How Bad Can Your Data Be? Convexity and Variance Maximization

Jeffrey L. Stuart[1]
Department of Mathematics, Pacific Lutheran University, Tacoma, WA 98447 USA[1]

Abstract

If a sample of m points (repetition allowed) is taken from a set T in n-dimensional space, it is well known that the minimum possible variance is zero, and that the minimum variance occurs precisely when all m points coincide. We investigate the maximum possible variance for a sample of size m. The maximum variance is well defined when T is closed and bounded (that is, T is compact). Our results provide bounds in terms of the maximum variance for the convex hull of T and for circumscribing $n-$spheres and n-cubes.

Keywords: variance maximization.

1 Introduction

Given a nonempty set T in n-dimensional space, select a sample S of consisting of m independently chosen random points (possibly repeated), and compute the variance for this sample. Everyone knows that if you are very, very lucky, the points will all coincide, and the variance will be zero, the absolute minimum possible. What if you were very unlucky? How big could the variance be?

Let S be the sample $S = \left\{X^{(1)}, X^{(2)}, \ldots, X^{(m)}\right\} \subseteq T \subseteq \mathbf{R}^n$. Since repetition is allowed, a sample can be a multiset rather than a set. Indeed, the sample S could consist of m copies of a single point from T. The mean of S is defined by

$$\overline{X}_S = \frac{1}{m}\sum_{j=1}^{m} X^{(j)}.$$

The variance of S is defined by

$$Var(S) = \frac{1}{m-1}\sum_{j=1}^{m} \left\|X^{(j)} - \overline{X}_S\right\|^2.$$

The sample S is called a *variance maximizing sample* for T if it maximizes the variance over all samples of size m from T.

The *n-sphere B centered at P with* radius $r > 0$ is the set $B = \{X \in \mathbf{R}^n : \|X - P\|_2 \leq r\}$.

The n-cube C centered at P with edge length $2s > 0$ is the set $C = \{X \in \mathbf{R}^n : p_i - s \leq x_i \leq p_i + s$ for $i = 1, 2, \ldots, n\}$.

2 Compactness

There are two natural restrictions on the nonempty set T:

- T is bounded. (Otherwise, choose sequence of sets S so that one point runs towards infinity, and variance will grow without bound).
- T is closed. (Otherwise , choose sequence of sets S so that they converge to a variance maximizing limit point set)

Equivalently, T is compact. Compactness enables us to measure the size of T. Specifically:

- The set diameter $diam(T)$ defined by $diam(T) = max\{||X - Y||_2 : X, Y \in T\}$ is well-defined and finite.
- The set radius $rad(T)$ defined as smallest radius of an n-sphere containing T is well-defined and finite.

The points X and Y in a compact set T are called a *diametric pair of points for* T if $diam(T) = \|X - Y\|_2$. If X and Y are a diametric pair, the line segment between X and Y is called a *diametric segment* of T.

Theorem 1 *Let T be a nonempty, compact set. Then $diam(T) \leq 2rad(T) \leq c \cdot diam(T)$ where $c < 2$ and c depends only on n (not on T). Further, there exists an n-sphere B of radius $r = \frac{c}{2}diam(T)$ such that$T \subseteq B$, and consequently,* $\max\{Var(S) : S \subseteq T\backslash\} \leq \max\{Var(S) : S \subseteq B\}$.

More generally, the maximum variance for a sample from a compact set T is bounded above by the maximum variance for a sample from any n-sphere, n-cube or other compact set containing T. It is also bounded above by the maximum variance of the convex hull of T. Among all maximizing samples S, there are several special types: samples S trapped in a low dimensional affine subspace such as a line or plane, and samples S whose convex hulls have maximal dimension (such as the four vertices of a tetrahedron in $\mathbf{R}^3$). Other variance maximizing samples of interest are those that maximize some interpoint distance measure such as the average distance or sum of squared distances.

3 Convexity

The nonempty set U is called *convex* if $aX + (1-a)Y \in U$ whenever $X, Y \in U$ and $0 \leq a \leq 1$. The point $Z \in U$ is called an *extreme point* for the convex set U if $Z = aX + (1-a)Y$ for some $X, Y \in U$ only if $a = 0$ or $a = 1$. The *convex hull* of a nonempty set U, denoted $conv(U)$, is the set $conv(U) = \{aX + (1-a)Y : X, Y \in U$ and

$0 \leq a \leq 1\}$. Examples of convex sets include n-cubes and n-spheres. The extreme points of an n-cube are the 2^n corners. The extreme points of the n-sphere are all boundary points. The convex hull of a circle is the disk. The convex hull of three noncolinear points is the triangle with those points as its corners (including its interior). The following proposition summarizes the properties of the convex hull.

Theorem 2 *Let T be a nonempty set in n-space.*

- *$conv(T)$ is contained in every convex set containing T.*
- *$T \subseteq U$ implies $conv(T) \subseteq conv(U)$.*
- *$T = conv(T)$ if and only if T is convex.*
- *$conv(T)$ is compact when T is compact.*
- *$conv(T)$ is the convex hull of the extreme points of T when T is compact.*
- *$rad(T) = rad(conv(T))$ when T is compact.*
- *$diam(T) = diam(conv(T))$ when T is compact.*
- *$diam(conv(T)) = \max ||X - Y||_2$ where the maximum is over all pairs of extreme points X, Y of T when T is compact.*

4 Variance Maximizing Sets

Lemma 3 *Let S be a variance maximizing sample for a convex, compact set T. If T contains more than one point, then no point in S coincides with the mean $\overline{X}_S$ of S.*

Theorem 4 *Let T be a nonempty, convex, compact set in n-dimensional space. If S is a variance maximizing sample for T with m points, then all of the points in S must lie on the boundary of T.*

4.1 Samples of Even Size

When the sample size is even, the variance is maximized by choosing a sample consisting of certain diametric pairs of points.

Theorem 5 *Let T be a nonempty, compact set in n-space. Let m be an even, positive integer. Let $X, Y \in T$ satisfy $diam(T) = ||X - Y||_2$. Let S be the sample consisting of $\frac{m}{2}$ copies of X and $\frac{m}{2}$ copies of Y. Then S is a variance maximizing sample for T.*

Theorem 6 *Let T be a compact set such that every diametric segment for T intersects at a common point. When m is even, any collection of $\frac{m}{2}$ diametric pairs of points for T is a variance maximizing sample for T.*

Corollary 7 *Let m be a positive, even integer. Any collection of $\frac{m}{2}$ diametric pairs of points on the boundary of the n-sphere is a variance maximizing set for the n-sphere. The maximum variance is for the n-sphere of radius r is*

$$\frac{mr^2}{m-1}.$$

Corollary 8 *Let m be a positive, even integer. Any collection of $\frac{m}{2}$ pairs of antipodal corner points on the n-cube is a variance maximizing set for the n-cube. The maximum variance for the n-cube with edge length $2s$ is*

$$\frac{mns^2}{m-1}.$$

4.2 Samples of Odd Size

When the sample size m is odd, choosing diametric pairs of points leaves a single, unbalanced point. Even with $\frac{m-1}{2}$ carefully selected diametric pairs of points from a nonempty, convex, compact set T, there may be no choice for the remaining point that yields a variance maximizing set. Consequently, there are few general results for the odd case.

Theorem 9 *Let m be an odd, positive integer. Let n be a positive integer with $n \geq 2$. Any collection of m points spaced equidistantly around a great circle on the boundary of the n-sphere is a variance maximizing sample for the n-sphere. The maximum variance for the n-sphere with radius r is*

$$\frac{mr^2}{m-1}.$$

Notice that the variance maximizing samples in the preceding result are contained in a 2-dimensional subspace of $\mathbf{R}^n$ independent of how large n is. Under certain, restricted circumstances, it is possible to find n-dimensional variance maximizing samples for the n-sphere.

Lemma 10 *Let n be is a positive, even integer, and let $m = n + 1$. Then the vertices of the regular n-simplex are a variance maximizing sample for the n-sphere that circumscribes the n-simplex.*

The maximum variance for an even sample for the n-cube differs from that for an odd sample:

Theorem 11 *Let m be an odd, positive integer. Any collection of $\frac{m-1}{2}$ pairs of antipodal corner points on the n-cube together with any single corner point is a variance maximizing set for the n-cube. The maximum variance for the n-cube with edge length $2s$ is*

$$\frac{(m+1)ns^2}{m}.$$

The definitions and basic results involving compact sets and compact, convex sets can be found in most texts on the theory of convex sets. (See *Rockafellar*, for example.) Discussions of distance measures on compact sets can be found in most texts on advanced calculus or metric topology. (See *Buck* or *Goldberg*, for example.) The results on variance maximization are in *Stuart*.

References

Buck, R.C., (1978), *Advanced Calculus, third ed.*, McGraw-Hill , New York.

Goldberg, R.C. (1976), *Methods of Real Analysis*, John Wiley and Sons, New York.

Rockafellar, R.T. (1970), *Convex Analysis*, Princeton University Press, Princeton, New Jersey.

Stuart, J.L. (2006), “Variance Maximization,” in preparation.