MAHALANOBIS DISTANCE

Def. The euclidian distance between two points $x = (x_1, \dots, x_p)^t$ and $y = (y_1, \dots, y_p)^t$ in the p-dimensional space $\mathbb{R}^p$ is defined as

$$d_E(x, y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_p - y_p)^2} = \sqrt{(x - y)^t(x - y)}$$

and $d_E(x, 0) = \|x\|_2 = \sqrt{x_1^2 + \cdots + x_p^2} = \sqrt{x^t x}$ is the euclidian norm of x.

It follows immediately that all points with the same distance of the origin $\|x\|_2 = c$ satisfy $x_1^2 + \cdots + x_p^2 = c^2$ which is the equation of a spheroid. This means that all components of an observation x contribute equally to the euclidian distance of x from the center. However in statistics we prefer a distance that for each of the components (the variables) takes the variability of that variable into accountwhen determining its distance from the center. Components with high variability should receive less weight than components with low variability. This can be obtained by rescaling the components

Denote

$$u = (\frac{x_1}{s_1}, \dots, \frac{x_p}{s_p}) \text{ and } v = (\frac{y_1}{s_1}, \dots, \frac{y_p}{s_p})$$

then define the distance between x and y as

$$d(x, y) = d_E(u, v) = \sqrt{(\frac{x_1 - y_1}{s_1})^2 + \cdots + (\frac{x_p - y_p}{s_p})^2} = \sqrt{(x - y)^t D^{-1}(x - y)}$$

where $D = \text{diag}(s_1^2, \dots, s_p^2)$. Now the norm of x equals

$$\|x\| = d(x, 0) = d_E(u, 0) = \|u\|_2 = \sqrt{(\frac{x_1}{s_1})^2 + \cdots + (\frac{x_p}{s_p})^2} = \sqrt{x^t D^{-1} x}$$

and all points with the same distance of the origin $\|x\| = c$ satisfy

$$(\frac{x_1}{s_1})^2 + \cdots + (\frac{x_p}{s_p})^2 = c^2$$

which is the equation of an ellipsoid centered at the origin with principal axes equal to the coordinate axes.

Finally, we also want to take the correlation between variables into account when computing statistical distances. Correlation means that there are associations between the variables. Therefore, we want the axes of ellipsoid to reflect this correlation. This is obtained by allowing the axes of the ellipsoid at constant distance to rotate. This yields the following general form for the statistical distance of two points

Def. The statistical distance or Mahalanobis distance between two points $x = (x_1, \ldots, x_p)^t$ and $y = (y_1, \ldots, y_p)^t$ in the p-dimensional space $\mathbb{R}^p$ is defined as

$$d_S(x, y) = \sqrt{(x-y)^t S^{-1}(x-y)}$$

and $d_S(x, 0) = \|x\|_S = \sqrt{x^t S^{-1} x}$ is the norm of x.

Points with the same distance of the origin $\|x\|_S = c$ satisfy

$$x^t S^{-1} x = c^2$$

which is the general equation of an ellipsoid centered at the origin. In general the center of the observations will differ from the origin and we will be interested in the distance of an observation from its center $\bar{x}$ given by $d_S(x, \bar{x}) = \sqrt{(x-\bar{x})^t S^{-1}(x-\bar{x})}$.

Result 1. Consider any three p-dimensional observations x, y and z of a p-dimensional random variable $X = (x_1, \ldots, X_p)^t$. The Mahalanobis distance satisfies the following properties

- $d_S(x, y) = d_S(y, x)$
- $d_S(x, y) > 0$ if $x \neq y$
- $d_S(x, y) = 0$ if $x = y$
- $d_S(x, y) \leq d_S(x, z) + d_S(z, y)$ (triangle inequality)

MATRIX ALGEBRA

Def. A p-dimensional square matrix Q is orthogonal if

$$QQ^t = Q^tQ = I_p \text{ or equivalently } Q^t = Q^{-1}$$

This implies that the rows of Q have unit norms and are orthogonal. The columns have the same property.

Def. A p-dimensional square matrix A has an eigenvalue λ with corresponding eigenvector $x \neq 0$ if

$$Ax = \lambda x$$

If the eigenvector x is normalized, which means that $\|x\| = 1$, then we will denote the normalized eigenvector by e.

Result 1. A symmetric p-dimensional square matrix A has p pairs of eigenvalues and eigenvectors

$$(\lambda_1, e_1), \ldots, (\lambda_p, e_p).$$

The eigenvectors can be chosen to be normalized ($e_1^t e_1 = \cdots = e_p^t e_p = 1$) and orthogonal ($e_i^t e_j = 0$ if $i \neq j$). If all eigenvalues are different, then the eigenvectors are unique.

Result 2. Spectral decomposition
The spectral decompositon of a p-dimensional symmetric square matrix A is given by

$$A = \lambda_1 e_1 e_1^t + \cdots + \lambda_p e_p e_p^t$$

where $(\lambda_1, e_1), \ldots, (\lambda_p, e_p)$ are the eigenvalue/normalized eigenvector pairs of A.

Example Consider the symmetric matrix

$$A = \begin{pmatrix} 13 & -4 & 2 \\ -4 & 13 & -2 \\ 2 & -2 & 10 \end{pmatrix}$$

From the characteristic equation $|A - \lambda I_3| = 0$ we obtain the eigenvalues $\lambda_1 = 9$, $\lambda_2 = 9$, and $\lambda_3 = 18$ The corresponding normalized eigenvectors are solutions of the equations $Ae_i = \lambda_i e_i$ for $i = 1, 2, 3$. For example, with $e_3 = (e_{13}, e_{23}, e_{33})^t$ the equation $Ae_3 = \lambda_3 e_3$ gives

$$\begin{aligned} 13e_{13} - 4e_{23} + 2e_{33} &= 18e_{13} \\ -4e_{13} + 13e_{23} - 23_{33} &= 18e_{23} \\ 2e_{13} - 2e_{23} + 10e_{33} &= 18e_{33} \end{aligned}$$

Solving this system of equations yields the normalized eigenvector $e_3 = (2/3, -2/3, 1/3)$. For the other eigenvalue $\lambda_1 = \lambda_2 = 9$ the corresponding eigenvectors are not unique. An orthogonal pair is given by $e_1 = (1/\sqrt{2}, 1/\sqrt{2}, 0)^t$ and $e_2 = (1/\sqrt{18}, -1/\sqrt{18}, -4/\sqrt{18})$. With these solutions it can now easily be verified that

$$A = \lambda_1 e_1 e_1^t + \lambda_2 e_2 e_2^t + \lambda_3 e_3 e_3^t$$

Def. A symmetric $p \times p$ matrix A is called **nonnegative definite** if

$$0 \leq x^t A x \text{ for all } x \in \mathbb{R}^p.$$

A is called **positive definite** if

$$0 < x^t A x \text{ for all } x \neq 0.$$

It follows (from the spectral decomposition) that A is positive definite if and only if all eigenvalues of A are strictly positive and A is nonnegative definite if and only if all eigenvalues are greater than or equal to zero.

Remark The Mahalanobis distance of a point was defined as $d^2_S(x,0) = x^t S^{-1} x$ which does implies that all eigenvalues of the symmetric matrix S^{-1} have to be positive.

From the spectral decomposition we obtain that a symmetric positive definite p-dimensional square matrix A equals

$$A = \sum_{i=1}^{p} \lambda_i e_i e_i^t = P \Lambda P^t$$

with $P = (e_1, \ldots, e_p)$ a p-dimensional square matrix whose columns are the normalized eigenvectors of A and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ a p-dimensional diagonal matrix whose diagonal elements are the eigenvalues of A. Note that P is an orthogonal matrix. It follows that

$$A^{-1} = P \Lambda^{-1} P^t = \sum_{i=1}^{p} \frac{1}{\lambda_i} e_i e_i^t$$

and we define the **square root** of A by

$$A^{1/2} = \sum_{i=1}^{p} \sqrt{\lambda_i} e_i e_i^t = P \Lambda^{1/2} P^t$$

Result 3. The square root of a symmetric, positive definite $p \times p$ matrix A has the following properties

- $(A^{1/2})^t = A^{1/2}$ (that is, $A^{1/2}$ is symmetric).
- $A^{1/2} A^{1/2} = A$.
- $A^{-1/2} = (A^{1/2})^{-1} = \sum_{i=1}^{p} \frac{1}{\sqrt{\lambda_i}} e_i e_i^t = P \Lambda^{-1/2} P^t$
- $A^{1/2} A^{-1/2} = A^{-1/2} A^{1/2} = I_p$
- $A^{-1/2} A^{-1/2} = A^{-1}$

Result 4. Cauchy-Schwarz inequality
Let $b, d \in \mathbb{R}^p$ be two p-dimensional vectors, then we have that

$$(b^t d)^2 \leq (b^t b)(d^t d)$$

with equality if and only if there exists a constant $c \in \mathbb{R}$ such that $b = cd$.

Result 5. Extended Cauchy-Schwarz inequality
Let $b, d \in \mathbb{R}^p$ be two p-dimensional vectors and B a p-dimensional positive definite matrix, then

$$(b^t d)^2 \leq (b^t B b)(d^t B^{-1} d)$$

with equality if and only if there exists a constant $c \in \mathbb{R}$ such that $b = cB^{-1}d$.

Proof. The result is obvious if $b = 0$ or $d = 0$. For other cases we use that

$$b^t d = b^t B^{1/2} B^{-1/2} d = (B^{1/2} b)^t (B^{-1/2} d)$$

and apply the previous result to $(B^{1/2}b)$ and $(B^{-1/2}d)$. □

Result 6. Maximization Lemma
Let B be a p-dimensional positive definite matrix and $d \in \mathbb{R}^p$ a p-dimensional vector, then

$$\max_{x \neq 0} \frac{(x^t d)^2}{x^t B x} = d^t B^{-1} d$$

with the maximum attained when there exists a constant $c \neq 0$ such that $x = cB^{-1}d$

Proof. From the previous result, we have that

$$(x^t d)^2 \leq (x^t B x)(d^t B^{-1} d)$$

Because $x \neq 0$ and B positive definite, $x^t Bx > 0$, which yields

$$\frac{(x^t d)^2}{x^t Bx} \leq d^t B^{-1} d$$

for all $x \neq 0$. From the extended Cauchy-Schwarz inequality we know that the maximum is attained for $x = cB^{-1}d$. □

Result 7. Maximization of Quadratic forms
Let B be a p-dimensional positive definite matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ and associated normalized eigenvectors $e_1, \dots, e_p$. Then

$$\begin{aligned} \max_{x \neq 0} \frac{x^t Bx}{x^t x} &= \lambda_1 \text{ (attained when } x = e_1) \\ \min_{x \neq 0} \frac{x^t Bx}{x^t x} &= \lambda_p \text{ (attained when } x = e_p) \end{aligned}$$

More general,

$$\max_{x \perp e_1, \dots, e_k} \frac{x^t Bx}{x^t x} = \lambda_{k+1} \text{ (attained when } x = e_{k+1}, k = 1, \dots, p-1)$$

Proof. We will proof the first result. $B = P \Lambda P^t$ and denote $y = P^t x$. Then $x \neq 0$ implies $y \neq 0$ and

$$\frac{x^t Bx}{x^t x} = \frac{y^t \Lambda y}{y^t y} = \frac{\sum_{i=1}^p \lambda_i y_i^2}{\sum_{i=1}^p y_i^2} \leq \lambda_1$$

Now take $x = e_1$, then $y = P^t e_1 = (1, 0, \dots, 0)^t$ such that $y^t \Lambda y / y^t y = \lambda_1$. □

Remark Note that since

$$\max_{x \neq 0} \frac{x^t Bx}{x^t x} = \max_{\|x\|=1} x^t Bx$$

the prevoius results shows that λ_1 is the maximal value and λ_p is the smallest value of the quadratic form $x^t Bx$ on the unit sphere.

Result 8. Singular Value Decomposition
Let A be a $(m \times k)$ matrix. Then there exist an $m \times m$ orthogonal matrix U, a $k \times k$ orthogonal matrix V and an $m \times k$ matrix Λ with entries (i,i) equal to $\lambda_i \geq 0$ for $i = 1, \ldots, r = \min(m,k)$ and all other entries zero such that

$$A = U \Lambda V^t = \sum_{i=1}^{r} \lambda_i u_i v_i^t.$$

The positive constants λ_i are called the singular values of A. The (λ_i^2, u_i) are the eigenvalue/eigenvector pairs of AA^t with $\lambda_{r+1} = \cdots = \lambda_m = 0$ if $m > k$ and then $v_i = \lambda_i^{-1} A^t u_i$. Alternatively, (λ_i^2, v_i) are the eigenvalue/eigenvector pairs of $A^t A$ with $\lambda_{r+1} = \cdots = \lambda_k = 0$ if $k > m$

RANDOM VECTORS

Suppose that $X = (X_1, \ldots, X_p)^t$ is a p-dimensional vector of random variables, also called a **random vector**. Each of the components of X is a univariate random variable X_j $(j = 1, \ldots, p)$ with its own marginal distribution having expected value $\mu_j = E[X_j]$ and variance $\sigma_j^2 = E[(X_j - \mu_j)^2]$.

The **expected value of** X is then defined as the vector of expected values of its components, that is

$$E[X] = (E[X_1], \ldots, E[X_p])^t = (\mu_1, \ldots, \mu_p)^t = \mu.$$

The population **covariance matrix of** X is defined as

$$\text{Cov}[X] = E[(X - \mu)(X - \mu)^t] = \Sigma.$$

That is, the diagonal elements of Σ equal $E[(X_j - \mu_j)^2] = \sigma_j^2$. The off-diagonal elements of Σ equal $E[(X_j - \mu_j)(X_k - \mu_k)] = \text{Cov}(X_j, X_k)$ $(j \neq k) = \sigma_{jk}$, the covariance between the variables X_j and X_k. Note that $\sigma_{jj} = \sigma_j^2$.

The population correlation between two variables X_j and X_k is defined as

$$\rho_{jk} = \frac{\sigma_{jk}}{\sigma_j \sigma_k}$$

and measures the amount of linear association between the two variables.

The population **correlation matrix** of X is then defined as

$$\rho = \begin{pmatrix} 1 & \rho_{12} & \ldots & \rho_{1p} \\ \rho_{21} & 1 & \ldots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \ldots & 1 \end{pmatrix} = V^{-1} \Sigma V^{-1}$$

with $V = \text{diag}(\sigma_1, \ldots, \sigma_p)$.

It follows that $\Sigma = V \rho V$

Result 1. Linear combinations of random vectors
Consider X a p-dimensional random vector and $c \in \mathbb{R}^p$ then $c^t X$ is a one-dimensional random variable with

- $E[c^t X] = c^t \mu$
- $\mathrm{Var}[c^t X] = c^t \Sigma c$

In general, if $C \in \mathbb{R}^{q \times p}$ then CX is a q-dimensional random vector with

- $E[CX] = C\mu$
- $\mathrm{Cov}[CX] = C\Sigma C^t$