Appendix A

Multivariate Statistics

A.1 Characterizing and Displaying Multivariate Data

Let **x** represents a random vector of p variables measured on a sampling unit (subject or object), so

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

where each x_j, $j = 1, \ldots, p$ is a random variable. The population mean vector or *expected value of* **x**. It is defined as a vector of expected values of each variable,

$$\mathbb{E}(\mathbf{x}) = \begin{bmatrix} \mathbb{E}(x_1) \\ \mathbb{E}(x_2) \\ \vdots \\ \mathbb{E}(x_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \vec{\mu}$$

where μ_j is expected value of the jth variable. The population covariance matrix is defined as

$$\Sigma = cov(\mathbf{x}) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \ldots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \ldots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \sigma_{p3} & \ldots & \sigma_{pp} \end{bmatrix}$$

where the diagonal elements $\sigma_{jj} = \sigma_j^2$ are the population variances of the variables x_j, $j = 1, \ldots, p$, and the off-diagonal elements $\sigma_{jk} = cov(x_j, x_k)$, $j, k = 1, \ldots, p$, $j \neq k$ are the population covariances.

In a multivariate random sample of n individuals and p variables, the n *observation vector* are denoted by $\mathbf{x}_1, \ldots, \mathbf{x}_n$, where for each $i = 1 \ldots, n$

$$\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix}$$

The sample mean vector $\bar{\mathbf{x}}$ can be found either as the average of the n observations vectors or by calculating the average of each of the p variables separately:

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} x_{i1}/n \\ \sum_{i=1}^{n} x_{i2}/n \\ \vdots \\ \sum_{i=1}^{n} x_{ip}/n \end{bmatrix}$$

Thus, $\bar{x}_1$ is the mean of the n observations on the first variable, $\bar{x}_2$ is the mean of the second variable, and so on.
The sample covariance matrix is the matrix of sample variances and covariances of the p variables:

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & s_{13} & \ldots & s_{1p} \\ s_{21} & s_{22} & s_{23} & \ldots & s_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & s_{p3} & \ldots & s_{pp} \end{bmatrix}$$

where the diagonal elements $s_{jj} = s_j^2$ are the sample variances of the variables x_j, $j = 1, \ldots, p$, and the off-diagonal elements $s_{jk} = \hat{cov}(x_j, x_k)$, $j, k = 1, \ldots, p$, $j \neq k$ are the sample covariances of x_j and x_k,

$$s_{jj} = s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$$

and

$$s_{jk} = s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

All n observations vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ can be transposed to row vectors and listed in the *data matrix* $\mathbf{X}$ as follows:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x_1^T} \\ \mathbf{x_2^T} \\ \vdots \\ \mathbf{x_n^T} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \ldots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \ldots & x_{np} \end{bmatrix}$$

A.2 Principal Component Analysis PCA. Analysis based on the Covariance matrix

Principal component analysis is a statistical technique that is used to analyze the interrelationships among a large number of variables and to explain these variables in terms of a smaller number of variables, called principal components, with a minimum loss of information.
Let $\mathbf{x} = (\mathbf{x_1}, \ldots, \mathbf{x_p})^{\mathbf{T}}$ be a p dimensional random vector with covariance matrix Σ. Assume that the variables in $\mathbf{x}$ are measure in the same or comparable units (when it does not hold we should work with correlation matrix). The aim now is to find a new coordinate system or a new random vector $\mathbf{y} = (\mathbf{y_1}, \ldots, \mathbf{y_p})^{\mathbf{T}}$ from $\mathbf{x} = (\mathbf{x_1}, \ldots, \mathbf{x_p})^{\mathbf{T}}$ such that:

- the new variables $y_1, \ldots, y_p$ be uncorrelated, and
- the greatest variance of the variables in $\mathbf{x}$ comes to lie on the first coordinate y_1 (called the first principal component), the second greatest variance on the second coordinate y_2, and so on.

Statistically the problem is as follows: Write the jth principal component y_j, $j = 1 \ldots, p$, as a linear combination of the variables in $\mathbf{x}$, that is

$$y_j = \sum_{k=1}^{p} \beta_{jk} x_k = \beta_j^T \mathbf{x} \tag{A.1}$$

where $\beta_j^T = (\beta_{j1}, \ldots, \beta_{jp})$ is some vector of coefficients which maximize the variance of y_j,

$$\mathbb{V}(y_j) = \mathbb{V}(\beta_j^T \mathbf{x}) = \beta_{\mathbf{j}}^{\mathbf{T}} \mathbf{\Sigma} \beta_{\mathbf{j}}, \tag{A.2}$$

(here we assume that $\mathbb{E}(\mathbf{x}) = 0$) subject to the constraint

- $Cov(y_j, y_k) = 0 \quad \forall j \neq k$, and
- $\beta_j^T \beta_j = 1$ and $\beta_j^T \beta_k = 0$, for $j \neq k$.

We find such coefficients β_{jk} using the Spectral Decomposition Theorem of Linear Algebra. Before write this theorem, we recall some results which are in your lectures notes of algebra lineal MS2S01.

Theorem A.1 (Theorem 4.21) *Let A be a symmetric matrix in $M_n(\mathbb{R})$ (the vector space of all $n \times n$ matrices with real entries). Then the following hold:*

1. *All eigenvalues of A are real, although these don't need to be distinct.*

2. *A is diagonalizable, i.e., there exitst a non-singular matrix C (non-singular means that C has an inverse, so $det(C) \neq 0$) such that*

$$C^{-1}AC = D$$

where D is a diagonal matrix (see Definition 3.6 in the lectures notes of MS2S01)

3. *Suppose that $\mathbf{v}$ is an eigenvector for A associated to an eigenvalue λ, and $\mathbf{w}$ is an eigenvector for A associated to an eigenvalue μ. If $\lambda \neq \mu$, then $\mathbf{v} \cdot \mathbf{w} = 0$, so, $\mathbf{v}$ and $\mathbf{w}$ are orthogonal.*

Remark A.1 *Orthogonality is a concept that relies on the notion of an inner product. Abstractly: Inner product space The usual inner product on Euclidean space is the dot product, where you multiply each "coordinate" pairwise and then add, but you can have other inner products. For PCA we will need orthonormality, that is $\mathbf{v} \cdot \mathbf{w} = 0$, if $\mathbf{v} \neq \mathbf{u}$ and $\mathbf{v} \cdot \mathbf{w} = 1$, if $\mathbf{v} = \mathbf{u}$. See Chapter 4 in the lectures notes of MS2S01.*

Remark A.2

- *By Lemma 3.7 in the lectures notes of MS2S01, the diagonal entries of D coincide with the eigenvalues of A. What is more an eigenvalue λ appears on the diagonal of D a total of the algebraic multiplicity of λ.*

- *By Corollary 4.22 in the lectures notes of MS2S01, if A is an $n \times n$ symmetric matrix (that is $A = A^T$), then it is possible to find an orthogonal matrix $C \in M_n(\mathbb{R})$ (that is $C^{-1} = C^T$) such that*

$$C^T AC = C^{-1}AC$$

is diagonal.

The orthogonal matrix C will be the matrix whose columns are the eigenvectors of A.
Now we are ready to write the spectral Theorem.

Theorem A.2 *(Spectral Decomposition Theorem): Let A be a symmetric $n \times n$ matrix, then A has a spectral decomposition $A = CDC^{-1}$ where C is an $n \times n$ orthogonal matrix (that is $C^{-1} = C^T$, so $CC^T = I$ or $C^TC = I$) whose columns are unit eigenvectors $C_1, \ldots, C_n$ corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_n$ of A and D is the $n \times n$ diagonal matrix whose main diagonal consists of $\lambda_1, \ldots, \lambda_n$.*

Remark A.3 *The covariance and correlation matrices are symmetric matrices and positive semidefinite matrices. (A symmetric matrix A is said to be positive definite if $x^TAx > 0$ for all possible vectors x except $x = 0$. Similarly A is positive semi-definite if $x^TAx \geq 0$ for all $x \neq 0$). Then their eigenvalues are always real and non-negative values.*

But how is this theorem used to find the β_{jk}'s?
Well, since the covariance matrix is symmetric $p \times p$, by the Spectral Decomposition Theorem, it follows that

$$\Sigma = \beta D \beta^T \tag{A.3}$$

where here β is a $p \times p$ matrix whose columns are unit eigenvectors $\beta_1, \ldots, \beta_p$ corresponding to the eigenvalues $\lambda_1 \ldots, \lambda_p$ of Σ and D is the $p \times p$ diagonal matrix whose main diagonal consists of $\lambda_1, \ldots, \lambda_p$.
Since the columns vectors of β are orthonormal, that is, $\beta_j \cdot \beta_k = \beta_j^T \beta_k = 0$ if $j \neq k$ and $\beta_j \cdot \beta_k = \beta_j^T \beta_k = 1$ if $j = k$, and taking $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p$, then by (A.2), (A.3) and after some computations we can see

$$\mathbb{V}(\mathbf{y}) = \beta^T \Sigma \beta = \beta^T \beta D \beta \beta^T = D$$

so,

$$\mathbb{V}(y_j) = \lambda_j \quad \text{and} \quad cov(y_j, y_k) = 0, j \neq k.$$

On the other hand, recall that the trace of a matrix A, written $Tr(A)$, is the sum of the diagonal entries of the matrix. Let A be an $n \times n$ matrix with complex entries, and let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of A (counted with multiplicity). Then by properties of eigenvalues we have

$$Tr(A) = \lambda_1 + \cdots + \lambda_n,$$

that is, the trace of a symmetric matrix is the sum of its eigenvalues. Applying this to the matrix Σ what we have is

$$\sum_{j=1}^{p} \sigma_j^2 = \sum_{j=1}^{p} \lambda_j.$$

Thus, the total variance of the variables in **x** can be expressed as $trace(\Sigma) = \sum_{j=1}^{p} \lambda_j$ which is also the total variance for **y**. Then, the portion of the total variance (of **x** or **y**) explained by the jth principal component y_j is

$$\frac{\lambda_j}{\sum_{j=1}^{p} \lambda_j}.$$

Our goal is to find a reduced number of principal components that can explain most of the total variance, i.e. we seek a value of m that is as low as possible but such that the ratio $\sum_{j=1}^{m} \lambda_j / \sum_{j=1}^{m} \lambda_j$ is close to 1.

Remark A.4 *Since the population covariance Σ is unknown, we will use the sample covariance matrix* **S** *as an estimate and proceed as above using* **S** *in place of Σ.*

A.3 Factor Analysis

In this model we again consider p variables $x_1, \ldots, x_p$ and observed data for each of these variables. Our objective is to identify m factors $f_1, \ldots, f_m$, preferably with $m \leq p$ as small as possible, which explain the observed data more succinctly. We next suppose that each x_i can be represented as a linear combination of the factors as follows:

$$x_j = \beta_{j0} + \sum_{k=1}^{m} \beta_{jk} f_k + \varepsilon_j \tag{A.4}$$

where the ε_j are the components which are not explained by the linear relationship. We assume that

- $\mathbb{E}(\varepsilon_j) = 0$, $\mathbb{V}(\varepsilon_j) = \phi_j$, $j = 1 \ldots, p$ and $cov(\varepsilon_j, \varepsilon_k) = 0, j \neq k$.
- The factors are independent with mean 0 and variance 1, i.e $\mathbb{E}(f_k) = 0$, $\mathbb{V}(f_k) = 1$, $k = 1, \ldots, m$, and $cov(f_k, f_\ell) = 0, k \neq \ell$.
- In addition, we assume that $cov(\varepsilon_j, y_k) = 0$ for all j and k,

We can consider (A.4) to be a series of regression equations. The coefficient β_{jk} is called the **loading** of the jth variable on the kth factor. The coefficient ε_j is called the **specific factor** for the ith variable.
Now observe that since

$$\begin{aligned}\mu_j = \mathbb{E}(x_j) =& \mathbb{E}\left(\beta_{j0} + \sum_{k=1}^{m} \beta_{jk} f_k + \varepsilon_j\right) \\ =& \mathbb{E}(\beta_{j0}) + \sum_{k=1}^{m} \beta_{jk}\mathbb{E}(f_k) + \mathbb{E}(\varepsilon_j) = \beta_{j0} + 0 + 0 = \beta_{j0},\end{aligned}$$

it follows that the intercept term $\beta_{j0} = \mu_j$, and so the regression equations can be expressed as

$$x_j = \mu_j + \sum_{k=1}^{m} \beta_{jk} f_k + \varepsilon_j. \tag{A.5}$$

Let $\beta = [\beta_{jk}]$ be the $p \times m$ **matrix of loading factors** and let $\varepsilon = [\varepsilon_j]$ be the $p \times 1$ column **vector of specific factors**, then we can express (A.5) as

$$\mathbf{x} = \vec{\mu} + \beta\mathbf{f} + \varepsilon.$$

where $\mathbf{x}$, $\mathbf{f}$ and $\vec{\mu}$ are the vectors of original variables, the factors and the mean of $\mathbf{x}$.

From the assumptions of the model and after some calculations we can see that

$$\sigma_j^2 = \mathbb{V}(x_j) = \beta_{j1}^2 + \cdots + \beta_{jm}^2 + \phi_j,$$

and

$$\sigma_{jk} = cov(x_j, x_k) = \sum_{\ell=1}^{m} \beta_{j\ell}\beta_{k\ell}$$

which play an importatn role in our development. From these equivalences it follows that the population covariance matrix Σ for **x** has the form

$$\Sigma = \beta\beta^T + \Phi$$

where $\beta = [\beta_{jk}]$ is the $p \times m$ matrix of loadings, Φ is the $p \times p$ diagonal matrix with ϕ_i in the ith position on the diagonal.
In addition we can also see that

$$cov(x_j, f_k) = \beta_{jk}.$$

The sum $\beta_{j1}^2 + \cdots + \beta_{jm}^2 = \sum_{k=1}^{m} \beta_{jk}^2 := h_j^2$ is call the **communality** of x_j. It refers to the *common variance* of the factors into the variance of x_j, while $\phi_j = \mathbb{V}(\varepsilon_j)$ is the specific variance also called *specificity, unique variance* or *residual variance*. *Thus, we have a partitioning of the variance of x_j into a component due to the common factors, called the communality, and a component unique to x_j, called the specific variance*.

A.3.1 Estimation of loading and communalities

There are different approach to estimate the loading and communalities. Here in this lectures notes we just describe one, namely, "Principal component method". This name is perhaps unfortunate in that it adds to the confusion between factor analysis and principal component analysis. We are not going to calculate any principal components. The reason for that name is that we will use the spectral theorem, as we did for PCA.

Remark A.5 *In the principal component approach, Φ is neglected, thus Σ is factor just as $\Sigma = \beta\beta^T$.*

In order to factor Σ in that way, we use the spectral decomposition, Theorem A.2, so we can write

$$\Sigma = CDC^T,$$

where C is an orthogonal matrix constructed with normalized eigenvectors of Σ as columns and D is a diagonal matrix with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ of Σ on the diagonal.

Since the eigenvalues of a positive semidefinite matrix are nonegative real numbers, then we can factor D into

$$D = D^{1/2}D^{1/2}.$$

With this factoring of D, then we can write

$$\Sigma = CDC^T = CD^{1/2}D^{1/2}C^T = \left(CD^{1/2}\right)\left(CD^{1/2}\right)^T,$$

which is of the form $\beta\beta^T$, but we do NOT define $\beta = \left(CD^{1/2}\right)$ because this is a $p \times p$ matrix and we are seeking for a $m \times p$ matrix, $m < p$. We therefore define $D_1 = diag(\lambda_1, \ldots, \lambda_m)$, m the largest eigenvalues $\lambda_1 \geq \ldots, \lambda_m$ and $C_1 = (c_1, \ldots, c_m)$ the matrix with the corresponding eigenvectors as its columns. We then approach β as

$$\hat{\beta} = C_1 D_1^{1/2} = \left(\sqrt{\lambda_1}c_1, \ldots, \sqrt{\lambda_m}c_m\right).$$

Observe that the jth diagonal element of $\hat{\beta}\hat{\beta}^T$ is the sum of the squares of the jth row of $\hat{\beta}$ or $\sum_{k=1}^{m} \hat{\beta}_{jk}^2$. Hence to complete the approximation of Σ, we define

$$\hat{\phi}_j = \sigma_j^2 - \sum_{k=1}^{m} \hat{\beta}_{jk}^2$$

and write

$$\Sigma \approx \hat{\beta}\hat{\beta}^T + \hat{\Phi} \tag{A.6}$$

where $\hat{\Phi} = diag(\hat{\phi}_1, \ldots, \hat{\phi}_p)$. Thus in (A.6) the variances on the diagonal of Σ are modeled exactly, but the off-diagonal covariances are only approximated. This is the challenge of factor analysis.
In this method of estimation, the sums of the squares of the rows and columns of $\hat{\beta}$ are equal to the communatlities and the eigenvalues, respectively. That is,

$$\hat{h}_j^2 = \sum_{k=1}^{m} \hat{\beta}_{jk}^2$$

and

$$\lambda_k = \sum_{j=1}^{p} \hat{\beta}_{jk}^2,$$

this last equality is due to the fact that the normalized eigenvectors (columns of C) have length 1.

Thus the kth factor contributes $\hat{\beta}_{jk}^2$ to σ_j^2. The contribution of the kth factor to the "total" variance, $tr(\Sigma) = \sigma_1^2 + \cdots + \sigma_p^2$, is therefore, λ_k. Then, the proportion of the total variance due to the kth factor is

$$\frac{\lambda_k}{tr(\Sigma)} = \frac{\lambda_k}{\sum_{i=1}^{m} \lambda_i}.$$

In practice we use the correlation matrix R instead Σ. Often that gives better results. Recall that $tr(R) = p$, where p is the number of variables.
There are other approaches to estimation of the loading and communalities. For example, the **Iterated Principal Factor Method** improves the estimates of communality. The idea is to use the previous calculation to obtain initial communality estimates. Then, apply an iterative process to obtain new communalities until the communality estimates converges.

A.4 Cluster Analysis (Clustering algorithms)

In cluster analysis we search for patterns in a data set by grouping the (multivariate) observations into clusters. The goal is to find an optimal grouping for which the observations or objects within each cluster are similar, but the cluster are dissimilar to each other. Two common approaches to clustering the observations vectors are hierarchical clustering and partitioning.
Hierarchical clustering: It typically starts with n clusters, one for each observation, and end with a single cluster containing all n observations. Is possible also reverse this process, that is, start with a single cluster containing all n observations and end with n clusters of a single item each.
Partitioning clustering: It simply divides the observations in g clusters. This can be done by starting with an initial partitioning or with cluster enters and then reallocating the observations according to some optimal criterion.

A.4.1 Measures of Distance

Since cluster analysis attempts to identify the observation vectors that are similar and group them into clusters, many techniques use an index of *similarity or proximity* between each par of observations. A common distance function is the Euclidian distance between two vectors $\mathbf{x} = (x_1, \ldots, x_p)^T$ and $\mathbf{y} = (y_1, \ldots, y_p)^T$ defined as

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

A.4.2 Hierarchical clusters

Single Linkage (Nearest Neighbor)

In the single linkage method, the distance between two clusters A and B is defined as the minimum distance between a point in A and a point in B:

$$D(A, B) = min\{d(\mathbf{x}_i, \mathbf{x}_j), \text{ for } \mathbf{x}_i \text{ in A and } \mathbf{x}_j \text{ in } B\}$$

where $d(\mathbf{x}_i, \mathbf{x}_j)$ is the Euclidean distance.
In this algorithm, at each step, the distance D is found for every pair of clusters, and the two clusters with smallest distance are merged. The number of clusters is therefore reduced by 1. After two clusters are merged, the procedure is repeated for the next step: the distance between all pair of clusters are calculated again, and the pair with minimum distance is merged into a single cluster.
The result of a hierarchical clustering procedure can be displayed graphically using a *tree diagram*, also known as a *dendrogram*, which shows all the steps in the hierarchical procedure, including the distances at which clusters are merged. Different definitions of distances between clusters produce different methods, for instance

Complete Linkage and Average Linkage

In the Complete Linkage algorithm the distance between two clusters A and B is defined as the maximum distance between a point in A and a point in B, so

$$D_1(A, B) = max\{d(\mathbf{x}_i, \mathbf{x}_j), \text{ for } \mathbf{x}_i \text{ in A and } \mathbf{x}_j \text{ in } B\}$$

At each step, the distance D_1 is found for every pair of clusters, and the two clusters with the smallest distance are merged.
In the average link approach, the distance between two clusters A and B is defined as the average of the $n_A n_B$ distances between the n_A points in A and the n_B points in B:

$$D_2(A, B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{i=1}^{n_B} d(\mathbf{x}_i, \mathbf{x}_j)$$

where the sum is over all $\mathbf{x}_i$ in A and all $\mathbf{x}_j$ in B. At each step, we join the two clusters with the smallest distance D_2.

A.4.3 Partitioning

k-Means

The basic k-means clustering algorithm is defined as follows:

- Step 1: Choose the number of clusters k
- Step 2: Make an initial selection of k centroids (there are many ways to do that, for example select at random k observations that are at least a Euclidean distance r apart)
- Step 3: Assign each data element in $\mathbf{X}$ to its nearest centroid (in this way k clusters are formed one for each centroid, where each cluster consists of all the data elements assigned to that centroid)
- Step 4: For each cluster make a new selection of its centroid
- Step 5: Go back to step 3, repeating the process until the centroids don't change (or some other convergence criterion is met)

There are various choices available for each step in the process.
An alternative version of the algorithm is as follows:

- Step 1: Choose the number of clusters k
- Step 2: Make an initial assignment of the data elements to the k clusters (there are many ways to do that)
- Step 3: For each cluster select its centroid
- Step 4: Based on centroids make a new assignment of data elements to the k clusters
- Step 5: Go back to step 3, repeating the process until the centroids don't change (or some other convergence criterion is met)

Bibliography

[1] Alvin C. Rencher. Methods of Multivariate Analysis. Second Edition. While Series in Probability and Statistics, 2002.