A correlation-based distance

Jean-Luc Falcone and Paul Albuquerque

1 Computer Science Department, University of Geneva, 1211 Geneva 4, Switzerland
2 I³, Ecole d’Ingénieurs de Genève, HES-SO, 1202 Geneva, Switzerland
e-mail: jean-luc.falcone@cui.unige.ch, albuquer@eig.unige.ch

Abstract. In this short technical report, we define on the sample space \( \mathbb{R}^D \) a distance between data points which depends on their correlation. We also derive an expression for the center of mass of a set of points with respect to this distance.

1 Preliminaries

For a sample point \( \mathbf{x} = (x_1, \ldots, x_D) \in \mathbb{R}^D \), we define the average

\[
\bar{x} = \frac{1}{D} \sum_{i=1}^{D} x_i
\]

and the standard deviation

\[
\sigma_x = \sqrt{\frac{1}{D} \sum_{i=1}^{D} (x_i - \bar{x})^2} = \frac{1}{\sqrt{D}} \| \mathbf{x} - \bar{\mathbf{x}} \|
\]

of its components and we set \( \bar{\mathbf{x}} = (\bar{x}, \ldots, \bar{x}) \).

We now restrict our attention to \( \mathbb{R}^D \setminus \text{Diag} \) where

\[
\text{Diag} = \{(x_1, \ldots, x_D) \in \mathbb{R}^D \mid x_1 = \cdots = x_D \}.
\]

To \( \mathbf{x} \in \mathbb{R}^D \setminus \text{Diag} \), we associate the centered and reduced variable

\[
\mathbf{x}^* = \frac{\mathbf{x} - \bar{\mathbf{x}}}{\sigma_x} = \sqrt{D} \frac{\mathbf{x} - \bar{\mathbf{x}}}{\| \mathbf{x} - \bar{\mathbf{x}} \|}
\]

Consequently, \( \bar{x}^* = 0 \) and \( \sigma_{x^*} = 1 \), and we have

\[
\sigma_{x^*}^2 = \frac{1}{D} \sum_{i=1}^{D} (x_i^*)^2 = 1 \Leftrightarrow \sum_{i=1}^{D} (x_i^*)^2 = D
\]
The geometric interpretation of this transform is that \( x^\ast \) lies on the \( D \)-dimensional hypersphere \( S^D(\sqrt{D}) \subset \mathbb{R}^D \) of radius \( \sqrt{D} \) centered at the origin.

The correlation between two sample points, \( x = (x_1, \ldots, x_D) \) and \( y = (y_1, \ldots, y_D) \), in \( \mathbb{R}^D \setminus \text{Diag} \) is given by

\[
corr(x, y) = \frac{\sum_{i=1}^{D} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{D} (x_i - \bar{x})^2 \sum_{i=1}^{D} (y_i - \bar{y})^2}}.
\]

which can also be expressed as

\[
corr(x, y) = \frac{(x - \bar{x}) \cdot (y - \bar{y})}{\|x - \bar{x}\| \|y - \bar{y}\|} = \frac{1}{D} (x^\ast \cdot y^\ast)
\]

where \( x \cdot y \) stands for the scalar product of \( x \) and \( y \).

2 A distance based on correlation

We propose the following correlation-based distance

\[
d(x, y) = \sqrt{1 - (corr(x, y))^2} = \sqrt{1 - \frac{(x^\ast \cdot y^\ast)^2}{D^2}} \tag{1}
\]

for \( x, y \in \mathbb{R}^D \setminus \text{Diag} \). Note that \( 0 \leq d(x, y) \leq 1 \).

The following properties of a metric distance

\[
d(x, x) = 0
\]

\[
d(x, y) = d(y, x)
\]

\[
d(x, z) \leq d(x, y) + d(y, z),
\]

must be verified.

We have

\[
d(x, x) = \sqrt{1 - \frac{(x^\ast \cdot x^\ast)^2}{D^2}} = \sqrt{1 - \frac{D^2}{D^2}} = 0
\]

and, obviously, \( d(x, y) = d(y, x) \).
The main feature of this distance is that strong correlation corresponds to small distance. Indeed,

\[
[\text{corr}(x, y)]^2 = 1 \iff \exists \mu \neq 0, \delta \in \mathbb{R} \text{ s.t. } x_i = \mu y_i + \delta, \forall i
\]

\[
\iff x^* = \pm y^*
\]

\[
\iff d(x, y) = 0.
\]

which also means that the distance \(d\) is degenerate, since \(d(x, y) = 0 \Rightarrow x \neq y\).

The triangle inequality \(d(x, z) \leq d(x, y) + d(y, z)\) requires some explanations. A preliminary remark is that

\[
d(x, y) = \sqrt{1 - \frac{(x^* \cdot y^*)^2}{D^2}} = \sqrt{1 - \frac{[D \cos(\alpha)]^2}{D^2}} = \sqrt{1 - \cos^2(\alpha)}
\]

where 0 \leq \alpha \leq \pi is the angle between \(x^*\) and \(y^*\).

Replacing \(y^*\) by \(-y^*\) and \(z^*\) by \(-z^*\) if necessary, we can assume that the angles \(\alpha\) between \(x^*\) and \(y^*\) and \(\beta\) between \(y^*\) and \(z^*\) belong to \([0, \pi/2]\). Consider the point \(\hat{z}\) obtained by rotating \(z^*\) around the axis defined by \(y^*\), into the plane determined by \(x^*\) and \(y^*\), but opposite to \(x^*\) with respect to \(y^*\). The angle between \(y^*\) and \(\hat{z}\) is still \(\beta\). However, the angle between \(x^*\) and \(\hat{z}\), which equals \(\alpha + \beta\), is greater than the one between \(x^*\) and \(z^*\). Therefore,

\[
d(x, z) \leq \sin(\alpha + \beta) = \sin(\alpha) \cos(\beta) + \sin(\beta) \cos(\alpha)
\]

\[
\leq \sin(\alpha) + \sin(\beta) = d(x, y) + d(y, z)
\]

As previously mentioned, the distance \(d\) is degenerate on \(\mathbb{R}^D \setminus Diag\) or on \(\mathbb{S}^D(\sqrt{D})\). However, we obtain a non-degenerate distance on the projective space \(\mathbb{P}^D\) (i.e. the space of lines through the origin in \(\mathbb{R}^D\)).

3 The center of mass

Onwards, we will assume that all variables are centered and reduced. Hence, we restrict the sample space to the \(D\)-dimensional hypersphere \(\mathbb{S}^D(\sqrt{D}) \subset \mathbb{R}^D\) of radius \(\sqrt{D}\) centered at the origin. We will omit the \(^*\) notation.
We compute the center of mass \( g \in S^D(\sqrt{D}) \) of a set of \( N \) points \( \{x_j\}_{j=1}^N \) on \( S^D(\sqrt{D}) \). By definition, the center of mass minimizes the average square distance to a set of points. We therefore want to minimize the expression

\[
F(g) = \frac{1}{N} \sum_{j=1}^{N} [d(g, x_j)]^2 = 1 - \frac{1}{ND^2} \sum_{j=1}^{N} (g \cdot x_j)^2
\]  

under the constraint

\[
H(g) = 1 - \frac{1}{D} g \cdot g = 0
\]  

that \( g \) lies on \( S^D(\sqrt{D}) \).

We solve this problem using the method of Lagrange multipliers. The gradients of \( F \) and \( H \) must satisfy

\[
\nabla F(g) = \lambda \nabla H(g),
\]

or equivalently

\[
\frac{\partial}{\partial g_k} F(g) = \lambda \frac{\partial}{\partial g_k} H(g) \quad (k = 1, \ldots, D).
\]  

Equation (4) can be rewritten as

\[
\frac{1}{ND} \sum_{j=1}^{N} x_{jk} (x_j \cdot g) = \frac{1}{ND} \sum_{j=1}^{N} \left( x_{jk} \sum_{i=1}^{D} x_{ji} g_i \right)
\]

\[
= \sum_{i=1}^{D} \left( \frac{1}{ND} \sum_{j=1}^{N} x_{jk} x_{ji} \right) g_i = \lambda g_k
\]  

If we define the \( D \times D \) matrix \( M = (m_{ik}) \) by

\[
m_{ik} = \frac{1}{ND} \sum_{j=1}^{N} x_{jk} x_{ji} \quad (i, k = 1, \ldots, D),
\]

then equation (5) becomes

\[
\sum_{i=1}^{D} m_{ik} g_i = \lambda g_k \quad (k = 1, \ldots, D)
\]
or equivalently
\[ Mg = \lambda g \]

Thus, minimizing \( F \) (eq. 2) under the constraint \( H \) (eq. 3) reduces to finding the eigenvectors of \( M \). The eigenvector, correctly normalized in order to satisfy \( H \), for which \( F \) is minimum, yields the center of mass of the set of \( N \) points \( \{ x_j \}_{j=1}^N \) on \( S^D(\sqrt{D}) \). The matrix \( M \) being symmetric, all its eigenvalues are real.