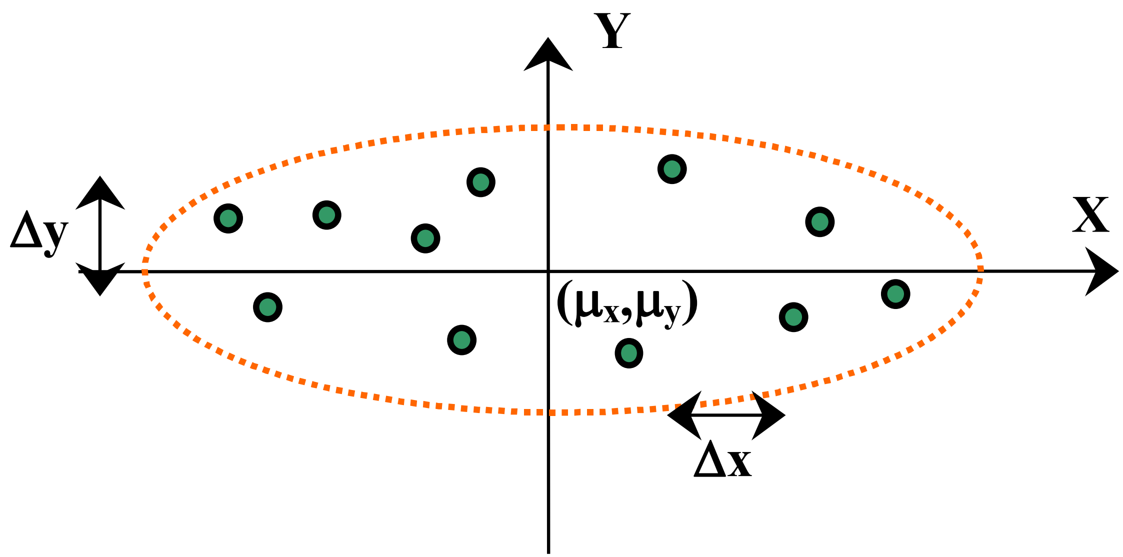
Q’



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

Mahalanobis Distance

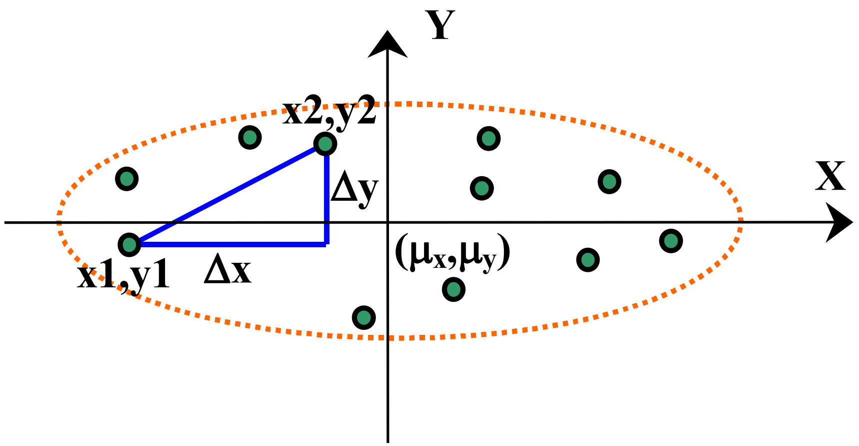
- Euclidian distance weights all dimensions (variables) equally, however, statistically they may not be the same:



Euclidian distance $\Delta x = \Delta y$

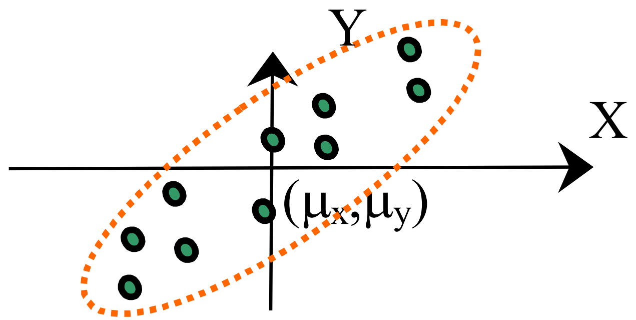
However statistically $\Delta x < \Delta y$

Mahalanobis Distance



- It is easy to see that for low (or zero) covariance we can can normalise the distances by dividing by the variance:
- $\Delta'x = (x_2 - x_1) / \mathrm{sqrt}(\sigma_{xx})$ $\Delta'y = (y_2 - y_1) / \mathrm{sqrt}(\sigma_{yy})$
- Dist = $\mathrm{sqrt}(\Delta'x^2 + \Delta'y^2)$
- In this case the covariance matrix is diagonal and the distance between two points can be written:
- $\mathrm{sqrt}[(x_2-x_1, y_2-y_1)^T \Sigma^{-1} (x_2-x_1, y_2-y_1)]$ = the Mahalanobis distance

Mahalanobis Distance



- The Mahalanobis distance = sqrt[$(x_2-x_1,y_2-y_1)^T \Sigma^{-1} (x_2-x_1,y_2-y_1)$] also works for high co-variance.
- Notice how the measure changes as co-variance increases
- $\sum$ is a covariance matrix

Mahalanobis and Multivariate Outliers

- Mahalanobis is a multidimensional version of a z-score. It measures the distance of a case from the centroid (multidimensional mean) of a distribution, given the covariance (multidimensional variance) of the distribution.

- A case is a multivariate outlier if the probability associated with its value is 0.001 or less. This value follows a chi-square distribution with degrees of freedom equal to the number of variables included in the calculation.

- Mahalanobis requires that the variables be metric, i.e. interval level or ordinal level variables that are treated as metric.

Gromov-Hausdorff distance

Allow for arbitrary embedding space $(\mathbb{X}, d_{\mathbb{X}})$

$$d_{\mathsf{GH}}(\mathcal{Q},\mathcal{S}) \;=\; \inf_{\substack{\mathbb{X} \\ \varphi:\mathcal{S}\to\mathbb{X} \\ \psi:\mathcal{Q}\to\mathbb{X}}} \; d_{\mathsf{H}}^{\mathbb{X}}(\varphi(\mathcal{S}),\psi(\mathcal{Q}))$$

where φ, ψ are isometric embeddings.

- Satisfies the metric axioms with $c = 2$
- Consistent to sampling: if $\mathcal{S}^r$ is an r-covering of $\mathcal{S}$, then

$$|d_{\mathsf{GH}}(\mathcal{Q},\mathcal{S}) - d_{\mathsf{GH}}(\mathcal{Q},\mathcal{S}^r)| \le r$$

- Computation: intractable

M. Gromov, 1981

Gromov-Hausdorff distance

For compact surfaces, there exists an equivalent definition in terms of metric distortions:

$$d_{\mathsf{GH}}(\mathcal{Q},\mathcal{S}) = \inf_{\substack{\varphi:\mathcal{S}\to\mathcal{Q}\\ \psi:\mathcal{Q}\to\mathcal{S}}} \max\{\mathrm{dis}\,\varphi,\ \mathrm{dis}\,\psi, \mathrm{dis}\,(\varphi,\psi)\}$$

where:

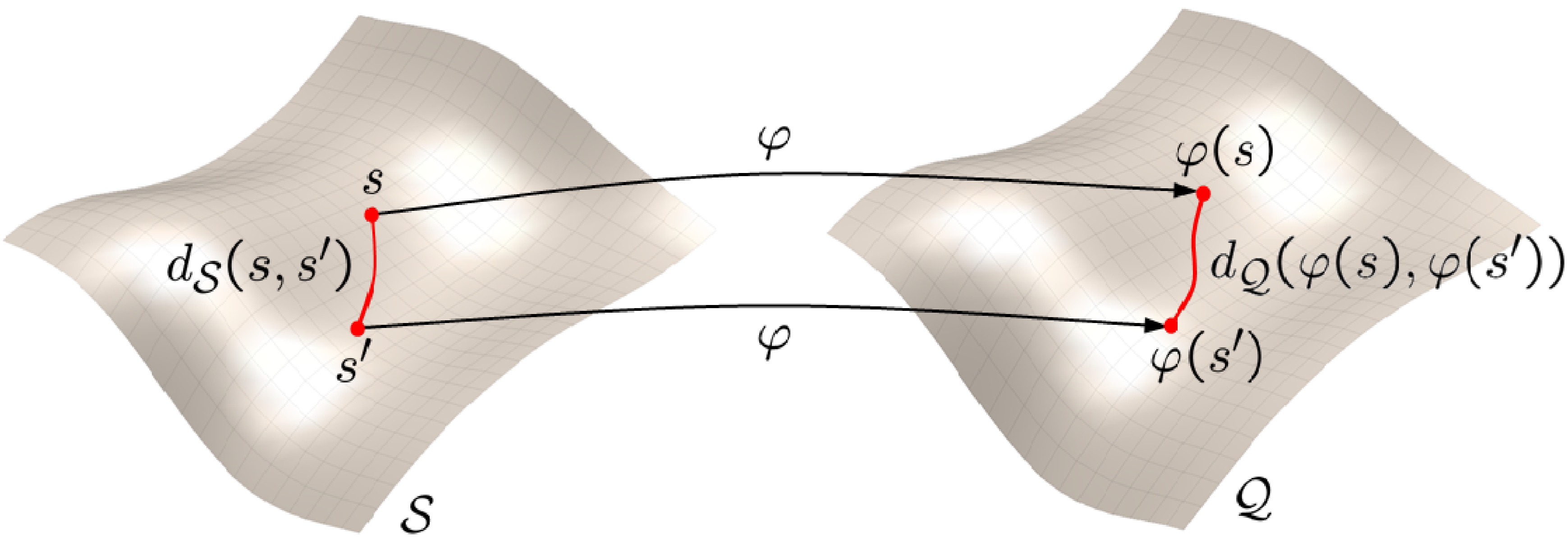
$$\mathrm{dis}\,\varphi = \sup_{s,s'\in\mathcal{S}} \left| d_{\mathcal{S}}(s,s') - d_{\mathcal{Q}}(\varphi(s),\varphi(s')) \right|$$

$$\mathrm{dis}\,\psi = \sup_{q,q'\in\mathcal{Q}} \left| d_{\mathcal{Q}}(q,q') - d_{\mathcal{S}}(\psi(q),\psi(q')) \right|$$

$$\mathrm{dis}\,(\varphi,\psi) = \sup_{s\in\mathcal{S},\, q\in\mathcal{Q}} \left| d_{\mathcal{S}}(s,\psi(q)) - d_{\mathcal{Q}}(q,\varphi(s)) \right|$$

Gromov-Hausdorff distance

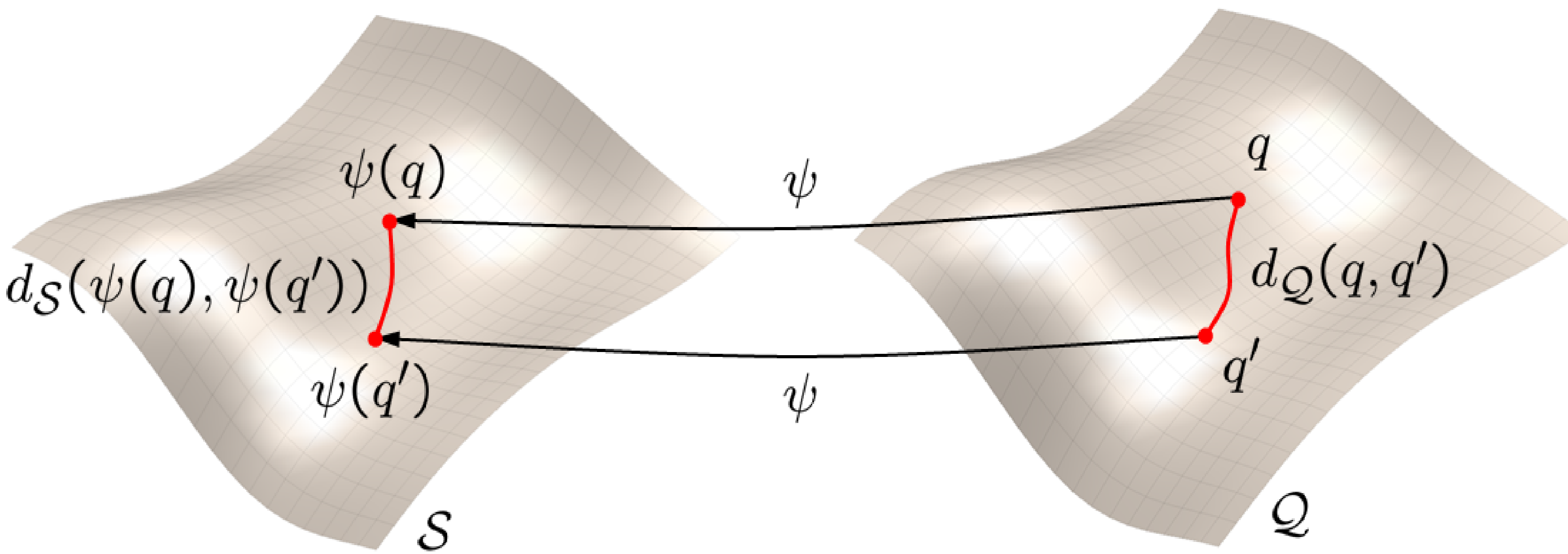
dis φ measures how isometrically can $\mathcal{S}$ be embedded into $\mathcal{Q}$



$$\text{dis}\,\varphi = \sup_{s,s'\in\mathcal{S}} \left| d_{\mathcal{S}}(s,s') - d_{\mathcal{Q}}(\varphi(s),\varphi(s')) \right|$$

Gromov-Hausdorff distance

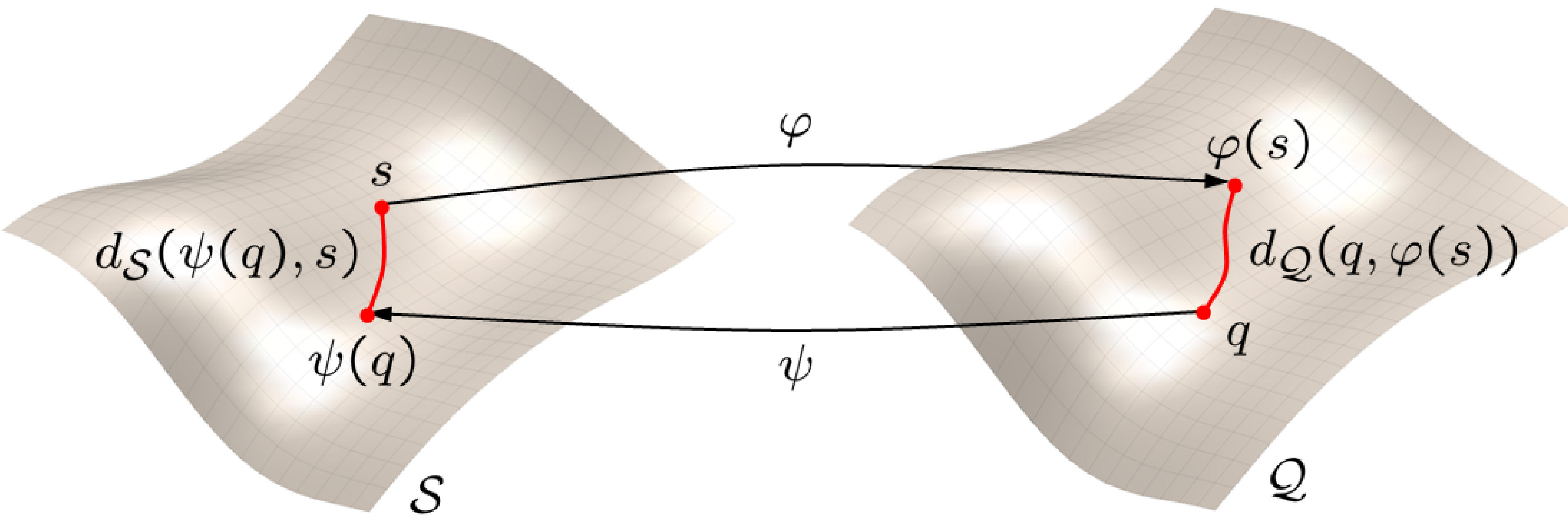
dis ψ measures how isometrically can $\mathcal{Q}$ be embedded into $\mathcal{S}$



$$\text{dis}\,\psi = \sup_{q,q' \in \mathcal{Q}} \left| d_{\mathcal{Q}}(q, q') - d_{\mathcal{S}}(\psi(q), \psi(q')) \right|$$

Gromov-Hausdorff distance

$\mathrm{dis}\,(\varphi, \psi)$ measures how far φ and ψ are from being one the inverse of the other



$$\mathrm{dis}\,(\varphi, \psi) = \sup_{s \in \mathcal{S},\, q \in \mathcal{Q}} \left| d_{\mathcal{S}}(s, \psi(q)) - d_{\mathcal{Q}}(q, \varphi(s)) \right|$$

References

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

[2]Hausdorff distance, Wikipedia encyclopedia,

http://en.wikipedia.org/wiki/Hausdorff_distance