Very high dimensional spaces are of simple structure

- Curse of dimensionality: no! In fact: Very high dimensional spaces are full of symmetries.
- In the (high dimensional) limit, everything has a simplex structure, which is a particular case of ultrametricity.
- For n points, m dimensions, we are concerned with (approximately) fixed n, but $m \to \infty$. See:
  - F. Murtagh, "On ultrametricity, data coding, and computation", Journal of Classification, 21, 167-184, 2004.
  - P. Hall, J.S. Marron and A. Neeman, "Geometric representation of high dimensions, low sample size data", JRSS B, 67, 427–444, 2005.
  - J. Ahn, J.S. Marron, K.E. Muller and Y.-Y. Chi, "The high dimension, low sample size geometric representation holds under mild conditions", Biometrika, to appear, 2007.
  - J. Ahn and J.S. Marron, "The direction of maximal data piling in high dimensional space", preprint, 2005.

- CC Aggarwal et al., "On the surprising behavior of distance metrics in high dimensional spaces", Proc. 8th Intl. Conf. on Database Theory, 4-6 Jan. 2001, pp. 420–434: "Recent research results show that in high dimensional space, the concept of proximity, distance or nearest neighbor may not even be qualitatively meaningful."
- Breuel, 2007 (citn. next slide): if nearest neighbor point y for some query point x has distance $d(x, y)$, then an ϵ-approximate nearest neighbor y' is s.t. $d(x, y') \leq (1 + \epsilon)d(x, y)$.
- "... the relationship between approximation and 'cost' of a solution need not be linear. For example, the cost of picking an ϵ-approximate nearest neighbor could be proportional not to the difference of distances between the optimal answer and the approximation, but to the volume of the shell between the two, that is, as $(1 + \epsilon)^{m-1}$, where m is the dimension of the space."

- *From the curse of dimensionality to the crisis of dimensionality.*
- Rather than nearest neighbor, determine ϵ-approximate one.
- Difference between query and nearest neighbor is $1 + \epsilon$, compared to exact solution.
- Great! But what is the cost of this approximation? Is it ϵ, or is it related to the volume between two shells, $(1 + \epsilon)^{m-1}$?
- Consider ratio of nearest to furthest neighbor. In high dimensions these tend to be equal! (We will see this in the following slides.)
- If $1 + \epsilon >$ this ratio, then any point is an ϵ-approximate neighbor.
- See: T. Breuel, "A note on approximate nearest neighbor methods", arXiv:cs/0703101 (2007). Or Beyer et al., "When is nearest neighbors meaningful?", ICDT, 217-235 (1999).

- Ahn et al.: "when $m >> n$, under a mild assumption, the pairwise distances between each pair of data points are approximately identical so that the data points form a regular n-simplex. In a binary classification setting, the training data from each class becomes two simplices ... any reasonable classification method will find the same [discriminant result] when m becomes very large."
- The mild condition for simplex structure formation, as $m \longrightarrow \infty$ is that directionality of the Gaussian cloud is "diffuse", defined in terms of eigenvalues:

$$\sum_{j}^{m} \lambda_j^2 / \left(\sum_{j}^{m} \lambda_j \right)^2 \longrightarrow 0 \text{ as } m \longrightarrow \infty$$

Then it is shown that the covariance matrix approaches a constant times the identity matrix.

- Why do we lay importance on the fact that the high dimensional simplex additionally defines an ultrametric topological embedding?
- Recall that ultrametric topology requires any triangle to be either (i) equilateral, or (ii) isosceles with small base.
- The equilateral case corresponds fine with the simplex structure.
- But it is useful to us to hang on to the isosceles with small base case, too, for inter-cluster relationships.
- We will look later at examples to support this.

- D.L. Donoho and J. Tanner, "Neighborliness of randomly-projected simplices in high dimensions", Proc. Natl. Acad. Sci., 102, 9452–9457, 2005.
- For a Gaussian cloud, "not only are the points on the convex hull, but all reasonable-sized subsets span faces of the convex hull".
- Intuitively, if all points fly apart from one another as dimensionality grows, then (i) each point is a vertex of the convex hull of the cloud of points; (ii) each pair of points generates an edge of the convex hull; and (iii) sets of points form a regional face of the convex hull.
- Conclude: "This is wildly different than the behavior that would be expected by traditional low-dimensional thinking."