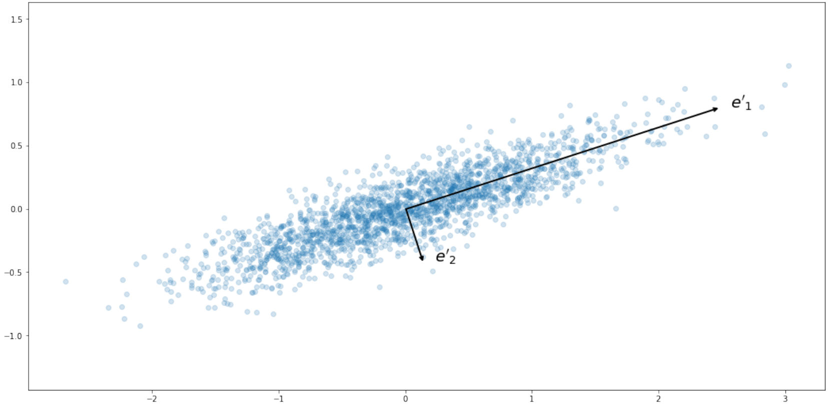
Pricipal Component Analysis

- Principal Component Analysis is a method to reduce a dimensionality of the dataset, while preserving most of the information which it carries. It can be seen then as a compression technique.
- Key idea: some features are correlated, so their presence is redundant.
- Solution: Rotate the feature space in such a way, that the new space basis corresponds to new, uncorrelated features. Reduce the new basis down to N vectors which account for as much of the variability in the dataset as possible. Project dataset onto these vectors, obtaining an approximated representation of N dimensions.

Principal Component Analysis



For a toy example above, vectors e_1' and e_2' were chosen as a new basis. We see, that vector e_1' accounts for much higher variability than vector e_2'. If we were to represent points from the dataset with a single number (1D feature space), the value which is the most distinguishable for dataset points is their projection onto vector e_1'.

Silhouette Metric

Let dataset X be a set divided into n clusters C_j. For every element $x_i \in C_j$ we will define:

$$a(x_i) = \frac{1}{1 - |C_j|} \sum_{x_k \in C_j} dist(x_k, x_i),$$

$$b(x_i) = \min_{l \neq j} \frac{1}{|C_l|} \sum_{x_k \in C_l} dist(x_k, x_i).$$

Then we can define a single element silhouette metric as:

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max(a(x_i), b(x_i))}.$$

And total clustering metric:

$$a(\{C_j\}) = \frac{1}{|X|} \sum_{x_k \in X} s(x_k).$$

Silhouette Metric

$$a(x_i) = \frac{1}{1 - |C_j|} \sum_{x_k \in C_j} dist(x_k, x_i), \quad b(x_i) = \min_{l \neq j} \frac{1}{|C_l|} \sum_{x_k \in C_l} dist(x_k, x_i).$$

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max(a(x_i), b(x_i))}, \quad a(\{C_i\}) = \frac{1}{|X|} \sum_{x_k \in X} s(x_i).$$

Interpretation:

- For given element x_i, the value $a(x_i)$ tells us, how close it is to other elements of the cluster.
- The value $b(x_i)$ indicates, what is the distance to the closest other cluster.
- With good clustering, clusters should be compact (which implies small $a(x_i)$) and distinct (which implies high $b(x_i)$).
- We define the metric $s(x_i)$, which approaches 1 for $a(x_i) \ll b(x_i)$ and -1 for $a(x_i) \gg b(x_i)$.

Further reading

Theory:

- http://people.duke.edu/ hpgavin/SystemID/References/Gillies-PCA-notes.pdf

Example applications:

- https://blog.insito.me/why-pca-and-genetics-are-a-match-made-in-heaven-6042ea027cf0
- https://www.aanda.org/articles/aa/abs/2013/05/aa20961-12/aa20961-12.html
- https://pdfs.semanticscholar.org/30f1/ceb3139129f0a96b0638e999113f46b32e7d.pdf