Separation of Clouds by a Plane

This note concerns attempts to draw a (hyper-)plane that partitions a given point-set into two subsets, called "clusters", like two clouds in an otherwise clear sky. Of course, such attempts are futile if the given points do not fall into two separable clusters, so any partitioning algorithm described herein may be merely a solution in search of a problem.

Given are K points $z_1, z_2, z_3, \ldots, z_K$ in a *Euclidean* space of arbitrary but finite dimension. Actually each z_k is a column vector drawn from the origin o to a point in that space, but we shall ignore this distinction and assemble the vectors into a matrix $Z := [z_1, z_2, z_3, \ldots, z_K]$ whose name shall be used for the point-set too. We characterize any (hyper-)plane Π in that space by choosing a nonzero linear functional (row) n^T and scalar constant $ß$ thus: every point x on Π satisfies $n^T x = ß$. Such a plane Π partitions the point-set Z into two clusters distinguished well just when $ß$ falls into a relatively wide gap between the elements of the row $n^T Z$; were its elements sorted in monotonic order, differences between elements that straddle $ß$ would be noticeably bigger than differences between adjacent elements on the same side of $ß$. However, point-sets cannot all be partitioned that well by planes.

Failure Modes:
Any algorithm designed to find a separating plane must fail or at least falter when no relatively wide gap exists among the elements of $n^T Z$ no matter how n^T is chosen. This can happen when clouds overlap a little; perhaps two clouds would be well-separated if some relatively few points were deleted from Z. Even if the clouds are quite distinct they can be situated, like the "c" and "O" in the character "©", where no plane can separate them. Perhaps the given point-set consists of three or more well-separated clusters; then separating planes may be abundant though none is so much better than all others that it deserves to be singled out by the algorithm.

Evidently separation is a matter of degree, and sometimes ill-defined even when conspicuous.

Moreover, separation alone lends itself to misinterpretation. For instance, a border separates the U.S.A. from Mexico, and separates also almost all of their citizens although many of each country's citizens reside on the other's side of the border. A census that counted just adult bodies would over-estimate the numbers of eligible voters in border states. Here is another instance: Suppose the scores achieved by a student on each of several standardized tests are assembled into a column vector z; and suppose the array Z is assembled from the scores of K students coming from two different high-schools, but mixed up so that nobody knows from which school any particular score-vector comes. Suppose too that the score-vectors form two clusters that both abut a separating plane though their centers are well-separated, suggesting strongly that the students come from two distinct populations. If someone jumped to the conclusion that each cluster's students almost all come from the same high-school, he would over-estimate the differences between the two high-schools' students' scores by attributing to one high-school all the scores on its side of the separating plane, thus omitting this school's scores on the other side of the plane while attributing to it some of the other school's scores.

Dividing Points on a Line into Two Clouds:

Let " $n^T x = ß$ " be the equation of a plane $\prod$ that does partition set Z into two clusters, and suppose n^T is known; how can $ß$ be determined? The elements of row $n^T Z$ can be viewed as a set of points on a line and, ideally, $ß$ divides that set into two clusters, one substantially above and the other substantially below $ß$. This much separation is probably too much to demand. Instead, a value $ß$ can be deemed acceptable if it lies in a neighborhood where elements of $n^T Z$ are sparse compared with neighborhoods above and below $ß$ where elements are dense. Here "neighborhood", "sparse" and "dense" are terms too vague to define an algorithm; they merely convey the intention behind the procedure to be described next.

Let $s^T := [s_1, s_2, \ldots, s_K]$ be the row obtained from $n^T Z$ by sorting it into, say, ascending order, so that $s_1 \le s_2 \le \ldots \le s_K$. If $1 \le i < j \le K$, the *Average Gap* between s_i and s_j is defined here to be $(s_j - s_i)/(j-i)$; it is roughly the reciprocal of the average density of elements of $n^T Z$ between s_i and s_j. For some integer k between 1 and K, a plot of $(s_{i+k} - s_i)/k$ against $i = 1, 2, \ldots, K-k$ should show comparatively small Average Gaps where elements of $n^T Z$ are dense in the neighborhood between s_i and s_{i+k}, and large Average Gaps where elements are sparse. If Z deserves to be partitioned into two clusters by $\prod$, the plot should show a pronounced peak between two regions where the Average Gap is comparatively low.

If the increment k is chosen too big, the fluctuations in the plot of Average Gap will be too few and too subdued to locate a gap between clusters. If k is chosen too small, the plot of Average Gap my fluctuate too wildly to locate that separating gap. A plausible initial choice for the increment k is roughly $\sqrt{K}$, and a plausible initial estimate for the separating gap is within the neighborhood where the Average Gap is maximized between neighborhoods where the Average Gap is comparatively low. If k is big enough, the maximizing neighborhood will be determined uniquely; then, plotting the Average Gap again around that neighborhood with a smaller choice of increment k will produce a narrower maximizing neighborhood with a bigger Average Gap.

Thus, plots of Average Gap for a sequence of diminishing incrememts k will produce a nested sequence of narrowing neighborhoods with growing Average Gaps all bigger than those in neighborhoods on both sides. When k gets down to 1 the separating gap in which $ß$ belongs will have been located.

Finding the Direction of a Separating Plane:

In a Euclidean space, where the length of a vector x is $||x|| := \sqrt{(x^T x)}$, the vector n is normal (perpendicular) to the plane $\prod$ whose equation is $n^T x = ß$. The distance from $\prod$ to a point z is $||z - \prod|| := |n^T z - ß|/||n||$ because $\bar{x} := z - n(n^T z - ß)/n^T n$ is the point in $\prod$ closest to z. This is so because $\bar{x}$ satisfies the equation of $\prod$, and every other point x in $\prod$ can easily be shown to satisfy $||z-x||^2 = ||z-\bar{x} + \bar{x} - x||^2 = ||z-\bar{x}||^2 + ||\bar{x}-x||^2 \ge ||z-\bar{x}||^2$.

An alternative form of the equation of Π is $n^T(x–c) = 0$ for any point c in Π, for which $n^Tc = ß$. This alternative form lends itself better to the solution of a *Least-Squares Problem*:

Given $n \neq o$ and a set (matrix) $Z := [z_1, z_2, z_3, \ldots, z_K]$ of points,
choose c to minimize $\sum_j \|z_j–\Pi\|^2$.

This sum of squared distances can be rewritten as

$$\sum_j \|z_j–\Pi\|^2 = \sum_j (n^T(z_j – c))^2/n^Tn = n^T(Z – cu^T)(Z–cu^T)^Tn/n^Tn$$

in which $u^T := [1, 1, \ldots, 1]$ with K elements. A minimizing choice for c turns out to be $ç := Zu/K$, regardless of n. This $ç$ is the average (or *mean*) of the points Z, their center of gravity. It minimizes the sum-of-squares because

$$(Z – cu^T)(Z–cu^T)^T = (Z – çu^T)(Z–çu^T)^T + K(c–ç)(c–ç)^T .$$

In short, among all parallel planes with the same normal n, the plane that minimizes the sum of squared distances from the point-set Z passes through its center of gravity regardless of n.

A plane that partitions the points Z into two separated clusters should be as far as possible from both clusters while passing between them, though it need not pass through Z's center of gravity. This thought motivates the following *MaxiMin Problem*:

Choose the normal n to a plane Π through $ç$ in such a way as to maximize $\sum_j \|z_j–\Pi\|^2$.

In other words, choose c and n to find $\max_n \min_c n^T(Z–cu^T)(Z–cu^T)^Tn/n^Tn$. This sum-of-squares is maximized when the normal $n = ñ$, an eigenvector of $(Z – çu^T)(Z–çu^T)^T$ belonging to its largest eigenvalue; another way to put it is that the maximizing $ñ$ is the singular vector belonging to the biggest singular value of $Z – çu^T$. In its *Singular Value Decomposition* $Z – çu^T = PVQ^T$, where $P^TP = Q^TQ = I$ (an identity matrix) and V is a positive diagonal matrix with the nonzero singular values of $Z – çu^T$ on its diagonal in descending order, $ñ$ is the first column of P.

If we presume that this $ñ$ is the normal to a plane Π that partitions Z into two clusters, the choice of $ß$ to select Π can be achieved as discussed earlier.

Example:
A perfectly partitionable set $Z = [b, b, \ldots, b, d, d, \ldots, d]$ consists of M repetitions of a point b and $K–M$ repetitions of another point d. Now $ç := (Mb + (K–M)d)/K$ is the center of gravity of Z, so $b = ç – (K–M)e$ and $d = ç + Me$ where $e := (d–b)/K$. Then $Z – çu = er^T$ where r^T is a row starting with M repetitions of $M–K$ followed by $K–M$ repetitions of M. The singular vector belonging to the biggest (and only nonzero) singular value of er^T is e, which is the best normal for a plane that separates d from b. On the other hand, any normal not perpendicular to e would work too if not so well, so this example is not a hard test of the procedure described above.

Warning:
Despite its success on one example, the procedure described above has a disquieting property. If an invertible linear operator L maps the set Z to another set LZ it should map a separating plane Π to another separating plane $L\Pi$ of the set LZ. The equation "$n^T x = ß$" satisfied by points x in Π should transform into an equation "$n^T L^{-1} y = ß$" satisfied by points $y = Lx$ on $L\Pi$. But the procedure described above will generally get a normal different from $n^T L^{-1}$ for the plane that separates the clusters of LZ; the new separating plane will not match $L\Pi$ in general, nor will the two clusters of LZ be images of the two clusters of Z.

Perhaps LZ does not deserve to be separated by $L\Pi$. Suppose the points of Z are the four vertices of a nondegenerate tetrahedron with one vertex at the origin o. Any such tetrahedron can be mapped to any other by an invertible linear transformation L. A tetrahedron whose vertex at o is well separated from the three others can be mapped to a tetrahedron whose vertex at o is close to two others but far from the third. This is a case when the clustering of a point-set is changed drastically by a linear map tantamount to a non-orthogonal change of coordinates.

A nonlinear map can change clustering utterly; for instance, the "c" and "O" in the character "©" that cannot be separated by a line are mapped to easily separable sets by a change to polar coordinates centered inside the "c". In general, attempts to separate non-convex clusters by a plane seem unlikely to succeed unless the clusters can be circumscribed by nonintersecting convex surfaces mostly farther apart than their diameters, not like two coins lying flat one on the other.

In other words, the separability of a point-set into two clusters may depend upon the coordinate system chosen for the points. Also important is the precision with which points are located, since the gap between clusters may be misleading if the points' locations are in error by much more than the gap. The procedure described above makes sense only if the choice of coordinate system has this property: The (in)significance of a small perturbation of any point's position is roughly independent of the point and of the perturbation's direction, and therefore determined almost entirely by the perturbation's length in roughly the same way for every point.