Euclidean Cloud

Cory Brunson, from Brigitte Le Roux and Henry Rouanet

November 6, 2015

This vignette (loosely or closely?) follows the treatment in Chapter 3 of Le Roux and Rouanet's *Geometric Data Analysis*, with drastically simplified exposition but employing the same example data.

The central object of GDA is a point cloud in a Euclidean space, which encodes a set of statistical observations arranged in a table, or two-dimensional array (for instance, a contingency table in Correspondence Analysis or an Individuals–Variables table in Principal Components Analysis). By convention, the columns of the table correspond to the coordinate axes of the Euclidean space while the rows correspond to the points.

3.1 Basic Statistics

Rigorously, a *point cloud* in a Euclidean space $\mathcal{U}$ consists of a set J of labels and a mapping $M^J : J \to \mathcal{U}$ that takes each label $j \in J$ to its corresponding point $M^j \in \mathcal{U}$. The points are assigned positive weights ω_j, which may be assumed unitary when not specified. The most common weights will be absolute frequencies n_j, $n = \sum_{j \in J} n_j$, and their associated relative frequencies $f_j = \frac{n_j}{n}$.

As a running example, we load the *Target example* dataset, used throughout the chapter, whose coordinates are provided in Exercise 3.4:

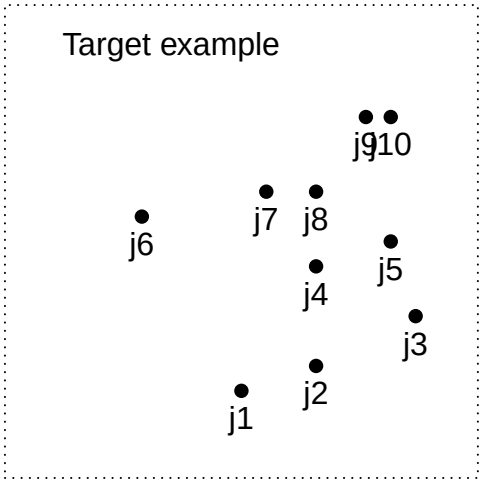
```
data(Target)
print(Target)
```

```
##     x1  x2
## j1   0 -12
## j2   6 -10
## j3  14  -6
## j4   6  -2
## j5  12   0
## j6  -8   2
## j7   2   4
## j8   6   4
## j9  10  10
## j10 12  10
```

```
print(class(Target))
```

```
## [1] "matrix"
```

```
T_freq <- rep(1 / nrow(Target), nrow(Target))
```



The labels consist of j_1 through j_{10}. Because these data occupy two dimensions, i.e. the labels are mapped into $\mathcal{U} = \mathbb{R}^2$, they constitute a *plane cloud*.

3.1.1 Mean Point

The *mean point* of the cloud M^J in $\mathcal{U}$ is meant to be a proxy point location for the entire cloud. Given an arbitrary point $P \in \mathcal{U}$, the mean point is defined as the point arrived to from P via the sum of the weighted vectors $f_j \overrightarrow{PM^j}$ from P to each point M^j—that is, as the (unique!) point $G \in \mathcal{U}$ for which

$$\overrightarrow{PG} = \sum_{j \in J} f_j \overrightarrow{PM^j}$$

doesn't depend on the choice of point $P \in \mathcal{U}$. The mean point satisfies the *Barycentric property* that

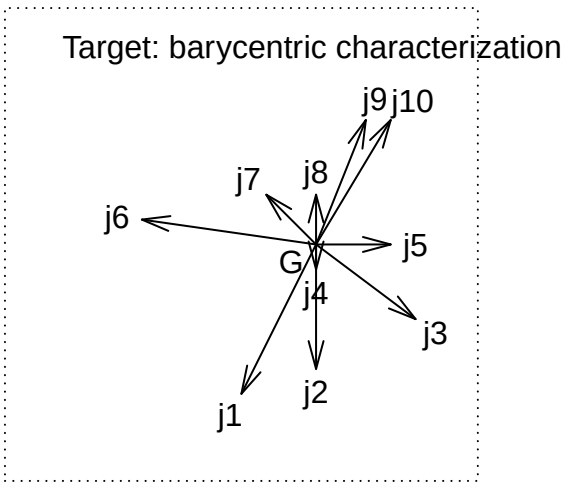
$$\sum_{j \in J} f_j \overrightarrow{GM^j} = \overrightarrow{0},$$

which can be checked by substituting $P = G$ in the definition.[1]

```
G <- barycenter(Target)
print(G)
```

```
##      [,1] [,2]
## [1,]    6    0
```

[1] These and other formulas are given in more general terms in *GDA*, in which the points M^j are weighted by masses ω_j. I may be forced to adopt this generality, but i haven't yet.



3.1.2 Inertia, Sum of Squares, Variance and Contributions

The *inertia* of a cloud M^J with respect to a point P is defined, as in other settings, as the sum-squared distance of each point M^j from P:

$$In^P M^J = \sum_{j \in J} In_j^P = \sum_{j \in J} \omega_j (PM^j)^2.$$

The inertia for $\omega_j = n_j$ is the *sum of squares* and that for $\omega_j = f_j$ is the *mean of squares*. The *variance* of the cloud is its mean of squares from its barycenter, and can be interpreted as the sum of the *absolute contributions* of the individual points in the cloud:

$$VarM^J = \sum_{j \in J} Cta_j = \sum_{j \in J} f_j (GM^j)^2.$$

The *relative contributions of the points are, then, the fractions $Ctr_j = \frac{Cta_j}{VarM^J}$.

```
P <- c(0, 0)
P_in <- inertia(point = P, cloud = Target)
G_in <- inertia(point = G, cloud = Target)
T_var <- inertia(G, cloud = Target,
                           weights = T_freq)
print(c(center_inertia = P_in,
        barycenter_inertia = G_in,
        variance = T_var))
```

```
##     center_inertia barycenter_inertia           variance
##               1280                920                 92
```

```
Ctas <- T_freq * apply(Target, 1, function(m) {
    sum((m - G) ^ 2)
})
Ctrs <- Ctas / T_var
print(cbind(Ctas, Ctrs))
```

```
##     Ctas        Ctrs
```

```
## j1  18.0 0.195652174
## j2  10.0 0.108695652
## j3  10.0 0.108695652
## j4   0.4 0.004347826
## j5   3.6 0.039130435
## j6  20.0 0.217391304
## j7   3.2 0.034782609
## j8   1.6 0.017391304
## j9  11.6 0.126086957
## j10 13.6 0.147826087
```

The **first Huyghen's theorem** states that the inertia of a cloud M^J with respect to a point P can be decomposed into the sum of the variance of the cloud (its inertia with respect to its barycenter G) and the squared distance between G and P:

$$\sum_{j \in J} f_j (PM^j)^2 = (PG)^2 + \sum_{j \in J} f_j (GM^j)^2.$$

(See the illustration below.) A consequence is the *metric characterization of the barycenter*, which has that the barycenter minimizes the mean of squares of the cloud. That is, across $P \in \mathcal{U}$, the quantity $\sum_{j \in J} f_j (PM^j)^2$ is minimized when $P = G$.

```
huyghen.test(point = P, cloud = Target, weights = T_freq)
```

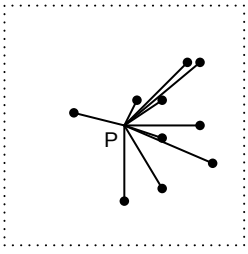
```
## [1] TRUE
```

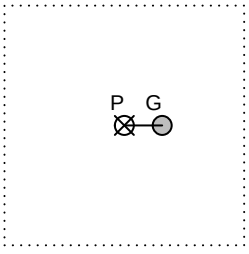
```
## $lhs
## [1] 128
##
## $rhs
## [1] 128
```

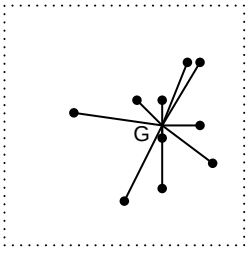
```
huyghen.test(point = G, cloud = Target, weights = T_freq)
```

```
## [1] TRUE
```

```
## $lhs
## [1] 92
##
## $rhs
## [1] 92
```







3.2 Projected Clouds

The *(orthogonal) projection* of a point cloud M^J in a space $\mathcal{U}$ onto a subspace $\mathcal{H}$ is the cloud $H^J = \{H^j\}_{j \in J}$ consisting of its points' projections, and called the *projected cloud*. An important property of the projected cloud is that its barycenter G' is the projection of the barycenter G of M^J.

3.2.1 Variance in a Direction

The variances of the projected clouds of M^J onto two equidimensional parallel subspaces $\mathcal{H}$ and $\mathcal{H}'$ are equal, and are called the *variance in the direction* of $\mathcal{H}$ (or of $\mathcal{H}'$). A special case is the variance in the direction of a line, which is equal to

$$\sum_{j \in J} f_j \frac{\langle \overrightarrow{GM^j} \mid \overrightarrow{\alpha} \rangle^2}{\| \overrightarrow{\alpha} \|^2}$$

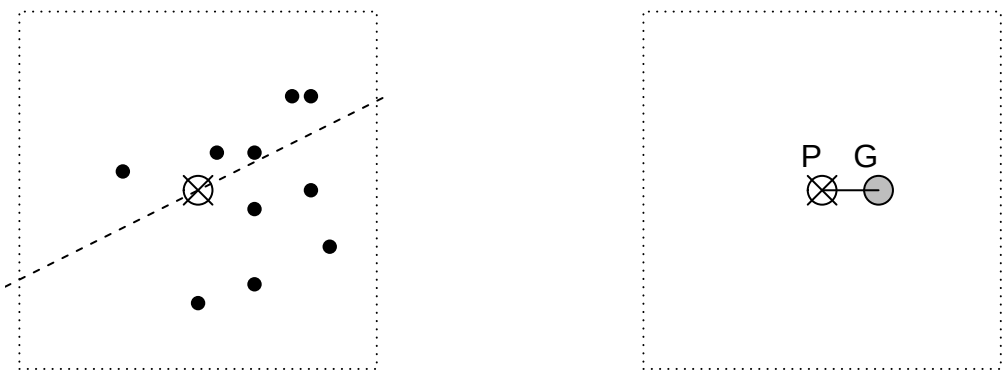
and is called the *variance of axis*.

```
axis <- rbind(c(0, 0), c(2, 1))
axis_var <- direction.variance(cloud = Target, weights = T_freq,
                               subspace = axis)
print(axis_var)
```

```
## [1] 48.8
```

```
print(direction.variance(cloud = Target, weights = T_freq,
                         subspace = affine.decomposition(axis)$linear.subspace,
                         type = "linear"))
```

```
## [1] 48.8
```



3.2.2 Residual Square Mean

The first Huyghen's theorem can be extended from "point" variance to directional variance by way of projection: Given a cloud M^J and a subspace $\mathcal{H}$, the *residual square mean*

$$\sum_{j \in J} f_j (M^j H^j)^2$$

of M^J with respect to $\mathcal{H}$ is the weighted mean of the squares of the "residuals" $M^j H^j$—the distances from the points of M^J to their projections in H^J. The **general Huyghen's theorem** then states that the residual

square mean with respect to $\mathcal{H}$ can be decomposed into the sum of the residual square mean with respect to a same-dimensional parallel space $\mathcal{H}'$ through the barycenter G of M^J and the squared distance from G to $\mathcal{H}$:

$$\sum_{j\in J} f_j(M^jH^j)^2 = \sum_{j\in J} f_j(M^jA^j)^2 + (GG')^2,$$

where G' is the projection of G onto $\mathcal{H}$.

```
huyghen.general.test(cloud = Target, weights = T_freq, subspace = axis)
```

```
## [1] TRUE
```

```
## $lhs
## [1] 50.4
##
## $rhs
## [1] 50.4
```

3.2.3 Fitted and Residual Clouds

When a projected cloud A^J (in a subspace $\mathcal{A}$) is used to model its preimage M^J, it is called a *fitted cloud*, and the vectors $\overrightarrow{A^jM^j}$, whose lengths are the residuals mentioned above, are called *residual deviations*. This leads to the geometric data–model–error formulation

$$\forall j \in J \quad M^j = A^j + \overrightarrow{A^jM^j},$$

where each $\overrightarrow{A^jM^j}$ is orthogonal to $\mathcal{A}$. Consequently, the cloud R^J consisting of points $R^j = G' + \overrightarrow{H^jM^j}$ (where G', as above, is the barycenter of H^J) depends only on the choice of $\mathcal{H}$. R^J is called the *residual cloud*, and it is constructed so that G is its barycenter. This yields the orthogonal decomposition

$$\overrightarrow{GM^j} = \overrightarrow{GA^j} + \overrightarrow{GR^j}.$$

It is straightforward to check that the variance of R^J equals the residual square mean of M^J with respect to $\mathcal{A}$. The distance formula, applied pointwise to the orthogonal decomposition, then yields the variance decomposition

$$VarM^J = VarA^J + VarR^J$$

into the cloud's *total variance*, its *fitted variance*, and its *residual variance*.

```
cloud.decomposition(cloud = Target, subspace = axis)
```

```
## $fitted.cloud
##     [,1] [,2]
## j1  -4.8 -2.4
## j2   0.8  0.4
## j3   8.8  4.4
## j4   4.0  2.0
## j5   9.6  4.8
## j6  -5.6 -2.8
## j7   3.2  1.6
## j8   6.4  3.2
## j9  12.0  6.0
```

```
## j10 13.6  6.8
##
## $residual.cloud
##       x1    x2
## j1  10.8  -9.6
## j2  11.2 -10.4
## j3  11.2 -10.4
## j4   8.0  -4.0
## j5   8.4  -4.8
## j6   3.6   4.8
## j7   4.8   2.4
## j8   5.6   0.8
## j9   4.0   4.0
## j10  4.4   3.2
```

```
cloud.variance.test(cloud = Target, weights = T_freq, subspace = axis)
```

```
## [1] TRUE
```

```
## $lhs
## [1] 92
##
## $rhs
## [1] 92
```

3.2.4 Variables Attached to an Axis

The linear formulation in the next section relies on several coordinatization schemes. Begin with an axis $(G, \overrightarrow{\alpha})$, through the barycenter of M^J in the direction of $\overrightarrow{\alpha}$, and the projected cloud A^J onto this axis. We then define four centered (having mean point zero) variables:

- The *covariant variable* $\alpha^J = (\alpha^j)_{j \in J}$ has *covariant coordinates* $\alpha^j = \langle \overrightarrow{GM^j} \mid \overrightarrow{\alpha} \rangle$ and variance $VarA^J \|\overrightarrow{\alpha}\|^2$. The term refers to the fact that these coordinates scale with $\overrightarrow{\alpha}$.
- The *calibrated variable* $y^J = (y^j)_{j \in J}$ has *calibrated coordinates* $y^j = \alpha^j / \|\overrightarrow{\alpha}\|$ and variance $VarA^J$.
- The *standard variable* $z^J = (z^j)_{j \in J}$ has *standard coordinates* $z^j = y^j / SDy^J$ and variance 1.
- The *axial variable* $t^J = (t^j)_{j \in J}$ has *axial coordinates* $t^j = \langle \overrightarrow{GM^j} \mid \overrightarrow{\alpha} \rangle / \|\overrightarrow{\alpha}\|^2$ and variance $Vart^J = A^J / \|\overrightarrow{\alpha}\|^2$. The t^j are the projections—*GDA* calls them the coordinates—of the A^j onto $(G, \overrightarrow{\alpha})$.