Unleashing the Power of Principal Component Analysis: Diagnosis of OA degenerative joint disease

Carsi González, Ana

Statistik und Simulation mit R (R2)
ana.carsi.gonzalez@mathe.uni-giessen.de

June 2023

Abstract

This report provides a theoretical and practical guide to the principal component analysis (PCA) method and its applications. Mathematical concepts behind PCA, its assumptions and limitations are discussed, as an introduction to its application in the classification of patients with advanced or early Osteoarthritis. Furthermore, recommendations for when to use PCA are provided, as an aim to guide researchers and data analysts looking to maximize insights and efficiency in their projects.

Keywords: dimensionality reduction, Karhunen-Loève transformation, Hotelling transformation, empirical orthogonal functions, Singular Value Decomposition, Principal Component Analysis, Synovitis, Osteoarthritis, ...

I. Introduction

In the present era of unprecedented data growth and demand for expedited analytical methodologies, PCA surged in popularity as a powerful tool for feature extraction. Particularly, it has been instrumental in identifying underlying patterns and reducing high-dimensional data sets into a smaller number of variables, easing the understanding and interpretation of the data. At the same time, other techniques of machine learning and data mining have been developed, whose use depends on the specific data set and of course, the research question being addressed.

For instance, Support Vector Machines (SVM) have been used as powerful classification algorithms that separate data points by finding a hyperplane that maximizes the margin between classes. This hyperplane is determined by finding the support vectors, which are data points located closest to the decision boundary. However, these algorithms have shown a higher computational cost than PCA, especially with larger data sets. Meanwhile, Independent Component Analysis (ICA) separates multivariate data into independent non-Gaussian subcomponents and is particularly useful in signal processing applications. Nevertheless, PCA has achieved a lead in demand and popularity over SVM and ICA due to its ability to handle larger complex data sets. It is also useful for data visualization and pre-processing, before applying other data mining techniques.

This report is divided in two sections: in the first one, a comprehensive overview of the mathematical concepts behind PCA and the advantages and disadvantages of this approach are provided. Secondly, an application of PCA is presented. This consists of a study to support current research on diagnosis of Osteoarthritis, for which PCA was applied to determine whether the inflammation of the Synovium joint tissue can help classify Osteoarthritis patients. This approach may be in many cases more efficient than the diagnosis of OA via Nuclear magnetic resonance (NMR). Finally, conclusions from both parts are stated in the last section.

II. Theoretical Background

PCA is applied to reduce the dimensionality of multivariate data while retaining most of its variability. It does so by transforming a large set of variables into a smaller group of linearly uncorrelated ones called **principal components**. Clearly the maximum number of principal components is the number of variables of the set.

Therefore, starting with p-dimensional

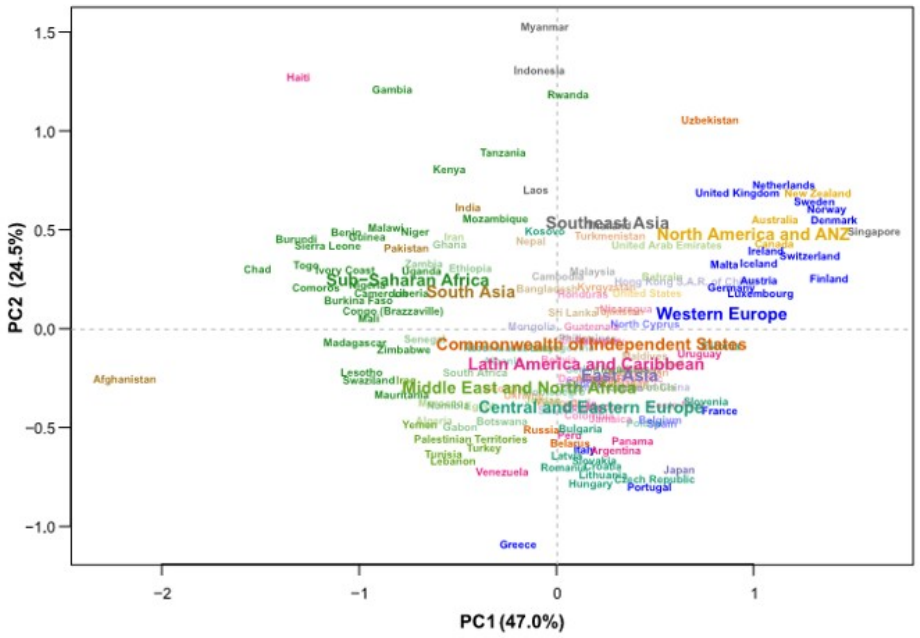

Figure 1: *Plot of multivariate data for 149 countries using the first two principal components as coordinate axes. The 82 countries that contribute more than average to the two-dimensional solution are shown in darker font and are generally further from the centre. The mean positions of the ten regions are added (each mean is at the centre of its label).*

vectors, one seeks to summarize them by projecting down into a q-dimensional subspace. The result will be the projection of the original vectors onto q directions, the principal components, which span the subspace [6].

As a first objective, PCA seeks to obtain the **standardized linear combination** (SLC) of the original variables which has the maximal variance, in order to encase the widest range of information. This may seem feasible, as a large variance eases the identification of disparities among the data points, therefore helping with their classification. Maximizing the variance turns out to be the same as looking for the projection with the smallest average distance (mean-squared) between the original vectors and their projections onto the principal components. This might seem the most logical approach at first, since for a representation to be successful, it must select an appropriate viewpoint, that is to say, by distorting the distances between individuals as little as possible.

An introductory example comes from the World Wealth Report conducted in 2021 as part of the Gallup World Poll in 149 countries, for which five indicators are considered: social support (abbreviated as Social), healthy life expectancy (Life), freedom to make one's own life choices (Choices), generosity of the general population (Generosity) and perception of internal and external corruption levels (Corruption). PCA determines the relationships between these five indicators, seeking a linear combination of the ones that has maximal variance. After this, a reorientation through the new two principal components is applied, resulting in the point cloud (**Figure 1**). In this case, the first PC_1 is the following linear combination of the five variables (factors), $PC_1 = 0.538$ Social $+ 0.563$ Life $+ 0.498$ Choices $- 0.004$ Generosity $- 0.38$ Corruption, and PC_2 is chosen to be linearly independent with the first one. Further details for this example are given in section B.b).

A. Mathematical framework

In order to extract the SLC with the maximal variance, PCA computes the eigenvalues and eigenvectors of the covariance matrix (or the correlation matrix in case that the variables are standardised). For this objective, Singular Value Decomposition (SVD) is applied. Then are the eigenvectors used to construct the principal components according to the greatest eigenvalues, which are sorted in descending order (and will all be non negative since the covariance matrix is positive semi-definite). Hereunder an introduction to the theoretical explanation of the method distinguishing between the population and the empirical case is provided. The exposition of these concepts refers to the source book from Mardia [1].

A.1 Population principal components

Definition 1 *Let* $\mathbf{X}$ *be a random vector with mean* μ *and covariance matrix* Σ*, then the* ***principal component transformation*** *is the transformation*

$$\mathbf{X} \longrightarrow \mathbf{Y} = \Gamma'(\mathbf{X} - \mu)$$

where Γ is orthogonal and

$$\Gamma'\Sigma\Gamma = \Delta$$

is diagonal, where $\lambda_1 \geq \lambda_1 \geq \cdots \geq \lambda_p \geq 0$. This representation of Σ follows from theorem 7 (Appendix). The i-th principal component may be defined as the i-th component of the vector $\mathbf{Y}$, namely as

$$\mathbf{y}_i = \gamma'_{(i)}(\mathbf{X} - \mu)$$

where $\gamma_{(i)}$ is the i-th column of Γ and may be called the i-th vector of **principal component loadings**. Later we may refer to this matrix as the **rotation matrix of the data set**.

In the following we describe further features of the theoretical approach of PCA that ease the computations:

Theorem 1 *Let* $\mathbf{X} \sim (\mu, \Sigma)$ *and* $\mathbf{Y}$ *as in Definition 1. Then*

(a) $\mathbf{E}(\mathbf{y_i}) = 0$*;*

(b) $\mathbf{V}(\mathbf{y_i}) = \lambda_i$*;*

(c) $\mathbf{Cov}(\mathbf{y_i}, \mathbf{y_j}) = 0, i \neq j$*;*

(d) $\mathbf{V}(\mathbf{y_1}) \geq \cdots \geq \mathbf{V}(\mathbf{y_p}) \geq 0$*.*

Moreover,

$$\sum_{i=1}^{p} \mathbf{V}(\mathbf{y_i}) = tr(\Sigma)$$

and

$$\prod_{i=1}^{p} \mathbf{V}(\mathbf{y_i}) = det(\Sigma).$$

Proof (a)-(d) are trivial given Definition 1. The two last claims follow from (b) and the fact that the trace of Σ is the sum of eigenvalues and the determinant is the product of them.

Furthermore, it is also important to note that, since one should aim to maximise the variability explained by the principal components, there is a bound on the maximum value that may be attained.

Theorem 2 *No SLC of* $\mathbf{X}$ *has a variance larger than* λ_1*, the variance of the first principal component.*

Proof Let the SLC be $\mathbf{a'X}$, where $\mathbf{a'a} = 1$. Since the eigenvectors of Σ constitute a basis for $\mathbb{R}^p$, we may write

$$\mathbf{a} = c_1\gamma_{(1)} + \cdots + c_p\gamma_{(p)} \tag{1}$$

Let $\alpha = \mathbf{a'X}$, then with the spectral decomposition follows:

$$V(\alpha) = \mathbf{a}'\Sigma\mathbf{a} = \mathbf{a}' \left(\sum_{i=1}^{p} \lambda_i \gamma_{(i)} \gamma'_{(i)} \right) \mathbf{a} \tag{2}$$

and plugging in (1) in (2), it follows

$$V(\alpha) = \sum_{i=1}^{p} \lambda_i c_i^2 \tag{3}$$

because of $\gamma_{(i),j} \cdot \gamma_{(i),k}$ being the Kronecker delta. Since $\mathbf{a'a} = 1$, $\sum c_i^2 = 1$, so the maximum of (3) is λ_1. Therefore the maximum is obtained when $c_1 = 1$ and the other constants are zero, i.e. the $V(\alpha)$ is maximized when $\mathbf{a} = \gamma_{(1)}$.

We may as well note that the maximum variance that is not explained by the $k < n$ principal components is obtained by creating the $k+1$ principal component in the direction of the eigenvector associated with the λ_{k+1} eigenvalue.

Theorem 3 *If $\alpha = \mathbf{u}'\mathbf{X}$ is a SLC of $\mathbf{X}$ which is uncorrelated with the first k principal components of $\mathbf{X}$, then the variance of α is maximized when it is the k+1-th principal component of $\boldsymbol{X}$.*

A.2 Sample principal components

After introducing the population-based framework of PCA, we present the sample-based counterpart from which the conclusions of the analysis may be obtained. Let $\mathbf{X}$ be a sample data matrix n x p, the *sample covariance matrix of* $\mathbf{X}$ may be written as

$$S = \frac{1}{n}(\mathbf{X}'\mathbf{X} - \frac{1}{n}\mathbf{X}'\mathbb{1}_p\mathbb{1}_n\mathbf{X}).$$

Defining

$$H = I_n - \frac{1}{n}(\mathbb{1}_p\mathbb{1}_n)'$$

then we can write

$$S = \frac{1}{n}\mathbf{X}'H\mathbf{X} = GLG'$$

which is a non negative definite matrix, with $G = (g_{(1)}, \ldots, g_{(p)})$ and $g_{(i)}$ the standardized eigenvector to the i-th eigenvalue of S. The decomposition is again obtained from the spectral theorem (7), so G is an orthogonal matrix (named as the **matrix of loadings**) and $L = diag(\lambda_1, ..., \lambda_n)$ with $\lambda_1 \geq ... \geq \lambda_n$.

Following again Definition 1 we obtain

$$\mathbf{Y} = (\mathbf{X} - \mathbb{1}_\mathbf{n}\bar{\mathbf{X}})G$$

the matrix of principal components, for which the columns represent uncorrelated linear combinations of the variables. The i-column of the vector $\mathbf{Y}$ is the i-th principal component:

$$\mathbf{y}_i = g_i'(\mathbf{X} - \mathbb{1}_n\bar{\mathbf{X}})$$

where g_i is the i-th column of G, and where $\mathbf{y}_i$,$\mathbf{y}_j$ are uncorrelated. After some calculations we obtain that the covariance matrix of $\mathbf{Y}$ is $S_Y = L$, that is, the columns of $\mathbf{Y}$ are uncorrelated and the variance of $\mathbf{y}_i$ is l_i. The r-the element y_{ri} of $\mathbf{y_i}$ represents the *score* of the i-the principal component of the r-th individual.

B. Further properties of principal components

Although the most significant properties of principal components have already been given, several other useful characteristics are presented as a heading for the following section.

(a) *The principal components are not scale-invariant.* One of the main disadvantages of PCA is that it is sensitive to the variables' scales. If the i-th variable is divided by a factor d_i then the covariance matrix of the new variables is $\mathbf{DSD}$, where $\mathbf{D} = diag(d_i^{-1})$. Then however, if $\mathbf{x}$ is an eigenvector of S, $\mathbf{D^{-1}x}$ is *not* an eigenvector of $\mathbf{DSD}$.

An illustrative example for this case is the following. Assume a data set with three variables, say, weight in pounds, height in feet and age in years. Let's assume we seek to obtain a principal component in ounces, inches and decades. Then two procedures may be considered:

(a) multiplying the component by 16, 12 and 1/10 and carry out the analysis.

(b) carry out the analysis and then multiply the elements of the component by 16, 12 and 1/10.

Then the principal components obtained after each procedure are not guaranteed to be equal. In fact, the first component of the original variables has different coefficients and different eigenvectors (not linearly dependant on the ones resulting from the other procedure) and the proportion of explained variance (inertia) also varies. In the following a counterexample for $p = 2$ is provided.

Counterexample Further details on this example can be obtained from section 8.2.3 of the bibliographic material [1]. Consider the population covariance matrix:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

where $\rho \geq 0$. The larger eigenvalue is

$$\lambda_1 = \frac{1}{2}(\sigma_1^2 + \sigma_2^2) + \frac{1}{2}\mathbf{\Delta},$$

where $\mathbf{\Delta} = ((\sigma_1^2 - \sigma_2^2)^2 + 4\ \sigma_1^2\ \sigma_2^2\ \rho^2)^{1/2}$. In the bibliography material it has been shown that the eigenvector to this eigenvalue is proportional to

$$(u_1, u_2) = (\sigma_1^2 - \sigma_2^2 + \Delta, 2\rho\sigma_1\sigma_2). \quad (4)$$

(see [1]). When $\sigma_1/\sigma_2 = 1$, the ratio u_2/u_1 derived from (4) is equal to 1. If $\sigma_1 = \sigma_2$ then the first variable is multiplied by a factor k and for scale-invariance we may seek for $u_2/u_1 = k$. However, substituting σ_1 with $k\sigma_1$ in (4) does not output the unity vector.

(b) *The sum of the first q eigenvalues divided by the sum of all the eigenvalues represents the* ***proportion of total variation*** *explained by the first q principal components, α_q:*

$$\frac{\lambda_1 + \cdots + \lambda_q}{\lambda_1 + \ldots \lambda_p} = \alpha_q.$$

The proportion of total variation defined here gives a quantitative measure of the amount of information retained by reducing from p to q dimensions.

Note to the example: Regarding the example in Figure 1., if each of the five variables is regressed on PC_1, their explained variances, being identical to the squared correlations with PC_1, are 0.680, 0.744,0.583, 0.000, and 0.341. Hence, the second variable (Life) makes the largest contribution to PC_1, whereas the fourth variable (Generosity) has almost none. The sum of these explained variances divided by the total is 0.470, so that the PC_1 has "explained" 47.0% of the total variance. Since 53.0% of the total variance has been left, a second linear combination of the variables is sought to explain as much of this residual variance as possible.

(c) *If the covariance matrix of* $\mathbf{X}$ *has rank* $q \leq p$*, then the total variation of* $\mathbf{X}$ *can be explained entirely by the first q principal components.* This follows from $\mathbf{\Sigma}$ having rank q precisely when the last $(p - q)$ eigenvalues of $\mathbf{\Sigma}$ are zero.

(d) *The vector subspace spanned by the first q principal components* $(1 \leq q < p)$ *has smaller mean square deviation from the population variables than any other k-dimensional subspace.* For the case $k = 1$, this claim is a reformulation of Theorem 2.

C. Correlation structure

We now examine the correlation between $\mathbf{X}$ and its vector of principal components $\mathbf{Y}$, assuming for simplicity that $\mathbf{X}$ and therefore $\mathbf{Y}$ have mean zero. Then

$$Cov(\mathbf{X}, \mathbf{Y}) = \mathbb{E}(\mathbf{X}\mathbf{Y}') =$$

$$= \mathbb{E}((\mathbf{X} - \mathbb{E}(\mathbf{X}))(\mathbf{Y} - \mathbb{E}(\mathbf{Y}))')$$

and since $\mathbf{Y} = \mathbf{\Gamma}'(\mathbf{X} - \mu)$, we have:

$$Cov(\mathbf{X}, \mathbf{Y}) = \mathbb{E}((\mathbf{X} - \mu)(\mathbf{X} - \mu)'\Gamma) = \mathbf{\Sigma\Gamma} = \mathbf{\Gamma\Delta\Gamma'\Gamma} = \mathbf{\Gamma\Delta}$$

which is due to the spectral decomposition of $\mathbf{\Sigma}$, $\mathbf{\Gamma'\Sigma\Gamma} = \mathbf{\Delta}$. x_i and y_j have variances σ_{ii} and λ_j respectively, and if their correlation is ρ_{ij}, then

$$\rho_{ij} = \frac{\gamma_{ij}\lambda_j}{(\sigma_{ii}\lambda_j)^{1/2}} = \gamma_{ij}\sqrt{\frac{\lambda_j}{\sigma_{ii}}} \quad (5)$$

where γ_{ij} is the i-th element of the j-th eigenvector of $\mathbf{\Sigma}$.

Therefore the covariance between x_i and the j-th principal component y_j is $\gamma_{ij}\lambda_j$. Adding up to the previous section D.b) we name the **proportion of variation of** x_i **explained by** y_j as ρ_{ij}^2. Since the elements of $\mathbf{Y}$ are uncorrelated, any set J of components explains the proportion

$$\rho_{iJ}^2 = \sum_{j \in J} \rho_{ij}^2 = \frac{1}{\sigma_{ii}} \sum_{j \in J} \lambda_j \gamma_{ij}^2. \quad (6)$$

which results from (5) when $\mathbf{\Sigma}$ is the correlation matrix (the variables are divided by their standard deviations), that is, $\sigma_{ii} = 1$ and $\rho_{ij} = \gamma_{ij}\sqrt{\lambda_j}$.

It is important to note that, since $Var(x_i) = \sigma_{ii} = \Sigma_{(i,i)} = (\mathbf{\Gamma\Delta\Gamma'})_{(i,i)}$, when

G includes all the components then is the ratio (6) equal to one. The **total variation** accounted by the components in G is the sum of all p elements of the **proportion of variation in each variable** explained by the components in G, that is,

$$\sum_{j\in G}\lambda_j = \sum_{i=1}^{p}\sigma_{ii}\rho_{iG}^2$$

since

$$\sum_{i=1}^{p}\sigma_{ii}\rho_{iG}^2 = \sum_{j\in G}\lambda_j\sum_{i=1}^{p}\gamma_{ij}^2$$

where

$$\sum_{i=1}^{p}\gamma_{ij}^2 = 1$$

because Γ is orthogonal.

D. Maximum likelihood estimation for normal data

For several cases may the sample distribution of eigenvalues and eigenvectors of a covariance matrix $\mathbf{S}$ be complicated. For this reason, some useful properties of the sample principal components for normal data may be considered.

Theorem 4 *For normal data when the eigenvalues of $\mathbf{\Sigma}$ are distinct, the sample principal components and eigenvalues are the **maximum likelihood estimators** of the corresponding population parameters.*

Proof The eigenvalues of $\mathbf{\Sigma}$ (and therefore its principal components) are related to $\mathbf{\Sigma}$ through a bijective one-to-one function, except when $\mathbf{\Sigma}$ has eigenvalues with multiplicity greater than one. The proof follows then as a result of the invariance of MLEs (see Appendix).

Note When the eigenvalues of $\mathbf{\Sigma}$ are not distinct the above theorem does not hold. In such cases, a certain arbitrariness in defining the eigenvectors of $\mathbf{\Sigma}$ is allowed and the eigenvalues of $\mathbf{\Sigma}$ may have a function to $\mathbf{\Sigma}$ that is different from the one that the eigenvalues of $\mathbf{S}$ have to $\mathbf{S}$. This means that for this case, principal components are in general *not* MLEs of their population cases.

The following theorem offers a solution for this case through a relation between population and sample eigenvalues.

Theorem 5 *For normal data with $k > 1$ not distinct eigenvalues λ of $\mathbf{\Sigma}$, then*

(a) the MLE of λ is $\bar{l}$, the arithmetic mean of the corresponding sample eigenvalues, and

(b) the sampe eigenvectors corresponding to the repeated eigenvalue λ are MLEs, although they are not the unique ones.

Note A further explanation of the estimation of the distribution of eigenvalues for Gaussian symmetric matrices may be found in [5].

E. Asymptotic distributions for normal data

The estimation of the distribution of eigenvalues may be used for small sample cases, however for large samples the following asymptotic result, based on the central limit theorem (see Appendix), provides useful distributions for the eigenvalues and eigenvectors of the sample covariance matrix $\mathbf{S}$.

Theorem 6 *(Anderson, 1963)*
Let $\mathbf{\Sigma}$ be a positive-definite matrix with distinct eigenvalues. Let $\mathbb{W}_p$ be the p-dimensional Wishart distribution and assume $\mathbf{M} \sim \mathbb{W}_p(\mathbf{\Sigma}, m)$ We set $\mathbf{U} = m^{-1}\mathbf{M}$. Consider spectral decompositions $\mathbf{\Sigma} = \mathbf{\Gamma}'\mathbf{\Delta}\mathbf{\Gamma}$ and $\mathbf{U} = \mathbf{GLG}'$ and let λ and $\mathbf{l}$ be the vectors of the diagonal elements in $\mathbf{\Delta}$ and $\mathbf{L}$. Then the following asymptotic distributions hold as $m \longrightarrow \infty$:

(a) $\mathbf{l} \sim N_p(\lambda, 2\mathbf{\Delta}^2/m)$; that is the eigenvalues of $\mathbf{U}$ are asymptotically normal, unbiased, and independent, with $l_i$having variance $2\lambda_i^2/m$, and similarly:

(b) $\mathbf{g_{(i)}} \sim N_p(\gamma_i, \mathbf{V_i}/m)$, where

$$\mathbf{V_i} = \lambda_i\sum_{j\neq i}\frac{\lambda_j}{(\lambda_j-\lambda_i)^2}\gamma_{(j)}\gamma_{(j)}';$$

that is, the eigenvectors of $\mathbf{U}$ are asymptotically normal and unbiased too, and have the stated asymptotic covariance matrix $\mathbf{V}_i/m$.

(c) The elements of $\mathbf{l}$ are asymptotically independent of the elements of $\mathbf{G}$.

Further details on this theorem are to be found in the bibliographic material [1].

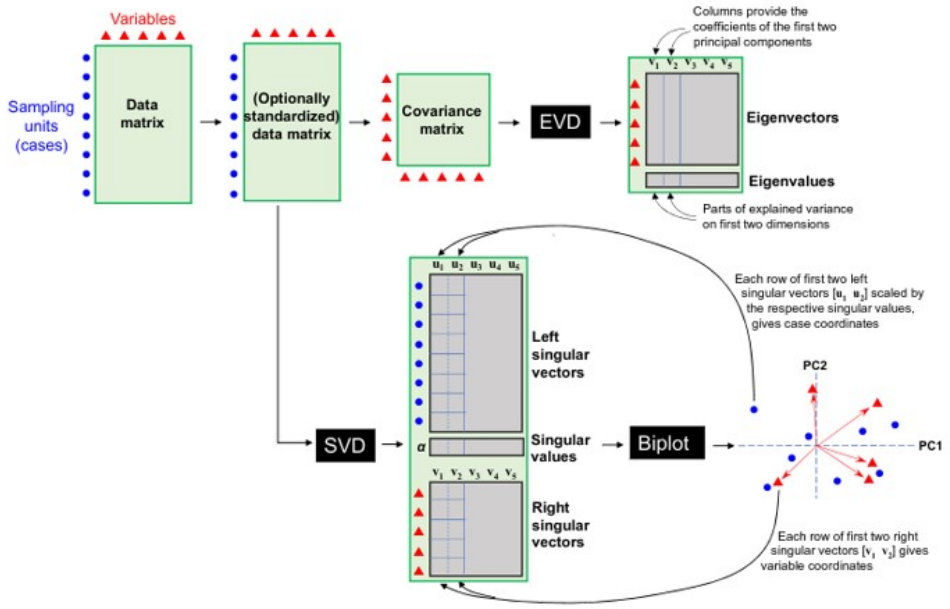

Figure 2: *Schematic view of the PCA workflow. The definition of the principal components (PCs) is obtained using the eigenvalue decomposition of the covariance matrix of the variables (section A). Standardization is optional, but centering is mandatory. For the lower pathway to be exactly equivalent to the upper one, the (optionally standardized) data matrix should be divided by* $\sqrt{n}$.

III. Application on Diagnosis of Osteoarthritis Progression

Osteoarthritis (OA) is the most prevalent arthritic disease and a leading cause of disability, affecting 32.5 million US adults and more than 40 million european adults (see Figure 3).

Research into the pathophysiology of osteoarthritis (OA) has focused on examining the cartilage and peri-articular bone on the knee, however, recent evidence supports a newer perspective: the clinical syndrome of "OA" affects as well the integrity of multiple joint tissues, in particular, the synovium (SM)[7].

The synovial lining consists of a thin layer of cells (macrophages) and vascularized connective tissue that form a complex structure, source of synovial fluid (SF) components. These ones are essential for normal cartilage and joint function and in fact, recent evidence [8] suggests that the inflammatory process of the synovium (synovitis) involves inflammatory mediators that represent potential targets for therapeutic interventions designed to reduce both symptoms and structural joint damage in OA patients. This report focuses on applying PCA to study the impact of synovial inflammation in OA, hoping to determine whether the substances that appear on it can be used to determine the progress of OA in adults. This approach may be in many cases more efficient and cheap than diagnosing OA with image detection, i.e. Nuclear Magnetic Resonance (NMR).

A. Method of the approach

For this purpose, a similar workflow as in Figure 2 has been used. For two records with multivariate data for the concentration of lipids in the blood serum and in the synovial fluid in the knee (KOA-Synovium.csv, KOA-Serum.csv), PCA was to be carried out.

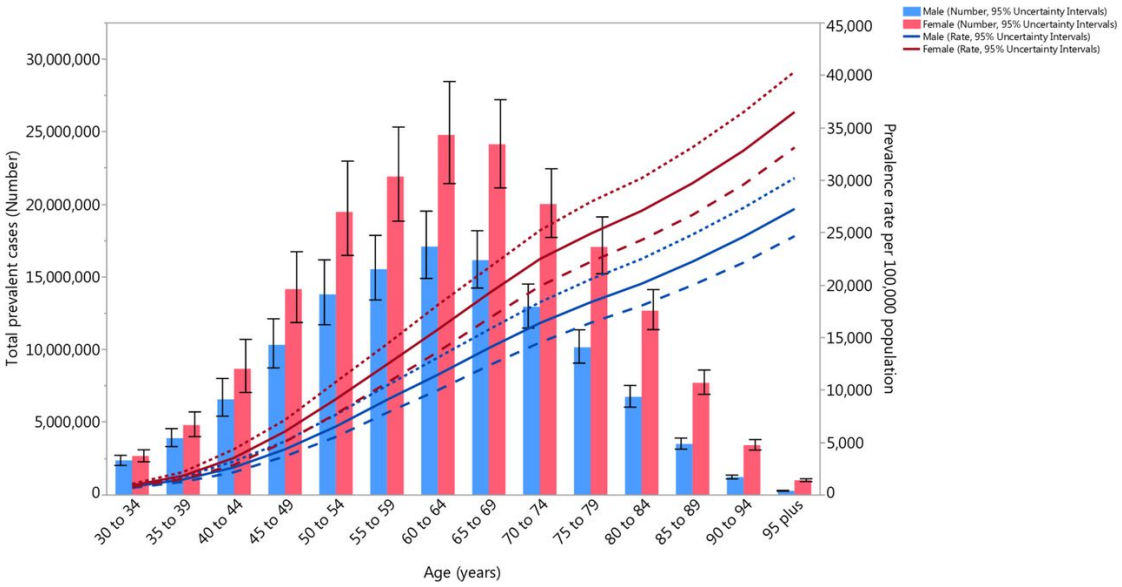

Figure 3: *Prevalence and Incidence of Osteoarthritis: BM Journals, Annals of Rheumatic Diseases*

A.1 Data Description and Analysis Parameters

The first data set (KOA-Synovium) contains the results of the analysis of 91 variables (lipids analysed in the synovium fluid) for 74 individuals, which are identified by an ID and Group. All individuals have been considered *active*, and the categorical variables are considered supplementary. The Group of the individuals (Control, Early OA, Late OA) identifies what the progression of the disease in the individual is. The approach is to determine whether a reduction in dimensionality of the data set, therefore reducing the number of variables analysed, is an appropriate approach to categorise the individuals in the groups they belong to.

A.2 Standardisation of the Variables

Since no evidence pointing out that some lipids should have different weights than others was found, and the 91 lipids were measured in the same units, standardisation was not considered a priority. However the approach depended on the functions used in R, leading to different results when different approaches are applied. Therefore, these have been saved in order to contrast the algorithms on which the R functions are based and propose different alternatives.

- **res.pca** operated over the matrix of 91 quantitative variables where missing values are completed with average values. The algorithm employed followed a different approach from **missMDA**, function in FactoMineR that gives an estimation on the number of dimensions needed to reconstitute the data, using the function **estimncpPCA**.

```
for (i in 1:n){
    # Obtain a submatrix with no NA
        values
    na.rows <- is.na(na.cols[,i])
    row.clean <- na.cols[!na.rows,]
    # Compute the mean for each
        column
    (mean.i <- mean(row.clean[,i]))
    print(mean.i)
    # Complete the original matrix
        with the average values of
        each variable for each
        individual
    for (j in 1:length(na.rows)){
     syn.clean[j, indices[i]] =
         mean.i
    }
  }
```

- **res.pca2** operated over the standardised and normalised matrix of 83 variables with the **princomp()** function of R, using the Singular Value Decomposition (SVD).

- **res.pca3** operated over the standardised and normalised matrix of 83 variables with the **PCA()** function from FactoMineR. The conclusions have been mainly extracted using this approach because of its high accuracy, which will be compared in the following.

It should be noted that the used functions are proved to provide a different degree of efficiency, since the latest is optimized for memory management and can handle larger datasets more efficiently, especially when using sparse matrix representations or incremental SVD. The function **princomp()** of R was finally no longer applied in the further purpose since for compatibility with the software S-PLUS makes use of two functions different from the ones of **PCA()**, and has shown less accuracy. In fact, the SVD protocol of **prcomp()** in R is universally preferred.

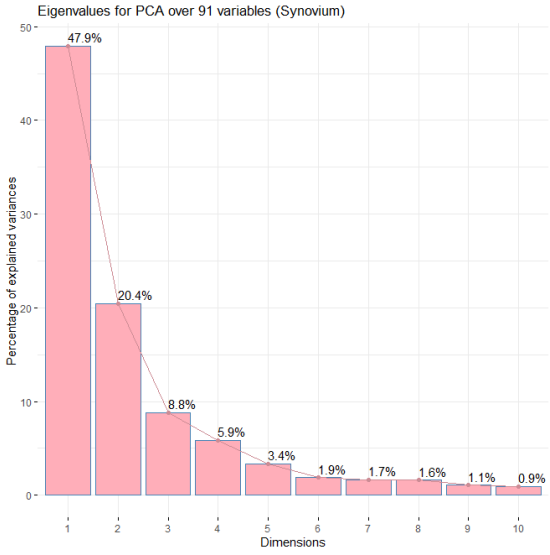

Figure 5: *PCA performed on the non-standardised data set of Synovium, not having removed the variables with NAs (but completed with average values)*

A.3 Computation of the principal components

For the first approach, the function by default in FactoMineR was applied.

```
res.pca <- PCA(syn.clean)
summary(res.pca, nb.dec = 3,
    nbElements = Inf)
```

The inertia explained by each component is represented by each eigenvalue of the correlation matrix, and then plotted in Figure 5. In the second case, PCA was applied after centering and standardising the data:

```
syn.clean2 <- syn.clean[,
    !contains.NA]
syn.norm <- scale(syn.clean2)
corr <- cor(syn.norm)
res.pca2 <- princomp(corr)
```

The object generated by this function provides the standard deviations of the components, the coordinates of the individuals (observations) on the principal components and **rotation matrix** or matrix of loadings, whose columns are the eigenvectors associated to the eigenvalues.

A.4 Comparison between the computations

As shown in Figure 5, 6 and 7, it looks reasonable to assume that the scaling method, although not considered a priority in the first place, did constitute a difference in the results. In fact, the not-scaled approach of Figure 7 shows that the variability explained by the first component is greater than 90%. This can be explained graphically by visualizing that not standardising the data lets the dimensions of greatest variability grow in their direction without considering their relation to the variables not included in the first component. Moreover, the results of Figure 5 and 6 are similar, since the same algorithm has been used in both approaches.

In fact, the proportion explained by the first components is greater when the columns containing NAs are removed, since then the number of dimensions by which we divide when calculating the inertia is lower, meaning that the proportion of explained variance is scaled. If we divide the proportion explained by the first dimension in Figure 6 by the one computed in Figure 5 we can see that the scaling is constant in every PC, as it should be expected. This means removing these variables affects proportionally every principal

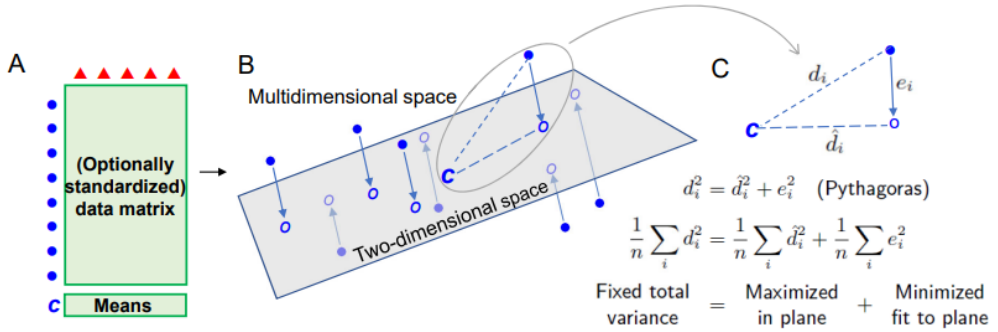

Figure 4: *Schematic view of dimension reduction in PCA. **A**. The rows of data, optionally standardized, and their mean c, define a cloud of points in a multidimensional space. **B**. The first two dimensions of the SVD identify the best-fitting two-dimensional plane in terms of least-squared distances between the plane and the points. **C**. Each multidimensional data point defines a right-angled triangle with its projection onto the plane. The maximization of average squared distances in the plane is equivalent to minimizing the average squared distances from the points to the plane (i.e., minimizing fit).*

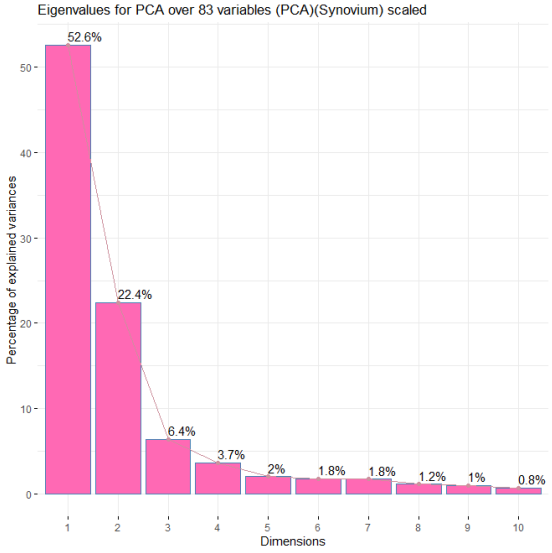

Figure 6: *PCA approached for the standardised data set of Synovium, having removed the variables with NAs, scaling the data.*

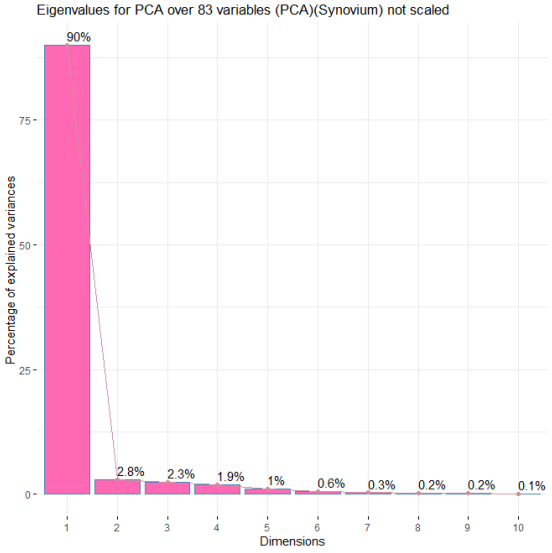

Figure 7: *PCA approached for the standardised data set of Synovium, having removed the variables with NAs, not scaling the data.*

component.

$$\frac{47.9}{52.6} = \frac{20.4}{22.4} = \frac{6.4}{8.8} = 0.9106.$$

Furthermore, PCA would have substituted the NAs columns with average values in the first case automatically as well.

A.5 Choosing the Number of Dimensions to Examine

As Figure 5 shows, the first two main principal components summarise 68% of the total inertia, i.e. 68% of the total variability is represented by the plane. However, for the second case in Figure 6 nearly 74% is explained, showing that the reduction of dimensionality to 2 dimensions is an accurate approach since it minimises the distortion of the new cloud of points after applying the rotation transformation given by PCA. This result is further commented in the following section.

A.6 Studying the cloud of variables

Representing the cloud of variables can be seen as an aid for interpreting the cloud of individuals, since the relations between variables distinguish individuals with opposite values for them. For this purpose, let F_s denote the coordinate vector of the n individuals on component s and $F_s(i)$ its value for individual i . This vector is also called the principal component of rank s, of dimension n (and thus can be considered a variable). One can then calculate the correlation coefficient between the vector F_s and the initial variables to interpret the relative positions of the individuals on the component of rank s. In this case, when the correlation coefficient between F_s and a variable $k \in 1, ..., p$ is positive, an individual with a positive coordinate on component F_s will likely have a high value for variable k.

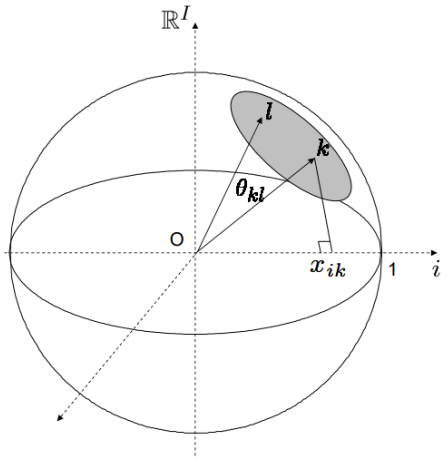

Figure 8: *Scatterplot of the variables in $\mathbb{R}^I$. Here, $\mathbb{R}^I$ stands for $\mathbb{R}^n$, since a variable is a point in the vector space with n dimensions (individuals). The variables k and l are located within the hypersphere.*

In the case of standardised PCA, the variables $1, ..., p$ are located within a hypersphere of radius 1 (see Figure 8) and the scalar product between two variables is expressed as

$$\sum_{i=1}^{n} x_{ik} \times x_{il} = ||k|| \times ||l|| \times cos(\theta_{kl}) \quad (7)$$

Since the data is centered (res.pca), the norm of each variable is equal to its standard deviation multiplied by the square root of n, that is:

$$\sum_{i=1}^{n} (x_{ik} - \bar{x_k}) \times (x_{il} - \bar{x_l}) = n \times s_k \times s_l \times cos(\theta_{kl}) \quad (8)$$

On the left-hand side of the equation we identify the covariance between the variables and after dividing each term in the equation by the standard deviation we get:

$$r(k, l) = cos(\theta_{kl}) \quad (9)$$

The parameter $cos(\theta_{kl})$ is as well given by the function **PCA()** in FactoMineR (squared) and represents the value of the projection of the variable (or individual if studying individuals) on the principal component. Therefore it allows us to measure the importance of the principal component for each observation, in the case of individuals, and the quality of the

representation for the variables on the factor map, if we are studying the variables:

```
head(res.pca3$var$cos2)
```

Selecting a bound for this parameter in the plotting method allows us to restrict ourselves to the variables whose representation is maximal, and therefore we obtain a representation on how significant for the study the analysed parameters are:

```
# Color by cos2 values: quality on the
    factor map
fviz_pca_var(res.pca, col.var = "cos2",
    select.var= list(cos2 = 0.85),
    gradient.cols = c("pink",
        "violetred", "lightpink4"),
    repel = TRUE # Avoid text overlapping
) + ggtitle("PCA point cloud for
    variables in the hypersphere of
    radius 1")
```

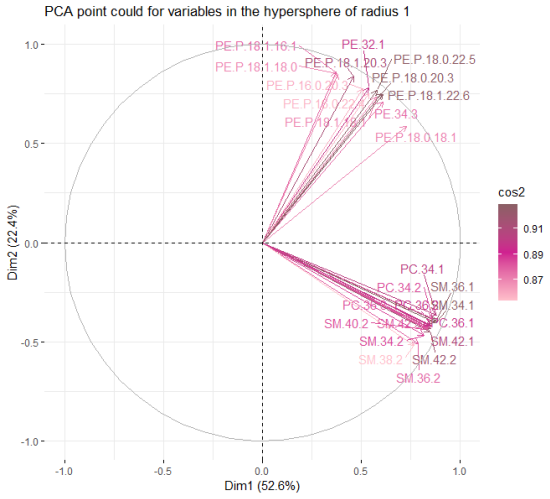

Figure 9: *Scatterplot of the subset of variables N_k more relevant for the study.*

A.7 Studying the cloud of individuals

The results for individuals can be extracted using the function **get_pca_ind()** in the Factoextra package, which provides a list of matrices containing all the results for the individuals (coordinates, correlation between individuals and axes, squared cosine and contributions). The cloud of individuals is a default output of the used function of PCA.

After applying PCA over quantitive variables, the categorical variable Group was added to the data frame. The results of the supplementary categorical variables can also be found in the output of the function **summary.PCA** or in the object **res.pca$quali.sup**. The table contains the coordinates, cosine-squared, and v-tests for each category of patients (Control, Early OA, Later OA).

```
# Add the groups
syn.groups <-
    cbind.data.frame(syn.clean2,
    groups)
head(syn.groups$groups)
res.pca <- PCA(syn.groups[, 1:83])
```

Confidence ellipses we also be drawn around the categories of the supplementary categorical variable, i.e. around the barycentre of the individuals characterised by the category Group (Figures 13 and 14). These ellipses enable us to visualise whether or not two categories differ significantly.

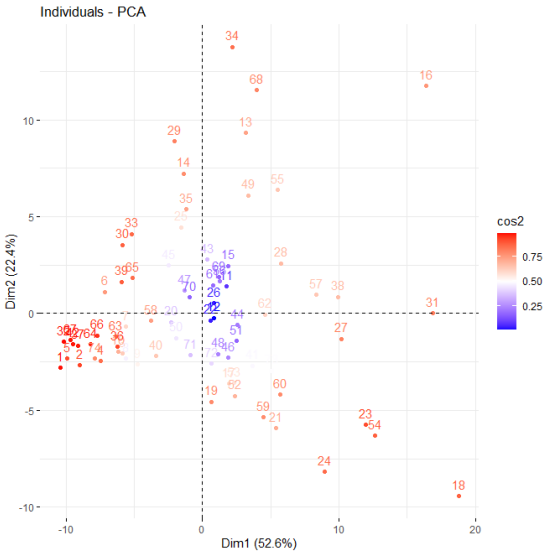

Figure 10: *Scatterplot of the individuals after applying PCA. We can picture which individuals are best represented in the plane created by the first two principal components.*

A.8 Note to the conclusions

In order for the confidence ellipses to be interpreted, the data had to be considered to be normally distributed. This was a reasonable approach since the data was sufficiently

large and we were interested in centres of gravity, and therefore averages. This was accomplished with

```
fviz_pca_ind(res.pca5, habillage
    = syn.groups$groups,
    addEllipses=TRUE,
    ellipse.level=0.65, col.var =
    "cos2", palette =
    c("hotpink", "lightpink",
    "lightpink4")) + ggtitle("PCA
    point cloud for patients
    (ellipses conf. 75%)")
```

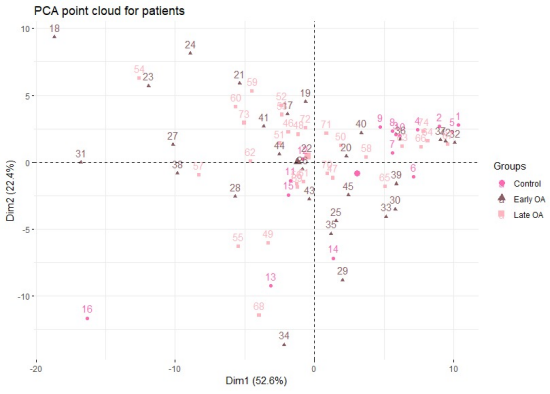

Figure 11: *Scatterplot of the individuals classified by colors for groups.*

When interpreting the results, the representations of both the cloud of individuals and the cloud of variables are to be analysed together. In fact, differences between individuals can be explained by the variables, and relationships between variables are illustrated by individuals. In fact, in Figure 12 have the variables and the individuals been plotted together.

We may note that the coordinates of individuals and variables are not constructed on the same space, and therefore in the biplot one should focus mainly on the direction of variable. In this case, it can be observed that the majority of individuals are plotted on the other side of the variables, which means they are inversely related to them (the majority of individuals have low values for the variables), and fewer individuals (e.g. 31, 16, 55, 60, 27) will have generally high values for the most of them. In fact, it can be expected that individual 31 had a higher value on average on most of the parameters (since its in the average direction of the variables), however 34 mainly only on the PE. P. 18 lipids (top of the plot).

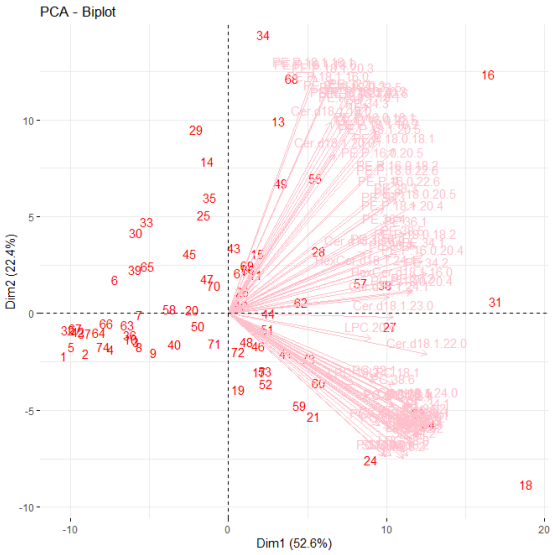

Figure 12: *Biplot of individuals and variables.*

This helps extract three conclusions:

(a) The individual 16 has a remarkable high average on the variables in comparison with other individuals, however, is located in the direction of lighter variables. In contrast is for example individual 60 in the fourth quadrant, which is positioned in the direction of the most relevant variables. Analysing the results on **summary.PCA** confirms this hypothesis by analysing their coordinates on the principal components.

(b) Individuals on the fourth quadrant will most likely have a different category than the ones on the third quadrant since their values of the most relevant variables are opposite, hypothesis that is to be confirmed once the ellipses are plotted.

(c) Since the majority of individuals have lower values for the variables (they are on the left-side of the plot) we can guarantee a certain uniformity in the relevance of the variables when categorising the individuals. This also helps establish relationships between variables that are on the same quadrant of the map, understanding that if their influence over

the individuals is similar, the ones with lighter color may be removed in order to save effort in the analysis.

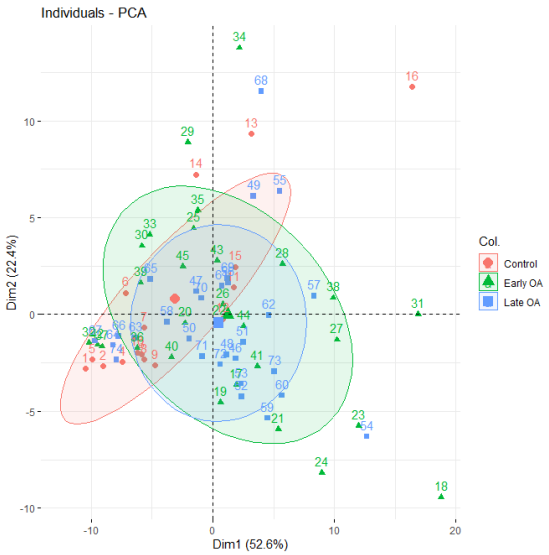

Figure 13: *Biplot of variables and individuals with ellipse confidence levels of 65%.*

A.9 Results of the study

One can conclude from Figure 9 that the most relevant parameters in the dimensionality reduction have been SM 36.1, SM 34.1 and PEP 18.1.22.6, providing a cos2 value higher than 0.95. Moreover, more than 80% of the variables where represented in the plot after filtering with cos2 higher than 0.7, which means that nearly 66 out of 83 variables where close to the plane pictured by the first two components in an angle of less than (1-0.7) * 90 degrees = 27 degrees.

On the whole, the first principal component was mainly constructed in the direction of the variables SM 36.1, SM 34.1 and PEP 18.1.22.6 which led to the construction of a plane that supported 74% of the variability (res.pca). However, the distinction between groups of individuals was not so easily conducted.

In fact, it seems plausible to accept that PCA was a usable approach for drawing the category of Control apart from the OA patients. However, the two categories of OA patients (Early and Later) where not so clearly distinguished after drawing the ellipses, since the Later OA one is included in the Early OA one on a 65% significance level.

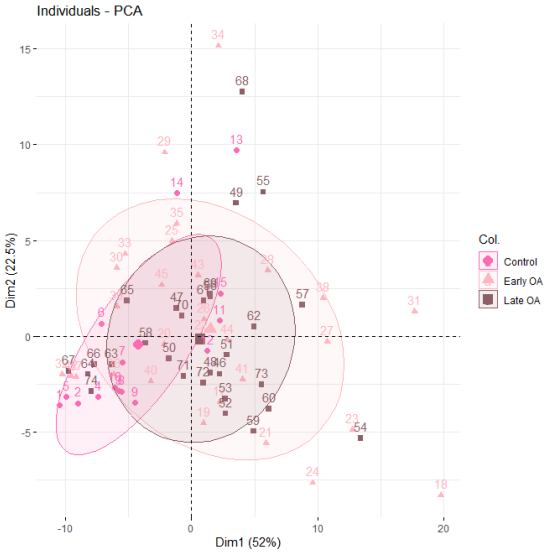

Figure 14: *Biplot of variables and individuals with ellipse confidence levels of 65% having removed the outlier individual 16. The ellipse that categorises the group Control changes its shape remarkably.*

IV. Conclusion

In this report PCA has been described from a theoretical and practical approach in order to clarify its potential when it comes to analysing not only individuals of a dataset, but also the relevance of the variables to their point cloud. In the first place, a theoretical approach based on the books [1] and [2] was conducted, leading to the analysis of a sanitary-oriented case. Finally, conclusions were extracted not only on the categorisation of OA patients but as well comparing the different tools that the libraries of R support to data scientists.

References

[1] K.V.Mardia et al., Multivariate Analysis, A series of Monographs and Textbooks. Academic Press, 1979, pp. 13-16, 213-234.

[2] J. R. Schott, Matrix Analysis for Statistics, Wiley series in probability and statistics, 1997, John Wiley & Sons, pp 84-150.

[3] Hothron Everitt, A Handbook of Statistical Analysis Using R, 3rd Ed. CRC Press, chpt. 15.

[4] Husson et al., Exploratory Multivariate Analysis Using R, 2nd Edition, CRC Press, chpt. 1.

[5] S. Chwartzman, Inference of Eigenvalues and Eigenvectors for Gaussian Symmetric Matrices, Harvard School of Public Health, Universidade de São Paulo and Stanford University.

[6] Greenacre, Michael Groenen, Patrick Hastie, Trevor Iodice D'Enza, Alfonso Markos, Angelos Tuzhilina, Elena. (2022). Principal component analysis.

[7] Sanchez-Lopez E, Coras R, Torres A, Lane NE, Guma M. Synovial inflammation in osteoarthritis progression. Nat Rev Rheumatol. 2022 May;18(5):258-275. doi: 10.1038/s41584-022-00749-9. Epub 2022 Feb 14. PMID: 35165404; PMCID: PMC9050956.

[8] Scanzello CR, Goldring SR. The role of synovitis in osteoarthritis pathogenesis. Bone. 2012 Aug;51(2):249-57. doi: 10.1016/j.bone.2012.02.012. Epub 2012 Feb 22. PMID: 22387238; PMCID: PMC3372675.

V. Appendix

A. Mathematical background (II)

A.1 Spectral Theorem Decomposition

The spectral theorem, or symmetric eigenvalue decomposition (SED) theorem, states that for any symmetric matrix, there are exactly n (possibly not distinct) eigenvalues, and they are all real; further, that the associated eigenvectors can be chosen so as to form an orthonormal basis. This constitutes a special case of the singular value decomposition theorem, not for square matrices. The result offers a simple way to decompose the symmetric matrix as a product of simple transformations.

Theorem 7 *For any symmetric matrix Σ in $\mathbf{S}^n$ it follows the symmetric eigenvalue decomposition*

$$\Sigma = \sum_{i=1}^{n} \lambda_i u_i u_i^T = U \Delta U^T$$

where $\Lambda = diag\ (\lambda_1, \ldots, \lambda_n)$ and the matrix of $U := [u_1, \ldots, u_n]$ is orthogonal (that is, $U^T U = U U^T = I_n$), and contains the eigenvectors of A, while the diagonal matrix Λ contains the eigenvalues of A.

A.2 MLEs invariance

Theorem 8 *(Invariance Property of Maximum Likelihood Estimators)*

Let a distribution be indexed by a parameter θ and τ a bijective transformation. Then if $\bar{\theta}$ is the MLE of θ, then the MLE of $\tau(\theta)$ is $\tau(\bar{\theta})$.

B. Central Limit Theorem

Theorem 9 *(Lindeberg–Lévy CLT)*
Suppose $\$X_1, \ldots, X_n, \ldots\}$ is a sequence of i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$. Then as n approaches infinity, the random variables $\sqrt{n}(\bar{X}_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$.

That means:

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right).$$

C. FactoMineR specifications and advantages

The election to make use of FactoMineR instead of the default functions **prcomp** and **princomp** has been taken according to the following reasons:

- Parallelization: FactoMineR allows execution of computations on multi-core processors.
- Incremental SVD: In situations where the data set is too large to fit into memory, FactoMineR implements an incremental SVD algorithm, processing in smaller batches, computing partial SVDs for each chunk. The partial SVD results are then combined to obtain the final principal components, allowing FactoMineR to handle large datasets that cannot be accommodated entirely in memory.

- Memory-mapped files: FactoMineR employs memory-mapped files to optimize memory usage, allowing the access to disk-based data instead of loading the entire data set into memory.
- Sparse matrix representation: FactoMineR avoids storing zero entries, significantly reducing memory consumption.