STAT 32940 Fall 2018

Mutlivariate Analysis Overview

Notes by: Ariel Boyarsky (aboyarsky@uchicago.edu)

Based on lectures by Dr. Lek-Heng Lim at the University of Chicago

Preface

The following are a condensed set of notes for a course in Multivariate Analysis. These notes are based on the handouts and lectures of Dr. Lek-Heng Lim. My additions are limited to some proofs and useful subject matters. Any mistakes are my own, please do not hesitate to email me if any are found. The material covers topics in advanced multivariate analysis also known as unsupervised learning. We omit a discussion of linear regression as we expect students will have already learned it. The only prerequisite is a standard course in linear algebra, though a course in real analysis would be helpful.

Contents

1 Basic Matrix Theory

We briefly review some familiar concepts that will be useful in the following sections.

1.1 Important Definitions

Definition 1.1 *(Orthogonality). Two vectors are said to be orthogonal iff* $a \cdot b = 0$. *That is perpendicular to each other.*

Definition 1.2 *(Mutual Orthogonality). A set of vectors,* $V = \{v_1, \ldots, v_n\}$, *are said to be mutually orthogonal if* $\forall\ a, b \in V$ *we have* $a \cdot b = 0$.

Definition 1.3 *(Orthonormality). A set of vectors* V *is orthonormal if* $\forall\ v \in V\ ||v||_2 = 1$ *and the set if mutually orthogonal.*

1.2 Norms

Definition 1.4 *A norm follows the following properties,*

1. $||x|| \geq 0$ *with equality* $\iff x = 0$ *(Positive Definiteness)*
2. $||ax|| = |a|||x||$ *for* $a \in \mathbb{R}, x \in \mathbb{R}^n$ *(Scalar)*
3. $||x + y|| \leq ||x|| + ||y||$ *(Triangle Inequality)*

A very useful class of norms are the L^p or p-norms.

Definition 1.5 *Take* $x \in \mathbb{R}^n$. *The p-norm is defined as,*

$$||x||_p = (|x_1|^p + \cdots + |x_n|^p)^{\frac{1}{p}}$$

The most famous of these norms is the $p = 2$ norm, more commonly referred to as the euclidean norm or the L^2 norm,

$$||x||_2 = (\sum_{i=1}^{n} |x_i|^2)^{1/2}$$

A useful fact about the relationship of norms and inner products follows.

Lemma 1.6 $||x||_2^2 = x^T x$.

1.3 Matrix Norms

Definition 1.7 *A matrix norm will satisfy,*

1. $||x|| \geq 0$ *with equality* $\iff x = 0$ *(Positive Definiteness)*

2. $||ax|| = |a|||x||$ *for* $a \in \mathbb{R}, x \in \mathbb{R}^n$ *(Scalar)*

3. $||x + y|| \leq ||x|| + ||y||$ *(Triangle Inequality)*

And sometimes a further property of submultiplicativity ($||AB|| \leq ||A||||B||$).

Definition 1.8 *(Matrix 2-Norm).*

$$||A||_2 = \max_{x \neq 0} \frac{||Ax||_2}{||x||_2}$$

Definition 1.9 *(Frobenius Norm).*

$$||A||_F = (\sum_i \sum_j |a_{ij}|^2)^{1/2}$$

Definition 1.10 *(Operator Norms). A class of norms called is operator or induced norms defined as,*

$$||A||_{p,q} = \max_{x \neq 0} \frac{||Ax||_p}{||x||_q} = \max\{||Ax||_p \ : \ ||x||_q \leq 1\}$$

1.4 Eigenvalue Decomposition (EVD)

Recall that the generalized eigenvalue problem is defined as for $A \in \mathbb{R}^{n \times n}$,

$$Ax = \lambda Bx \tag{1}$$

Setting, $B = I$ yields the familiar eigenvalue problem. Recall that we may find the eigenvalues using the characteristic polynomial:

$$P(A) = det(A - \lambda I)$$

Where the eigenvalues are the roots of the polynomial. To calculate the eigenvectors simply compute teh solutions to the system given by,

$$(A - \lambda I)[x_1, \ldots, x_n]^T = \mathbf{0}$$

Plugging in the eigenvalue for λ.

The eigenvalue decomposition takes the form,

$$A = Q \Lambda Q^{-1} \tag{2}$$

Where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$

Theorem 1.11 *(Spectral Theorem for symmetric matrices).*

Let $A \in \mathbb{R}^{n \times n}$ *is symmetric* ($A^T = A$) *iff*

$$EVD(A) = V \Lambda V^T$$

where $V^T = V^{-1}$ *that is* V *is orthogonal.*

Proof: *Proof in book (not nesc).* ■

Lemma 1.12 *This also implies that the eigenvalues of a real symmetric matrix are real.*

Proof: $\lambda < v, \bar{v} >=< \lambda v, \bar{v} >=< Av, \bar{v} >=< v, \bar{A}v >=< v, \bar{\lambda}\bar{v} >=< v, \bar{v} > \bar{\lambda} \implies \lambda = \bar{\lambda}$ ■

1.5 Singular Value Decomposition (SVD)

The singular value decomposition of a matrix $A \in \mathbb{R}^{n \times p}$ is defined as,

$$A = U\Sigma V^T \tag{3}$$

Where, $U \in \mathbb{R}^{n \times n}$ are the left singular values and $V \in \mathbb{R}^{p \times p}$ are the right singular values. And Σ is a diagonal matrix of singular values that is $\text{diag}(\sigma_1, \sigma_r)$ where $r = \text{rank}(A)$. There is also the condensed SVD where we remove the 0 columns and rows.

In condensed SVD we can write,

$$A = \sum_i^r \sigma_i \mathbf{u}_i \mathbf{v}_i$$

Theorem 1.13 *(Existence of SVD).*

Every matrix has a condensed SVD.

Proof: Take $A \in \mathbb{R}^{n \times p}$. Then define,

$$W = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} \in \mathbb{R}^{(n+p)\times(p+n)}$$

Notice this matrix is symmetric ($W = W^T$) as such by the spectral theorem for Hermitian matricies we have that,

$$W = Z\Lambda Z^T$$

Furthermore write,

$$W[x, y]^T = \sigma[x, y]^T \implies Ay = \sigma x, A^T x = \sigma y$$

Notice, if we apply W to z where we negate y we get,

$$W \begin{bmatrix} x \\ -y \end{bmatrix} = \begin{bmatrix} -Ay \\ A^T x \end{bmatrix} = \begin{bmatrix} -\sigma x \\ \sigma y \end{bmatrix} = -\sigma \begin{bmatrix} -x \\ -y \end{bmatrix}$$

Thus, $-\sigma$ is also an eigenvalue and $\Lambda = \text{diag}(\sigma_1, \ldots, \sigma_r, -\sigma_1, \ldots, -\sigma_r, 0, \ldots, 0)$. Since, eigenvectors are orthogonal (W is symmetric). Then if we scale such that $z^T z = 2$ we get the system,

$$x^T x + y^y = 2, x^T x - y^T y = 0 \implies x^T x = y^T y = 1$$

Now represent the Z matrix of normalized eigenvectors as,

$$\tilde{Z} = \frac{1}{\sqrt{2}} \begin{bmatrix} X & X//Y & -Y \end{bmatrix} \implies Z\Lambda Z^T = \tilde{Z}\Lambda\tilde{Z^T}$$

Now set $\tilde{\Lambda}$ such that we remove 0 diagonals. Then we can write,

$$\begin{aligned} W = Z\Lambda Z^T &= \tilde{Z}\tilde{\Lambda}\tilde{Z}^T \\ &= \frac{1}{2}\begin{bmatrix} X & X \\ Y & -Y \end{bmatrix}\begin{bmatrix} \Sigma_r & 0 \\ 0 & -\Sigma_r \end{bmatrix}\begin{bmatrix} X & X \\ Y & -Y \end{bmatrix}^T \\ &= \begin{bmatrix} 0 & X\Sigma_r Y^T \\ Y\Sigma_r X^T & 0 \end{bmatrix} \end{aligned}$$

As such, $A = X\Sigma_r Y^T$ and $A^T = Y\Sigma_r X^T$ as well as the fact this implies orthonormality of the columns yields the SVD. ■

A useful fact in helping calculate SVD by hand is a consequence of the above proof.

Lemma 1.14 *The square of the singular values of A are the eigenvalues of AA^T and A^TA. Simmilarly, the left singular values are given by the eigenvectors of AA^T and the right singular values are given by the eigenvectors of A^TA.*

Proof: Let $A = U\Sigma V^T$. Then, $AA^T = U\Sigma V^T V\Sigma U^T = U\Sigma^2 U^T = U\Sigma^2 U^{-1} \underset{(EVD)}{=} AA^T$.
The other way is analogous. ■

Lemma 1.15 *EVD is equivalent to SVD in symmetric positive definite matrices.*

Proof:

$$A = A^{1/2}A^{T\frac{1}{2}} = U\Sigma^{1/2}V^T V\Sigma^{1/2}U^T = U\Sigma U^T$$

Thus, $A = Q\Lambda Q^T$, so, $U = Q = V$ and $\Lambda = \Sigma$. ■

1.6 Subspaces

There are 4 fundamental subspaces. They are,

1. $\ker(A) = \{x \in \mathbb{R}^n \mid Ax = 0\}$
2. $\text{im}(A) = \{y \in \mathbb{R}^n \mid A = y\}$
3. $\ker(A^T)$
4. $\text{im}(A^T)$

Notice, $\ker(A^T) = \text{im}(A)^\perp$ and $\text{im}(A^T) = \ker(A)^\perp$.
To see this take $y \in \text{im}(A^T), x \in \ker(A), y = A^T u \implies x^T y = x^T(A^T u) = (Ax)^T u = 0 \implies \text{im}(A^T) \subset \ker(A)^\perp$.

Furthermore, $\mathbb{R}^p = \text{im}(A^T) \oplus \ker A$ and $\mathbb{R}^n = \text{im}(A) \oplus \ker A^T$

Lemma 1.16 *SVD may be used to get these subspaces. Such that,*

- $\ker(A) = span(v_{r+1}, \ldots, v_p)$
- $im(A) = span(u_1, \ldots, v_r)$
- $\ker(A^T) = span(u_{r+1}, \ldots, u_p)$
- $im(A^T) = span(v_1, \ldots, v_r)$

1.7 Moore-Penrose Pseudo Inverse

Theorem 1.17 *(Moore-Penrose).* $\forall\ A \in \mathbb{R}^{n\times p}\ \exists\ A^\dagger = X \in \mathbb{R}^{p\times n}$ *s.t.* ,

1. $(AX)^T = AX$
2. $(XA)^T = XA$
3. $AXA = A$
4. $XAX = X$

If A^{-1} exists then $A^\dagger = A^{-1}$. Just plug it in to see. However, $(AB)^\dagger \neq B^\dagger A^\dagger$ and $A^\dagger A \neq I$.

A very useful fact is that if $D = \text{diag}(d_1, \ldots, d_n)$ then $D^\dagger = \text{diag}(\delta_1, \ldots, \delta_n)$ where $d_i \neq 0$ then $\delta_i = 1/d_i$ otherwise $\delta_i = 0$.

This is nice because it means we can use SVD to easily calculate the psuedoinverse.

Lemma 1.18 *Let* $A = U\Sigma V^T$ *then,* $A^\dagger = V\Sigma^{-1}U^T$ *where* $\Sigma^{-1} = diag(1/\sigma_1, \ldots, 1/\sigma_r, 0, \ldots, 0)$.

Furthermore, if $A \in \mathbb{R}^{n\times p}$ then,

$$\text{rank}(A) = n \implies A^\dagger = A^T(AA^T)^{-1}$$

Or,

$$\text{rank}(A) = p \implies A^\dagger = (AA^T)^{-1}A^T$$

1.8 Projections

Definition 1.19 *(Projection Matrices).* $P \in \mathbb{R}^{n\times n}$ *is said to be a projection if it is idempotent* ($P^2 = P$).

Definition 1.20 *(Orthogonal Projection Matrices).* $P \in \mathbb{R}^{n\times n}$ *is said to be an orthogonal projection if it is a projection, i.e. idempotent* ($P^2 = P$)*, and symmetric* ($P^T = P$).

Notice that if P is a projection matrix then so is $I - P$. Furthermore, if P is an orthogonal projection matrix then if $\text{im}(P) = W$ and $\text{im}(I - P) = W'$ implies that $\mathbb{R}^n = W \oplus W'$.

Finally, we can show that $AA^\dagger$ is an orthogonal projection matrix.

2 Important Matrices in Multivariate Analysis

Before, we get into the methods of multivariate analysis. It is useful to review some basic statistical concepts and more specifically their matrix counterparts. We break these up into sample and population parameters.

2.1 Sample Mean Matrices

Definition 2.1 *(Sample Mean Vector). The sample mean vector of* $X \in \mathbb{R}^{n\times p}$ *is defined as* $\bar{x} = \frac{1}{n}X^T\mathbf{1}$. *Where* $\mathbf{1} \in \mathbb{R}^{n\times 1}$ *a vector of all 1's.* $\bar{x} \in \mathbb{R}^{p\times 1}$.

Definition 2.2 *(Matrix of means). The sample mean matrix of* $X \in \mathbb{R}^{n\times p}$ *is defined as* $\mathbf{1}\bar{x}^T = \frac{1}{n}\mathbf{1}\mathbf{1}^T X$. *Where* $\mathbf{1} \in \mathbb{R}^{n\times 1}$ *a vector of all 1's.* $\mathbf{1}\bar{x} \in \mathbb{R}^{n\times p}$.

Definition 2.3 *(Demeaning Matrix). The demeaning matrix demeans (mean centers) a data matrix by the means of the columns i.e.* $X - \mathbf{1}\bar{x}^T$. *It is defined as* $H := I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$

2.2 Sample Covariance Matrices

Definition 2.4 *(Sample Covariance Matrix). The sample covariance matrix is defined as* $S_n := \frac{1}{n}(X - \mathbf{1}\bar{x})^T(X - \mathbf{1}\bar{x}^T) = X^T HX$

Definition 2.5 *(Sample Variance Matrix). The sample variance matrix is defined as* $D := diag(S_n)$. *That is* S_n*'s diagonal elements.*

Definition 2.6 *(Sample Standard Deviations Matrix). The sample SD matrix is defined as* $D^{1/2}$.

Which of course implies,

Definition 2.7 *(Sample Correlation Matrix). The sample Variance matrix is defined as* $R := D^{-1/2}S_nD^{-1/2}$.

Notice, that if we have a small sample size and care about unbiased estimates we may switch $\frac{1}{n}$ to $\frac{1}{n-1}$.

2.3 Population Matrices

In the previous section we introduced sample parameter matrices. It is useful to review the difference between sample and population. Specifically, we may regard the sample as drawn from the population variables. Indeed, we often think of the population in terms of random variables.

Definition 2.8 *(Random Variable). Recall, that a random variable, X, is a measurable function from the sample space to* $\mathbb{R}$. *That is,*

$$X : \Omega \to \mathbb{R}$$

As such, samples are simply realization of the random variable of the form $X(\omega)$ where $\omega \in \Omega$.

When we deal with populations we use the expectation with respect to the Lebesgue measure,[1]

$$\mathbb{E}(X) = \int_\Omega X(\omega)P(d\omega)$$

We can use this to define the population mean matrix.

[1] For this course it is not important to know the details of Lebesgue integration.

Definition 2.9 *(Population Mean Vector). Let $X = [X_1, X_2, \ldots, X_p]^T$ be a random matrix. Then,*

$$\mathbb{E}[X] = \begin{bmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_p] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \mu \in \mathbb{R}^p$$

Using this we can define the population covariance matrix.

Definition 2.10 *(Population Covariance Matrix). Let $X = [X_1, X_2, \ldots, X_p]^T$ be a random matrix. Then, the population covariance matrix is given by $Cov(X) = \mathbb{E}[(X - \mu)(X - \mu)^T]$*

Now let's recall some useful facts.

1. $\mathbb{E}[aX] = a\mathbb{E}[X]$
2. $Var(aX) = a^T Cov(X)a$
3. $Cov(a^T X, b^T X) = a^T Cov(X)b$

3 Procrustes Analysis

We now move on to studying the applications of the theory we built above specifically the applications of it to the tools of multivariate analysis. These tools are also sometimes called unsupervised learning in the Machine Learning literature.

The goal of Procrustes analysis is to find a Q matrix that will minimize the distance between two matrices. There are several versions.

3.1 Orthogonal Procrustes Analysis

$$\min_{Q \in O(p)} ||A - BQ||_F \tag{4}$$

Where $O(p)$ denotes the class of orthogonal matrices. We can find the solution to this problem by expanding. That is,

$$||A - BQ||_F^2 = \text{tr}(A^T A) + \text{tr}(B^T B) - 2\text{tr}(Q^T B^T A)$$

Which implies we can solve this by solving for $Q = \text{argmax}_Q \text{tr}(Q^T B^T A)$.

Then let $C = B^T A$ such that, $\text{tr}(Q^T C) = \text{tr}(Q^T U \Sigma V^T) = \text{tr}(V^T Q^T U \Sigma)$. Then, since the trace of orthogonal matrices is bounded from above by 1 we have that

$$\max \text{tr}(V^T Q^T U \Sigma) = \sigma_1$$

But, then we just want to pick Q such that U and V go away, i.e. $V^T Q U = I$. Which in itself implies the solution,

$$Q = UV^T \tag{5}$$

A special case arises if $B = I$, specifically, we can just take $A = U\Sigma V^T$ and still $Q = UV^T$.

3.2 Symmetric Procrustes Analysis

Here we can define the problem as,

$$\min_{X=X^T} ||A - X||_F \tag{6}$$

In this case we can recall that we can write any matrix as a skew-symmetric matrix that is,

$$A = \frac{1}{2}(A + A^T) + \frac{1}{2}(A - A^T)$$

Thus, X will be the symmetric component, i.e. $X = \frac{1}{2}(A + A^T)$.

3.3 Best Rank-r Approximation

Here want to approx A with a matrix of rank r. Notice, the problem is trivial ($X = A$) if rank(()A) $= r$. We define the optimization problem,

$$\min_{\text{rank}(X)\leq r} ||A - X|| \tag{7}$$

There is a nice solution to this whenever $||\cdot||$ is orthogonally invariant. That is, $||AQ|| = ||A||$ when Q is an orthogonal matrix. The solution comes from the Eckart-Young theorem.

Theorem 3.1 *Let $A = U\Sigma V^T$. Then, whenever $||\cdot||$ is orthogonally invariant, the argument solution to equation 7 is given by,*

$$X = U\Sigma_r V^T \text{ where } \Sigma_r = diag(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)$$

and if $||\cdot|| = ||\cdot||_2$ then the minimized value is σ_{r+1}. Notice, for F-norm it is $\sqrt{\sigma_{r+1}^2 + \cdots + \sigma_{rank(A)}^2}$.

4 Principle Component Analysis (PCA)

4.1 Population PCA

$$a_k = \text{argmax}\{Var(a^T X) \mid ||a||_2 = 1, Cov(a^T X, a_1^T X) = \cdots = Cov(a^T X, a_{k-1}^T X) = 0\} \tag{8}$$

Notice that $a_k = q_k$ where $\Sigma = Cov(X)$ and $\Sigma = Q\Lambda Q^T$. The maximized value (i.e. the variance explained by the k'th PC) is given by λ_k eigenvalue.

4.2 SVD Method for Sample PCA

Let X_c be the mean centered data matrix. Then, $X_c = U\Sigma V^T$ yields all the information.

Notice that,

$$S_n = \frac{1}{n-1} X_c^T X_c = \frac{1}{n-1} V\Sigma^2 V^T$$

This implies that eigenvalues of the sample covariance matrix are given by $\sigma_k^2/(n-1)$. And $a_k = q_k = v_k$ are the PC vectors.

This method also avoids floating point errors if there is a value in the mean centered matrix close to 0.

4.3 SVD Projections

To project data onto j-th and k-th PCs simply plot $(\sigma_j u_{ij}, \sigma_k u_{ik})$. Also there is no need to compute variable PCA. To plot variables onto PCs, simply plot $(\sigma_j v_{ij}, \sigma_k v_{ik})$.

You can also do a biplot but then you need to decide which points to scale by singular values (you also don't have to and could just plot the u's and v's).

Notice these methods work because,

$$X_c = U\Sigma V^T = (U\sigma)V^T = \begin{bmatrix} \sigma_1 u_{11} & \cdots \\ \vdots & \ddots \end{bmatrix} [v_1, \ldots, v_p]$$

And recall we just want to project $P_w x = (x_i^T q_j)q_j + (x_i^T q_k)q_k$ such that with EVD we would plot $(x_i^T q_j, x_i^T q_k)$. This is of course much more expensive.

5 Factor Analysis

5.1 The Model

In matrix form we write,

$$X = \mu + LF + \epsilon$$

We also make the following assumptions.

Assumptions:

1. $\mathbb{E}[\epsilon] = 0$
2. $\mathbb{E}[\epsilon_i \epsilon_j] = 0$
3. $\mathbb{E}[F_i F_j] = 0$ $(Cov(F) = I)$
4. $\mathbb{E}[\epsilon F] = 0$
5. $\mathbb{E}[F_i] = 0$
6. $Var[F_i] = 1$

5.2 Some Intuition

In general there are too many free variables so finding L is impossible. So there are different ways to do it using approximations.

In finance, X is usually returns of some asset. The u's are called alphas and measure risk.

1. $\mu > 0$: Risk is too high for return.
2. $\mu < 0$: Risk is too low for return.
3. $\mu = 0$: Risk matches return properly.

The L is a loading matrix of betas telling you about return correlation with the factors (F) which are usually macroeconomic variables (GDP, Interest Rates, etc.), Stat (i.e. Fama-French 3 factor model), or unobserved (i.e. innate ability).

1. $l > 1$: High direct correlation with factors.
2. $l < 0$: Inverse correlation with factors.
3. $l \in [0, 1]$: Low correlation with factors.

5.3 PCA Approach

We can use PCA to satisfy the assumptions and approximate L.

5.3.1 Covariance Structure

First, we will derive covariance structure.

$$\begin{aligned} Cov(X) = \Sigma &= \mathbb{E}[(X-\mu)(X-\mu)] = \mathbb{E}[(LF+\epsilon)(LF+\epsilon)] \\ &= \mathbb{E}[LL^T FF^T] + 2\mathbb{E}[LF\epsilon] + \mathbb{E}[\epsilon^T \epsilon] \\ &= LL^T \mathbb{E}[F^T F] + \Psi \\ &= LL^T + \Psi \end{aligned}$$

Where $\Psi = \text{diag}(Var(\epsilon_i)) \; \forall \; i = 1, \ldots, n$.

Furthermore, we see that

$$Cov(X, F) = \mathbb{E}[(X-\mu)F] = \mathbb{E}[(LF+\epsilon)F^T] = \mathbb{E}[LFF^T] + \mathbb{E}[\epsilon F^T] = L$$

As such we can see that the key formula is,

$$\Sigma = LL^T + \Psi \tag{9}$$

5.3.2 Solving for L

To solve, fix $m < p$. Then $S = Q\Lambda Q^T$. Set $\Lambda_m = \text{diag}(\lambda_i, \ldots, \lambda_m)$ and $Q = [q_1, \ldots, q_m]$.

Then,

$$L = Q_m \Lambda_M^{1/2} = \sqrt{\lambda_i} q_i$$
$$\Psi = \text{diag}(S - LL^T)$$

Notice, this is a rank-m approximation. Because, $LL^T = Q_m \Lambda_m^{1/2} \Lambda_m^{1/2} Q_m^T = Q_m \Lambda_m Q_m^T$ and SVD = EVD in the case of symmetric matrices. Because of this based on the norm you are using, you know the value of the minimization i.e. $||\cdot||_2 \implies \lambda_{m+1}$.

The proportion of variability explained by each factor is given by $\lambda_i / \sum_i s_{ii}$. As always you can do this with correlation matrix and it would be equivalent to not only mean centering but also scaling the data by the inverse std. deviation.

6 Canonical Correlation Analysis (CCA)

This is very much like PCA except we care about correlation and we want to compare correlations of two data sets i.e. X and Y.

We need a bit of setup. Let $W = [X, Y]^T$ and $Cov(W) = \begin{bmatrix} Cov(X) & Cov(XY) \\ Cov(YX) & Cov(Y) \end{bmatrix}$.

The canonical correlation variables are given by $U = a^T X$ and $V = b^T Y$. The problem is defined as,

$$\begin{aligned} (U_k, V_k) &= \text{argmax}\{Corr(U_k, V_k)\} \\ &s.t.\ Var(U_k) = Var(V_k) = 1 \\ &\quad Cov(V_i, U_j) = Cov(V_i, V_j) = Cov(U_i, U_j) = 0 \end{aligned} \tag{10}$$

The value of the maximum is the canonical correlation ρ_k.

6.1 SVD Solution

First find,

$$G_{XY} = \Sigma_x^{-1/2} \Sigma_{xy} \Sigma_y^{-1/2} \tag{11}$$

Notice this is essentially the R matrix.

Then, $G_{XY} = U \Sigma V^T$ which then gives us

$$U_k = u_k \Sigma_x^{-1/2} X$$

And

$$V_k = v_k \Sigma_y^{-1/2} X$$

And the canonical correlation for the k-th variable is given by the singular value, $\rho_k = \sigma_k$.

Theorem 6.1 *(CCA Vector Relation). One property of CCA is that if $X' = MX + d$ and $Y' = NX + c$. Then, correlations are the same and the variables simply differ by M^{-T} and N^{-T}. That is,*

$$a'_k = M^{-T} a_k \qquad b'_k = N^{-T} b_k$$

The nice thing about this is that if you use correlation matrices instead of the Σ's you can get the variables because $a'_{k,cor} = V_X^{-1/2} a_{k,cov}$. The same is true for b just use $V_Y^{-1/2}$.

Aside (Square Root Calculation). You can calculate square roots by doing EVD and taking square roots of the eigenvalue matrix. Consider,

$$A = Q\Lambda Q^{-1} \implies A^{1/2} = Q\Lambda^{1/2} Q^{-1}$$

Where $\Lambda^{1/2} = \text{diag}(\sqrt{\lambda_i})$.

6.2 Alternative SVD Method

It is possible that G_{XY} as defined in Equation 11 is not easily retrievable especially when doing this by hand. In this case simply use,

$$H_{XY} = S_X^{-1/2} G_{XY} G^T XY S_X^{1/2} = S_X^{-1} S_{XY} S_Y^{-1} S_{YX}$$

and

$$H_{YX} = S_Y^{-1/2} G_{XY} G^T XY S_Y^{1/2} = S_Y^{-1} S_{YX} S_X^{-1} S_{XY}$$

Then, the eigenvalues of H_{XY} will be the eigenvalues of $G_{XY} G_{XY}^T$ and the eigenvalues of H_{YX} will be the eigenvalues of $G_{YX} G_{YX}^T$. Thus, there is no need to take square roots of matrices. Simply take the square root of the eigenvalues. Also notice that we can also do this with sample correlation matrices.

7 Linear Discriminant Analysis (LDA)

Supervised vs. Unsupervised Learning

This method is a departure from the previous multivariate techniques in that this is an example of supervised learning. That is, instead of inferring information directly from the data (e.g. unsupervised learning), we split data into test and training data.

7.1 Preliminaries

First we need to define some matrices. Suppose that we have g groups or classes.

The group mean matrix is defined as,

$$\bar{x}_i := \frac{1}{n_i} X_i^T 1$$

The total sum of squares matrix is defined as,

$$T := X^T H X$$

The within group sum of squares matrix is defined as,

$$W := \sum_{i}^{g} X^T H X$$

The between group sum of squares matrix is defined as,

$$B := \sum_{i}^{g} n_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})$$

Notice that $T = B + W$.

An alternate to the within group matrix is the pooled variance matrix defined as,

$$S_{\text{pool}} := \frac{1}{n-g} \sum_{i}^{g} (n_i - 1) S_i$$

Where the group variance matrix is defined as,

$$S_i := \frac{1}{n_i - 1} (X_i - 1\bar{x}_i)(X_i - 1\bar{x}_i)$$

Notice that $W = (n-g) S_{\text{pool}}$.

7.2 The Linear Discriminant Function

The linear discriminant function or Fischer discriminant function is defined as $f(a) = \frac{a^T B a}{a^T W a}$.[2] Using this function we can define the LDA optimization problem,

$$\max_{a \neq 0} \{ \frac{a^T B a}{a^T W a} \} \tag{12}$$

Theorem 7.1 *The solution to Equation 12 is given by the principle eigenvector of $W^{-1}B$ such that $a = q_1$ and the value of the function will be $\lambda_{\max}(W^{-1}B)$.*

Proof: The proof of this is a simple constrained optimization problem.

Notice that we can rewrite this as

$$\max_{a \neq 0} a^T B a \quad s.t. \quad a^T W a = 1$$

Then write the Lagrangian for the problem,

$$\mathcal{L} = a^T B a - \lambda (a^T W a - 1)$$

Then the first order condition is,

$$2Ba = 2\lambda W a \implies Ba = \lambda W a$$

Notice, that what we have here is the generalized eigenvalue problem (Notice that if $W = I$ then we get the familiar eigenvalue problem). As such, notice we can solve this by inverting W.

[2] This is the same R.A. Fischer responsible for Fischer Information!

$$W^{-1}Ba = \lambda a$$

Therefore we can satisfy the optimality condition by choosing a to be an eigenvector yielding the corresponding eigenvalue as the maximal value. Furthermore, it is clear that if we choose the $a = q_1$ (the principle eigenvector where $\lambda_1 > \cdots > \lambda_n$) we will solve the maximization problem. This also yields the proof of the theorem.

■

7.3 Classification Rule

So how do we classify new data? The classification rule is then given by (for a new vector t),

$$i = \text{argmin}\{|q_1(t - \bar{x}_j)| \ : \ \forall j \in 1, \ldots, g\}$$

7.3.1 Binary Classification

There is a nice simplification of the above process in the case of binary classification.

Consider that in a two group case we have,

$$\begin{aligned}(q_1^T(t - \bar{x}_2)) - (q_1^T(t - \bar{x}_1)) = & \\ &= (q_1^T(t - \bar{x}_2) - q_1^T(t - \bar{x}_1)))(q_1^T(t - \bar{x}_2) + q_1^T(t - \bar{x}_1)) \\ &= [q_1^T(\bar{x}_1 - \bar{x}_2)][q_1^T(2t - (\bar{x_1} + \bar{x_2}))] \\ &= 2[q_1^T(\bar{x}_1 - \bar{x}_2)][q_1^T(t - \frac{\bar{x}_1 + \bar{x}_2}{2})]\end{aligned}$$

Then notice that, $2[q_1^T(\bar{x}_1 - \bar{x}_2)] = (W^{-1}d)^T d = d^T W d > 0$. Where we let $d = (\bar{x}_1 - \bar{x}_2)$. Then we can write the classification rule as,

$$Class(t) = \begin{cases} 1 & (\bar{x}_1 - \bar{x}_2)^T W^{-1}(t - \frac{\bar{x}_1 + \bar{x}_2}{2}) > 0 \\ 2 & (\bar{x}_1 - \bar{x}_2)^T W^{-1}(t - \frac{\bar{x}_1 + \bar{x}_2}{2}) < 0 \end{cases} \tag{13}$$

Where in the binary case, $W = (n_1 - 1)S_1 + (n_2 - 1)S_2$

8 Correspondence Analysis

Todo

8.1 Application to HITS Algorithm

8.2 Application to Information Retrieval

9 Multidimensional Scaling

Definition 9.1 *(Euclidean Distance Matrix) The EDM is a symmetric matrix* $d_{ij} = [(x_i - x_j)^T(x_i - x_j)]^{1/2}$. *Where* $x_1, \ldots, x_n \in \mathbb{R}^p$. *As such we have,*

$$D = \begin{bmatrix} d_{11} & \ldots & d_{1p} \\ \vdots & \ddots & \vdots \\ d_{n1} & \ldots & d_{nn} \end{bmatrix}$$

Notice, however that we need not use the euclidean distance and that we may define a distance matrix using any metric we wish.

Furthermore, notice that the distance metric in invariant up to column wise translations. That is if each column is translated (+ or −) by some $c = [c_1, \ldots, c_p]$ the EDM does not change.

As such we may define that the data and mean centered matrix are equal (i.e. that $\bar{x} = 0$).

Definition 9.2 *(Gram Matrix) The Gram matrix (also known as the inner product matrix) is defined as,*

$$G = XX^T$$

Such that $g_{ij} = x_i^T x_j$

Theorem 9.3 *Given a symmetric matrix* $D = d_{ij} \in \mathbb{R}^{n \times m}$ *define*

$$G = g_{ij} = -\frac{1}{2}(d_{ij}^2 - \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 - \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2) + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2)$$

Then, D is a EDM iff G is positive semidefinite. And thus, G is the Gram matrix of X.

In the case where the distance matrix is the EDM this means we can recover X. Notice,

$$G = Q\Lambda Q^T = XX^T \implies X = Q\Lambda^{1/2}$$