Volume-based Landmark Selection For Dimensionality Reduction

FC+AR

July 16, 2008

1 Introduction

Problem The problem we are trying to address is as follows : we have a point set of a very large size N. This is in a high-dimensional space of dimension D. We are trying to bring this down to N points of dimension K where $K < D$.

Method The method we are trying to use is as follows : choose K of these N points as *landmarks*. Map each D-dimensional point, $\{p_i\}$, to a K-dimensional point (a K-tuple), $\{p'_i\}$, where the i-th co-ordinate in the tuple is the distance of the point p from the i-th landmark.

Error This method leads to some error. This is how the error is measured : each of the landmarks is known in R^d. p'_i represents a D-dimensional sphere in R^D since we know that the point is $|p'_i|$ (radius) away from the first landmark (centre). So, we have K D-dimensional spheres for each point $\{p'_i\}$. The surface-area of the $(D-K+1)$-dimensional sphere so formed will be the set of all possible points that p could have been, and hence is a measure of the error due to one point p_i. The total error can be measured in different ways (L_p norms) which will be talked about later.

Aim What we want to do is choose landmarks intelligently and quickly, so as to minimize the total error.

2 Clusters

Definition A cluster (C_i) is defined by the following assumptions : it is a sphere centred at centre c_i and radius R_i, and has w_i points distributed with density ρ_i. Let the number of clusters be N_C. The landmark, if chosen from a cluster, is chosen to be the centre of the cluster.

2.1 Motivation Example

The crosses on the map mark the distance of the point from the landmark (0,0,0). The line is to be read as (Error +- Standard Deviation), with the centre of the line marking the average error if that particular point is the second landmark.

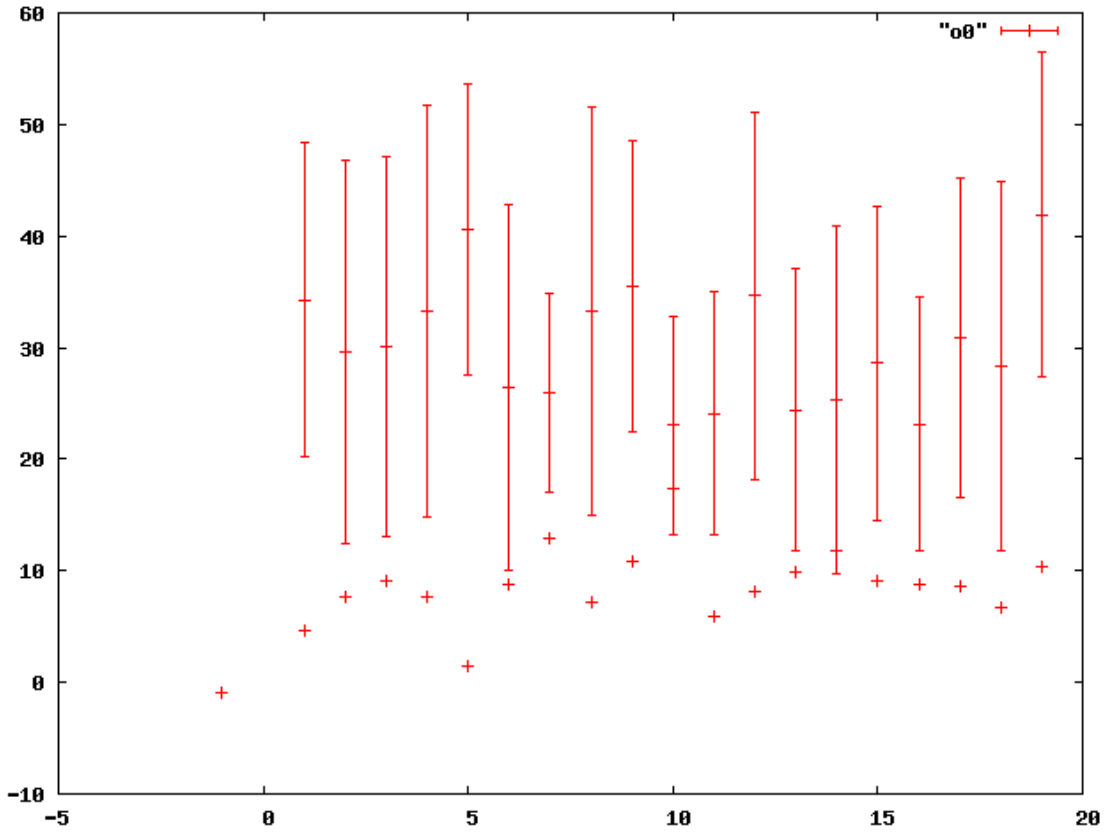

Figure 1: Example - random points from $(0, 0, 0)$ to $(10, 10, 10)$

In Figure 1, we have 20 randomly distributed points from $(0, 0, 0)$ to $(10, 10, 10)$. We choose $(0, 0, 0)$ as one of the landmarks, and see how the error behaves with respect to distance. Here are some observations.

When paired with 5, the distance is low, and the error is terribly high.

When paired with 19, the distance is high, and the error is terribly high.

When paired with 14, the distance is high, and the error is low.

When paired with 11, the distance is low, and the error is low.

We get similar results by fixing the first landmark as some other point as well.

Now, in Figure 2, we have 20 points distributed evenly into 3 clusters, centered at $(0, 0, 0)$, $(10, 10, 10)$ and $(5, 25, 14)$. Again, we choose $(0, 0, 0)$ as one of the landmarks. Here are some observations.

When paired with a point from 1-6 (C_1), the error is clearly higher than the rest.

When paired with a point from 7-13 (C_2), the error is clearly low and almost constant.

When paired with a point from 14-19 (C_3), the error is the lowest and almost constant.

So, the error is clearly lowest when we choose the two landmarks from different clusters. Also, the error was lower when we chose the landmarks from the clusters farthest apart. It also turned out that the dip in error when 0 was paired with a point from C_1 was because of collinearity, something we will look at afterwards.

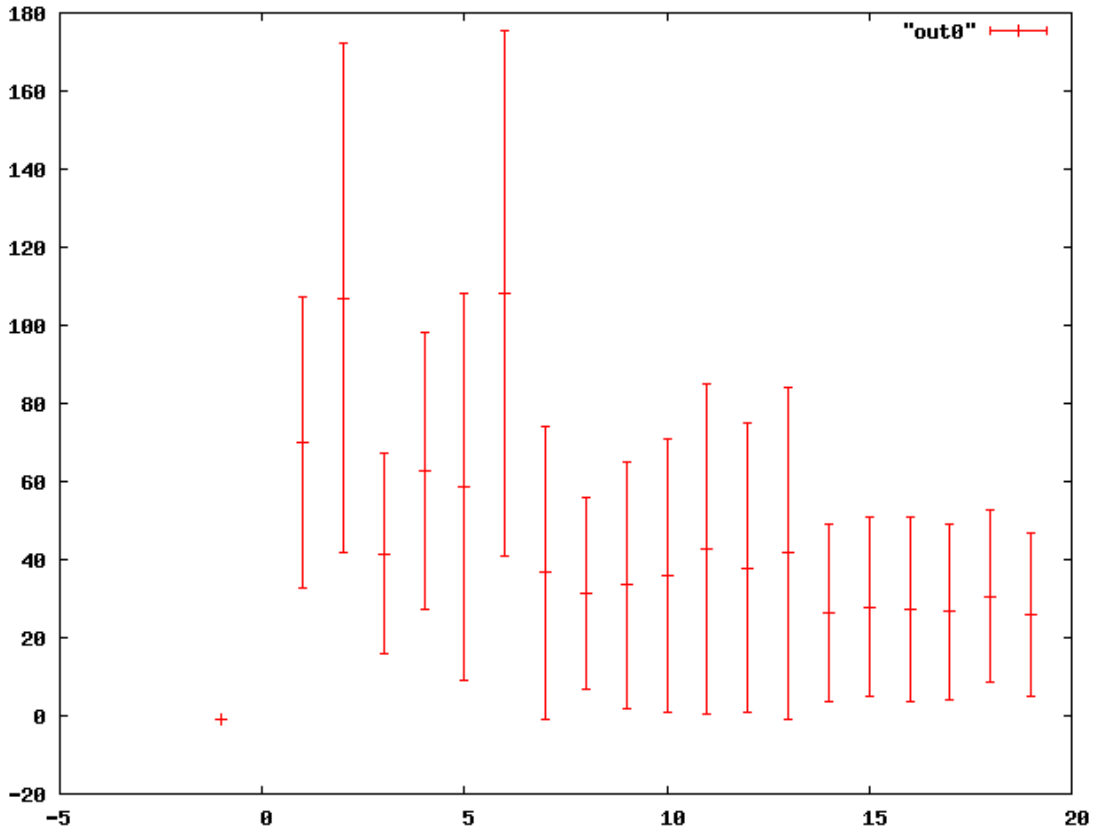

Figure 2: Example - three clusters centred at $(0, 0, 0)$, $(10, 10, 10)$ and $(5, 25, 14)$

Here are a list of the types of data sets used for experimentation, to get some intuitions:

- A linear point cloud distribution (all points lie on a straight line).
- A random point cloud in a cube with opposite corners $(-10, -10, -10)$ and $(10, 10, 10)$.
- A random sparse point cloud in a cube with opposite corners $(-100, -100, -100)$ and $(100, 100, 100)$.
- A 2-clustered point cloud around $(0, 0, 0)$ and (d, d, d) for different d.
- A 3-clustered point cloud around $(0, 0, 0)$, $(10, 10, 10)$ and $(5, 25, 14)$.
- An ellipsoidal point cloud distribution symmetric around $(0, 0, 0)$.

Some preliminary testing for the best pair of landmarks seemed to show consistent results for clustered point clouds. However, for sparse and random clouds, the results were interesting, but inconsistent. Hence, we shall first consider point clouds in clusters.

Sometimes, we may make further assumptions, like ρ_i being uniform, or the distance d_{ij} between two clusters C_i and C_j has the property $d_{ij} >> R_i, R_j$, but we will specify when it is required.

3 Classifications & Terminology

3.1 Dimensions

For any problem that is being solved, we will mention the original point space's (P) dimension D, the reduced point space's (P') dimension K, the number of points (N), and the number of clusters N_C

involved (if any).

3.2 Landmark Space

Discrete The landmarks can be only points from the original point set.

Continuous The landmarks can be any points in the D-space.

3.3 Error Type

First we shall consider single point estimators. With respect to an estimator point p, L_i is defined as the i-th root of the sum of the i-th powers of distances from the points (say $x \in P$) to p (say d_{xp}).

$$L_i = \sqrt[i]{\sum_{x \in P} d_{xp}^i}$$

L_∞ - Circumcenter Here as i tends to infinity, only the largest term of the summation counts. Hence L_∞ refers to the distance of the furthest point from the estimator point. Hence, minimizing L_∞ is basically *minimizing the maximum error.*

L_2 - Centroid This error refers to the root mean square error (RMS) involved. The centroid results in the LMS (Least Mean Square) error.

L_1 - Median/Fermat-Weber Point The total error is calculated just as the sum of the individual errors. Hence, the median or Fermat-Weber point minimizes the total sum of errors.

Note (Collinearity) If all the points lie on a line, then it doesn't matter which points you choose as landmarks - the total error is always zero since the spheres will always kiss. In other words, if there are some collinear points, choosing even two out of the K landmarks as points on the line will lead to zero error contribution for each of those points.

4 Solved Problems

4.1 Lemma 1 - $D = 2, K = 1, L_2$, continuous

The required point is nothing but the centroid of the data since all we need to do is minimise the root mean square error (it is a well-known result).

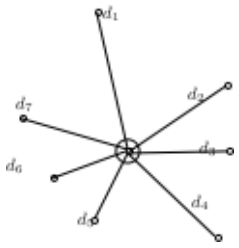

Figure 3: Centroid minimizes Σd_i^2 (Lemma 1)

Complexity $O(N)$

4.2 Lemma 2 - $D = 3, K = 2, N_C = 2, L_1,$ discrete

The total error is independent of the distance between the clusters.

Proof Let the clusters be C_1 and C_2. So the landmarks are their centres c_1 and c_2. The error due to a point is going to be the perimeter of the circle of intersection of the 2 $3D$-spheres formed due to that point.

Considering the z-axis to be the line joining c_1 and c_2. The error due to one point in C_1, which is at (r_1, θ, ϕ) is just going to be a circle, perpendicular to the z-axis, inside C_1 whose radius is $r_1.sin\phi$.

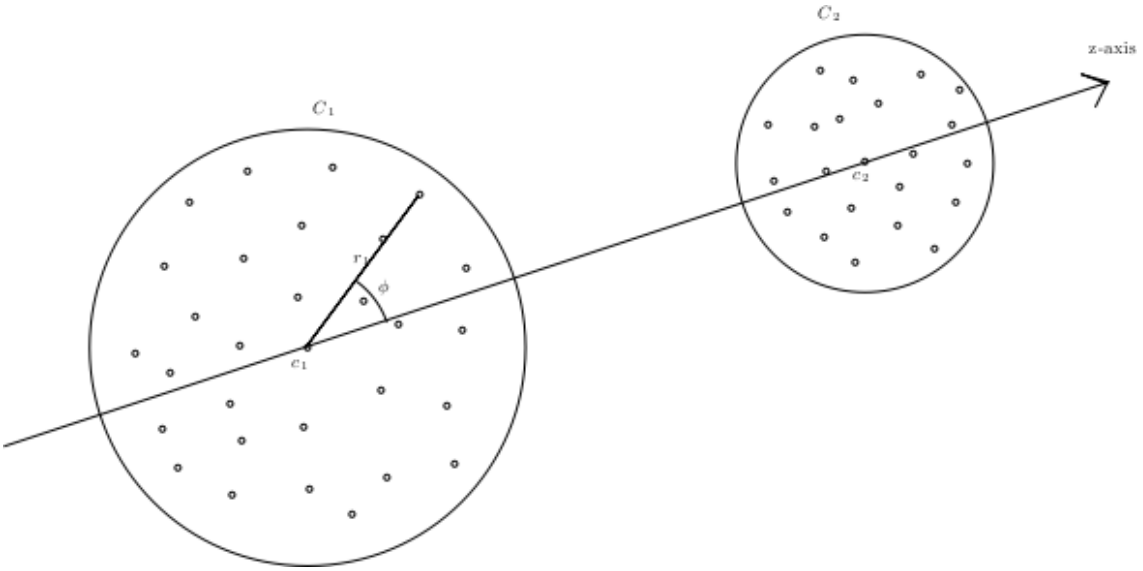

Figure 4: The 2 Cluster Case (Lemma 2)

Hence the total error due to C_1 is

$$E_1 = \int_\theta \int_\phi \int_{r_1} \rho_1 \; 2\pi r_1 sin\phi \; r_1^2 \; dr_1 \; sin\phi \; d\phi \; d\theta$$

which is basically independent of d_{12} whatever ρ_1 may be. Similarly E_2 is also independent of d_{12}. Therefore, the same holds for $E = E_1 + E_2$. Hence proved.

Verification Tried out two cases. In one, they were centred around (0,0,0) and (10,10,10). In the other, they were centered exactly the same way around (0,0,0) and (100,100,100). They lead to exactly the same average error of 7.62978. This proves the point above, and this will become clearer when we view the problem from the following different perspective.

Different Perspective Note that the circle of intersection of the two spheres formed due to one point is perpendicular to the line joining the two landmarks. Hence, it's radius is EXACTLY the perpendicular distance from the point to the line. This perpendicular distance is independent of where the clusters lie - just depends on the distribution of points within the cluster. Hence proved.

4.3 Lemma 3 - $D = 3, K = 2, N_C = 3, L_1$, discrete

Assume that all ρ_is are uniform and equal, clusters are equal in size, and that $d_{ij} >> R_i, R_j$. Then, the total (L_1) error is minimised when d_{ij} is maximised. In other words, if all clusters have approximately equal number of points, then we should choose the centres of the farthest two clusters as our landmarks.

Some Maths We shall calculate the radius of the circle of intersection of 2 spheres. Assume 2 intersections spheres of radii r_1 and r_2, with centres c_1 and c_2 and distance between centres d_{12}. Take one of the points, P, on the sphere. Let the angle Pc_2c_1 be θ. Then,

$$cos\theta = \frac{(r_2^2 + d_{12}^2 - r_1^2)}{2.r_2.d_{12}}$$

The required radius is $r_2.sin\theta$, which simplifies to a simple symmetric expression,

$$r_2.sin\theta = \frac{\sqrt{(d_{12} + r_1 + r_2)(d_{12} + r_1 - r_2)(d_{12} - r_1 + r_2)(-d_{12} + r_1 + r_2)}}{2.d_{12}}$$

Proof Let the clusters be C_1, C_2 and C_3. Assume the landmarks to be c_1 and c_2. Using a similar proof as above, $E_1 = e_1$ and $E_2 = e_2$ are independent of d_{12} and are constants.

$$e_1 = \frac{1}{2}\pi^3\rho_1 r_1^4$$

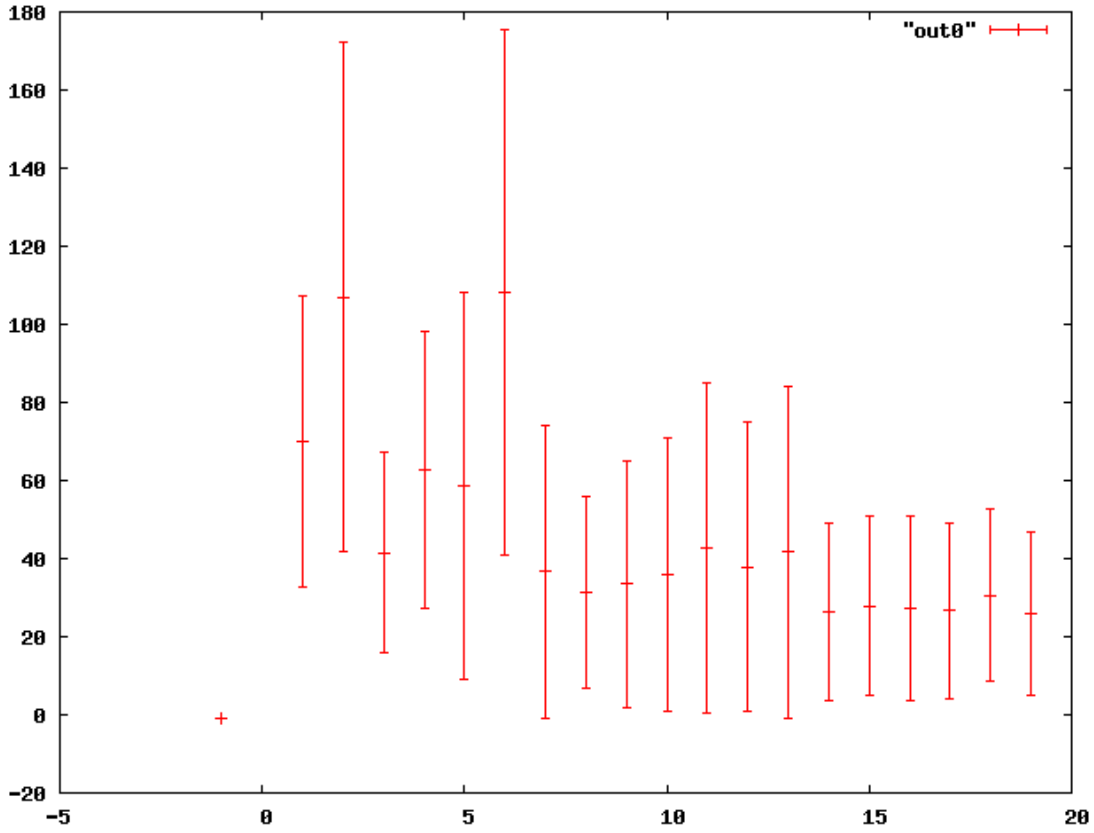

Figure 5: Verification V1 - One center is in 1st cluster

$$e_2 = \frac{1}{2}\pi^3 \rho_2 r_2^4$$

$$\rho = \frac{w}{\frac{4}{3}\pi r^3}$$

Let us calculate the error due to C_3. Given any point in C_3, at a distance r_1 and r_2 from the two centers, we can use the above calculated formula for the radius of intersection of the 2 spheres. Approximating r_1 as d_{13} and r_2 as d_{23}, and using the expression for all w_3 points, we get :

$$E_3 = w_3.2\pi.\frac{\sqrt{(d_{12}+d_{13}+d_{23})(d_{12}+d_{13}-d_{23})(d_{12}-d_{13}+d_{23})(-d_{12}+d_{13}+d_{23})}}{2.d_{12}}$$

We define

$$S_{ijk} = \sqrt{(d_{ij}+d_{ik}+d_{jk})(d_{ij}+d_{ik}-d_{jk})(d_{ij}-d_{ik}+d_{jk})(-d_{ij}+d_{ik}+d_{jk})} \quad (i \neq j \neq k)$$

It is defined to be 0 otherwise. Then

$$E_3 = \frac{w_3.\pi.S_{123}}{d_{12}}$$

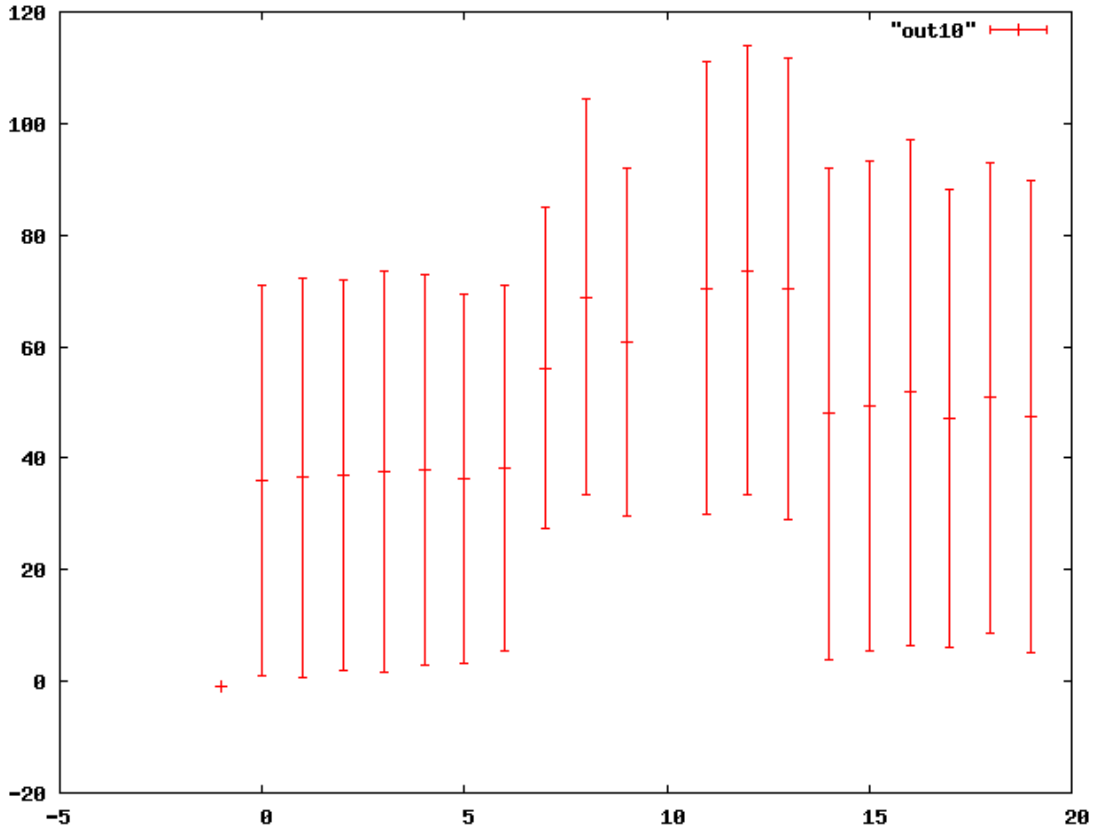

Figure 6: Verification V1 - One center is in 2nd cluster

Therefore

$$E = \frac{3}{8}\pi^2 w_1 r_1 + \frac{3}{8}\pi^2 w_2 r_2 + \frac{w_3.\pi.S_{123}}{d_{12}}$$

Since the clusters are assumed to be equal in size and equally dense, we have

$$E_{ij} = \frac{3}{4}\pi^2 wr + \frac{w\pi S_{ijk}}{d_{ij}}$$

Since $N_C = 3$, the term S_{ijk} is actually S_{123} and is common to E_{12}, E_{23}, E_{13}. Hence, the total error is minimised when the d_{ij} is maximised.

Verification V1 We look at the graphs of figures 3,4,5. We note that the clusters centers are at $(0, 0, 0)$, $(10, 10, 10)$ and $(5, 25, 14)$. Clearly, the largest distance is between clusters 1 and 3. From the graphs, when a landmark in cluster 2 is chosen, it does worse when paired with a landmark from both 3 or 1 as compared to when the landmarks are from clusters 1 and 3. Hence it practically agrees with the above claim.

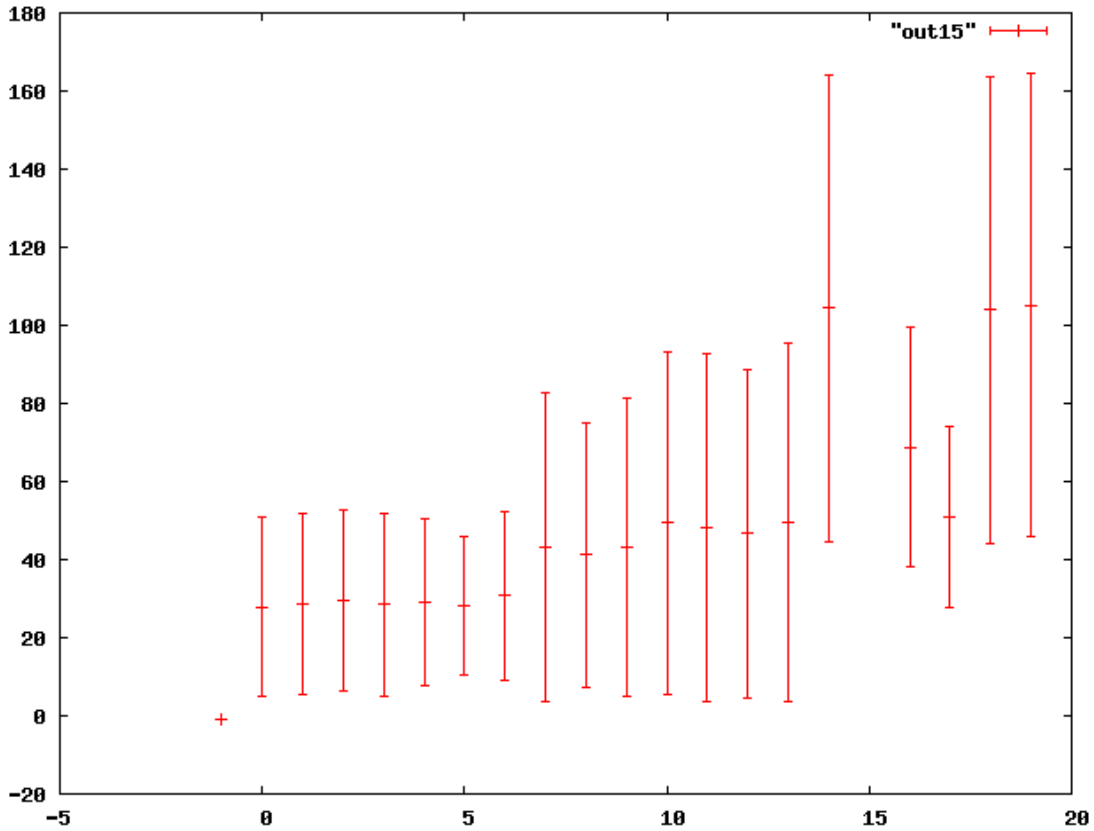

Figure 7: Verification V1 - One center is in 3rd cluster

Complexity It seems like there is no better way to solve this problem than to check all possible sets of 3 landmarks and choose the minimum which would take $O(n^3)$ time.

4.4 Problem 4 - $D = 3, K = 2, N_C = K, L_1$, discrete

Now, assuming that c_i and c_j are used as landmarks, we have

$$E_{ij} = e_i + e_j + \sum_{k=1}^{K} \frac{\pi.w_k.S_{ijk}}{d_{ij}}$$

$$E_{min} = min_{i \neq j} E_{ij}$$

Complexity It presently seems that the best method to solve the above optimization problem is the brute force method, taking $O(K^3)$.

4.5 Lemma 5 - $D = 3, K = 2, L_1$, continuous

Let us look at the problem in the alternate perspective as was suggested earlier. We are looking for a line L in 3D which minimizes the total sum of the perpendicular distances from points in P to it. Such a problem has already been solved. The algorithm (thanks to Dr. Yves Nievergelt) follows below.

Algebraically, the line is parallel to the right-singular vector for the largest singular value. In other words, the line is the intersection of the planes perpendicular to the right-singular vector for the smallest (TLS plane) and middle singular values.

First, fit a Total Least-Squares plane to the data. Second, project the data orthogonally onto the fitted plane. Finally, fit a Total Least-Squares line to the projections of the data in the plane.

Complexity $O(N)$

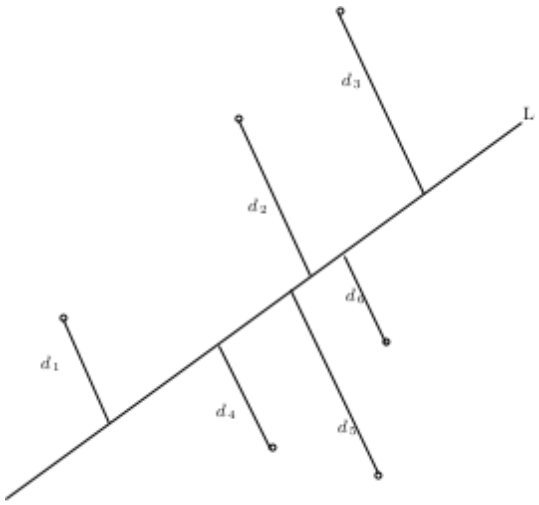

Figure 8: L minimizes Σd_i (Lemma 5)

4.6 Lemma 6 - $D = 3, K = 2, L_2$, continuous

Using the analogy in the previous case, we are looking for a line L in 3D which minimizes the total sum of the squares of perpendicular distances from points in P to it. This is nothing but the LMS/TLS line of P, which is a standard solved problem (and interestingly, passes through the centroid).

Complexity $O(N)$

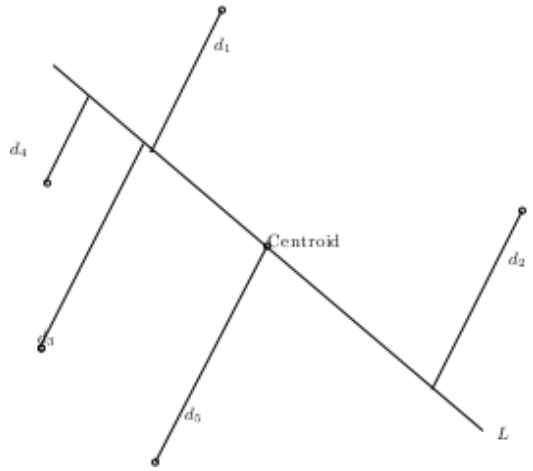

Figure 9: The LMS line L minimizes Σd_i^2 (Lemma 6)

4.7 Lemma 7 - $D = 4, K = 2, L_1$, continuous

Here each 4-D point i is mapped to a 2-tuple (a, b) using the distances to the two landmarks L_1, L_2. Hence, when we move back to 4D, we get two intersecting 3-spheres (S_a, S_b) of radius a and b. This intersection is a 2-sphere S_2 (the surface of a 3-sphere), which is perpendicular to the 4-D line L joining the two landmarks and equidistant from it.

The radius involved is the perpendicular distance from the S_2 to L, which is the perpendicular distance d_{iL} from the point i to L. Each error is proportional to d_{iL}^2, and hence we are looking for a line L' which minimizes $\Sigma d_{iL'}^2$. This, again, is nothing but the LMS/TLS line of P, which is a standard solved problem.

Complexity $O(N)$

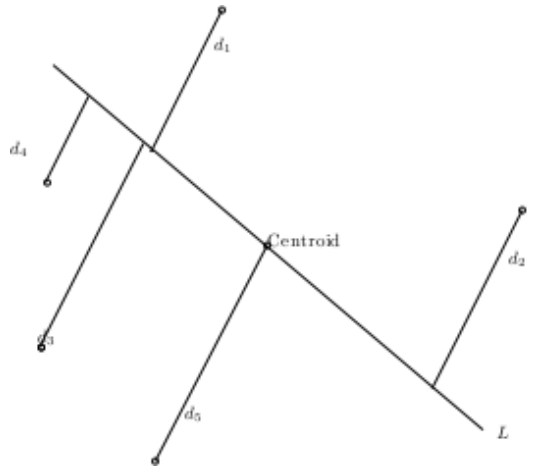

Figure 10: The LMS line L minimizes Σd_i^2 (Lemma 6)

4.8 Lemma 8 - $D = D, K = 1, L_\infty$, continuous

In [GÖ99], the problem of finding the smallest enclosing ball in any dimension has already been addressed (ie, finding the circumcenter of the point set to miminize L_∞. Hence this problem has also been solved.

Complexity Don't know!!!

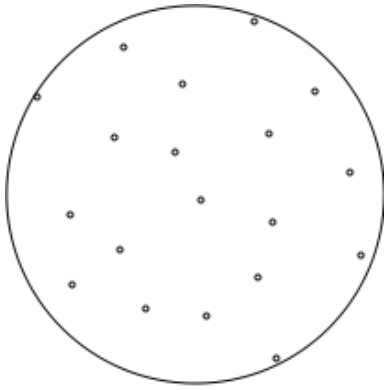

Figure 11: Minimum Enclosing Ball - L_∞ minimized (Lemma 8)

4.9 Lemma 9 - $D = 2, K = 1, L_\infty$, discrete

We can use the "farthest-neighbour Voronoi diagram". In the Voronoi diagram of the given points, let $f(p)$ be defined as the point which is farthest from p. The required point is that with minimum distance between p and $f(p)$.

Complexity $O(NlogN)$

4.10 Lemma 10 - $D = 3, K = 1, L_1$, continuous

Since we know the one landmark point L, after dimensionality reduction, the co-ordinate value of each point i is the distance to it in 3D. Hence, the point can lie anywhere on a sphere centered at the landmark and radius equal to the co-ordinate value of that point. Hence, the uncertainity is now the surface area of the sphere. The sum of all the errors is proportional to Σd_{iL}^2). The centroid is well known to minimise this value (the point which leads to LMS).

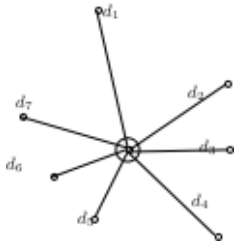

Figure 12: Centroid minimizes Σd_i^2 (Lemma 10)

Complexity $O(N)$

5 Survey Of Related Problems

Several problems regarding central points and different error measures have been discussed in literature, and here is a brief review of some results.

5.1 2-Center problem in 2-D (continuous)

Problem Given a point set P of N points on the plane, find 2 congruent discs, who's centers need not be in P, of smallest radius that cover all the points.

Complexity Randomized expected time $O(nlog^2n)$.

Reference [ASW97]

5.2 2-Center problem in 2-D (discrete)

Problem Given a point set P of N points on the plane, find 2 congruent discs, who's centers are in P, of smallest radius that cover all the points.

Complexity $O(n^{4/3}log^5n)$ time.

Reference [ASW97]

Note Turns out that the continuous problem is easier than the discrete problem. Why? - Imagine that you have fixed one disk of radius r. The possible location for the other disk is in the intersection of disks of radius r for all the points not covered by the first disk. It is much easier to check if this area is nonempty (you can choose any point here in the continuous case) than to see if a point lies within this area (which you would require for the discrete case).

5.3 k-Median Problem in Polygons

Problem Place k centers into a polygonal region P with holes, such that the overall average distance of all points $p \in P$ to their respective closest centers is minimised.

Complexity Proved to be NP-Hard for any general k.

Reference [FMW00]

More Algorithmic results for a continuous set of demand locations - for L_1 distances in the plane, we can determine an optimum center in $O(n)$ time for geodesic distances in simple polygons, to $O(n^2)$ for straight line distances in general polygonal regions, and $O(n^4)$ for geodesic distances in polygons with holes.

5.4 Fast Approximation to Fermat-Weber Problem

Theorem Given a set P on N points in $\mathbb{R}^D$ (fixed D), in deterministic $O(kNlogN)$ time and $O(kN)$ space, a point p' can be computed such that the value of $w(p')$ ($w(p') = \sum_{x \in P} d_{xp'}$) satisfies $(1-\epsilon)w(p) \leq w(p') \leq (1+\epsilon)w(p)$ where the point p minimizes $w(p)$, and k is a function of ϵ. The point p' can be computed with high probability in expected $O(n)$ time and space.

Reference [BMM03]

Relation For dimensionality reduction from dimension D to 1 for any fixed D, this gives a fast ϵ-approximation to the actual error.

5.5 Bregman Balls

Problem Find an approximation to the smallest enclosing ball of a D-dimensional point set P using a different distance metric called the Bregman divergence framework.

Complexity $O(\frac{DN}{\epsilon^2})$ for a $(1+\epsilon)r^*$ approximation, where r^* is radius of the smallest enclosing ball of the point set P.

Reference [NN06]

Relation Solves an approximation of the minimum enclosing ball problem with a different distance metric. Basically the Euclidean distance might not be a good reflection of the actual distance between D-dimensional points.

5.6 Center-Points

Problem Given a point set P of N points in a D dimensional space, the center point is a point x (not necessarily in the point set) such that any half-plane not containing x must have less than $\frac{DN}{D+1}$ points of P. In other words, every such half-plane must have at least $\frac{N}{D+1}$ points. Such a center point always exists.

Reference Algorithms in Combinatorial Geometry, by H. Edelsbrunner.

Note There doesn't seem to be a method to calculate such center-points, though they are proven to exist.

5.7 Robustness

The breakdown point is the proportion of data that must be moved to infinity so that the estimator will do the same. For example, in $\mathbb{R}^{\mu\nu}$, the median has a breakdown of $\frac{1}{2}$ while the mean has a breakdown of $\frac{1}{N}$. It has been shown that the maximum breakdown is $\frac{1}{2}$, so the median does well according to this robustness criterion. More importantly, if we are dealing with error-prone data like readings of protein energies, the median will keep you safest.

Reference Geometric Measures of Data Depth, by George Aloupis.

5.8 Convex-Hull Peeling

Different Definition A possible definition for a univariate median is to remove pairs of extreme points. Though I wonder if this will lead to any approximations. In larger dimensions, we can use a method called convex hull peeling.

Problem Compute the final point(s) remaining after the process of convex-hull peeling a point set P of N points.

Complexity Brute force is $O(N^2 logN)$ in $\mathbb{R}^{\nu}$. This has been improved over time to $O(NlogN)$.

Reference Geometric Measures of Data Depth, by George Aloupis

Relation Good alternate definition (intuitive). However, no relations yet with error norms. Also, the breakdown point of these methods cannot exceed $\frac{1}{D+1}$ in $\mathbb{R}^{\mathbb{D}}$.

References

[ASW97] Pankaj K. Agarwal, Micha Sharir, and Emo Welzl. The discrete 2-center problem. In *SCG '97: Proceedings of the thirteenth annual symposium on Computational geometry*, pages 147–155, New York, NY, USA, 1997. ACM.

[BMM03] Prosenjit Bose, Anil Maheshwari, and Pat Morin. Fast approximations for sums of distances, clustering and the fermat–weber problem. *Comput. Geom. Theory Appl.*, 24(3):135–146, 2003.

[FMW00] Sándor P. Fekete, Joseph S. B. Mitchell, and Karin Weinbrecht. On the continuous weber and k-median problems (extended abstract). In *SCG '00: Proceedings of the sixteenth annual symposium on Computational geometry*, pages 70–79, New York, NY, USA, 2000. ACM.

[G$\ddot{9}$9] Bernd Gärtner. Fast and robust smallest enclosing balls. In *ESA '99: Proceedings of the 7th Annual European Symposium on Algorithms*, pages 325–338, London, UK, 1999. Springer-Verlag.

[NN06] Frank Nielsen and Richard Nock. On approximating the smallest enclosing bregman balls. In *SCG '06: Proceedings of the twenty-second annual symposium on Computational geometry*, pages 485–486, New York, NY, USA, 2006. ACM.