Equivalency

1 Diagonalizability

Let $\mathbf{X}$ be a matrix whose rows are variables and columns are observations that are centered[1] about the mean. Then the covariance of the $\mathbf{X}$ is given by

$$\mathbf{\Sigma} = cov(\mathbf{X}) = \frac{1}{n-1}\mathbf{X}\mathbf{X}^{\mathbf{T}} \tag{1}$$

Assume we want to have a mapping $\mathbf{M}$ so that $\mathbf{Y} = \mathbf{M}\mathbf{X}$ has covariance that is diagonal. Then

- Assume there is no observation with all zero values
- Assume there is no two identical observations, i.e.columns are linearly independent.
- (Rare) things like that which will make covariance matrix singular.

Then since covariance matrix is symmetric, with positive entries, etc., it is diagonalizable, i.e. it can be written as

$$\mathbf{\Sigma} = \mathbf{U}^{-\mathbf{1}}\mathbf{D}\mathbf{U} = \mathbf{U}^{\mathbf{T}}\mathbf{D}\mathbf{U} \tag{2}$$

or equivalently

$$\mathbf{U}^{\mathbf{T}}\mathbf{\Sigma}\mathbf{U} = \mathbf{D} \tag{3}$$

where $\mathbf{D}$ is diagonal and, in this case due to properties of covariance matrix, the matrix $\mathbf{U}$ is orthogonal, i.e. $\mathbf{U}^{-\mathbf{1}} = \mathbf{U}^{\mathbf{T}}$, i.e. $\mathbf{U}\mathbf{U}^{\mathbf{T}} = \mathbf{I}$

The columns of $\mathbf{U}^{-\mathbf{1}}$ are eigenvectors of covariance matrix, entries of $\mathbf{D}$ are eigenvalues of covariance matrix, which are also variances of different variables in $\mathbf{X}$.

Assume we want to rotate the data, or represent them in a coordinate system where it is represented with variables that are orthogonal, i.e. the collinearity is killed. This is what PCA does.

Now lets pretend we do not know that. Lets say we want to transform $\mathbf{X}$ into $\mathbf{Y}$ by a mapping $\mathbf{M}^{\mathbf{T}}$, i.e. $\mathbf{Y} = \mathbf{M}^{\mathbf{T}}\mathbf{X}$, so that covariance of $\mathbf{Y}$ is diagonal matrix, $\hat{\mathbf{D}}$.

[1] This is why the covariance matrix of PCA data I made in your office was not diagonal, I did not centerize it!

So, we want $cov(\mathbf{Y}) = \frac{1}{n-1}\mathbf{Y}\mathbf{Y}^{\mathbf{T}}$, Lets take a look:

$$\begin{aligned}\hat{\mathbf{D}} = cov(\mathbf{Y}) &= \frac{1}{n-1}\mathbf{Y}\mathbf{Y}^{\mathbf{T}} \\ &= \frac{1}{n-1}\mathbf{M}^{\mathbf{T}}\mathbf{X}(\mathbf{M}^{\mathbf{T}}\mathbf{X})^{\mathbf{T}} \\ &= \frac{1}{n-1}\mathbf{M}^{\mathbf{T}}\mathbf{X}\mathbf{X}^{\mathbf{T}}\mathbf{M} \\ &= \mathbf{M}^{\mathbf{T}}\mathbf{\Sigma}\mathbf{M}\end{aligned} \tag{4}$$

So, we arrived at $\hat{\mathbf{D}} = \mathbf{M}^{\mathbf{T}}\mathbf{\Sigma}\mathbf{M}$. Hence, if you choose $\mathbf{M}$ to be the same as $\mathbf{U}$, then covariance of $\mathbf{Y}$ would be diagonal, and you have $\hat{\mathbf{D}} = \mathbf{D}$.

Since, the the decomposition given by Eq. (2) is eigen-decomposition of $\mathbf{\Sigma}$, we can see this is what PCA does.

2 Equivalency

Definition 2.1. Suppose the random vectors of $\mathbf{v}$ and $\mathbf{w}$ be drawn from a distribution whose associated covariance matrix is given by $\mathbf{\Sigma}$. Then define the Malanoblis distance as follows:

$$d = (\mathbf{v} - \mathbf{w})^{\mathbf{T}}\mathbf{\Sigma}^{-\mathbf{1}}(\mathbf{v} - \mathbf{w}) \tag{5}$$

Lets take a look:

$$\begin{aligned}\mathbf{d} &= (\mathbf{v} - \mathbf{w})^{\mathbf{T}}\mathbf{\Sigma}^{-\mathbf{1}}(\mathbf{v} - \mathbf{w}) \\ &= (\mathbf{v} - \mathbf{w})^{\mathbf{T}}(\mathbf{U}^{\mathbf{T}}\mathbf{D}\mathbf{U})^{-\mathbf{1}}(\mathbf{v} - \mathbf{w}) \\ &= (\mathbf{v} - \mathbf{w})^{\mathbf{T}}(\mathbf{U}^{-\mathbf{1}}\mathbf{D}^{-\mathbf{1}}\mathbf{U}^{-\mathbf{T}})(\mathbf{v} - \mathbf{w}) \\ &= (\mathbf{v} - \mathbf{w})^{\mathbf{T}}(\mathbf{U}^{\mathbf{T}}\mathbf{D}^{-\mathbf{1}}\mathbf{U})(\mathbf{v} - \mathbf{w}) \\ &= (\mathbf{U}(\mathbf{v} - \mathbf{w}))^{\mathbf{T}}\mathbf{D}^{-\mathbf{1}}(\mathbf{U}(\mathbf{v} - \mathbf{w}))\end{aligned} \tag{6}$$

Notice:

- The diagonal entries of $\mathbf{D}$ eigenvalues of covariance matrix, which are variance of variables in $\mathbf{X}$. So, if the data in $\mathbf{X}$ was scaled by their variances, then this distance was equivalent to Euclidean distance.
- Lets look at the last term in above equation:

$$\mathbf{U}(\mathbf{v} - \mathbf{w}) = \mathbf{U}\mathbf{v} - \mathbf{U}\mathbf{w} \tag{7}$$

  Each of these terms are mappings of $\mathbf{v}$ and $\mathbf{w}$ into the PCA space of the data $\mathbf{X}$.

Suppose $\mathbf{x}$ is a vector and we wish to represent it, in the column space of a matrix $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \ldots \mathbf{A}_N]$, where each $\mathbf{A}_i$ is a column of $\mathbf{A}$. So, we are looking for constants $y_1, y_2, \ldots, y_N$ so that

$$\mathbf{x} = y_1\mathbf{A}_1 + y_2\mathbf{A}_2 \cdots + \mathbf{A}_N y_N = \mathbf{A}\mathbf{y}$$

Hence, $\mathbf{y} = \mathbf{A^{-1}X}$. So, $\mathbf{y}$ is the mapping of $\mathbf{x}$ into column space of $\mathbf{A}$. Just like $\mathbf{Uv}$ which is mapping of $\mathbf{v}$ into column space of $\mathbf{U^{-1}}$ whose columns are eigenvectors of covariance, i.e. PCA.

P.S. I had to choose where to put the exponent {-1} to indicate inverse of the matrices, whether to the left or right of $\mathbf{\Sigma}$. Either way would have made some parts easier, but some other parts harder to follow.

Now, the question is, when you mentioned M. Distance takes care of scales and collinearity, where, in what context you learned that. Did the context refer to this definition of distance as M. distance, or they were talking about the distance between a point and distribution?