Revision: Matrix algebra

Vectors and matrices

As a shorthand notation, we shall be using $X \in \mathcal{M}_{p,n}$ to indicate that X is a matrix with p rows and n columns. A notation $\boldsymbol{x} \in \mathbb{R}^n$ will be used to indicate that $\boldsymbol{x}$ is a n-dimensional *column* vector. Of course, if $\boldsymbol{x} \in \mathbb{R}^n$, it also means that $\boldsymbol{x} \in \mathcal{M}_{n,1}$. *Transposition* will be denoted by $^\top$. After a transposition, from a matrix $X \in \mathcal{M}_{p,n}$ we get a new matrix $X^\top \in \mathcal{M}_{n,p}$. In particular, from a *column* vector $\boldsymbol{x} \in \mathbb{R}^n$ we arrive, after a transposition, to a *row* a vector $\boldsymbol{x}^\top \in \mathcal{M}_{1,n}$. It is well known that multiplication of a matrix (vector) with a scalar means multiplication of each of the elements of the matrix (vector) with that scalar. Also, two matrices (vectors) of the same dimension can be added (subtracted) and the result is a new matrix (vector) of the same dimension and elements which are the element wise sum (difference) of the elements of the matrices (vectors) to be added (subtracted).

The *Euclidean norm* of a vector $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \in \mathbb{R}^p$ is denoted by $\|\boldsymbol{x}\|$ and is defined as $\|\boldsymbol{x}\| = \sqrt{\sum_{i=1}^p x_i^2}$.

The *inner product* or, equivalently, the *scalar product* of two p-dimensional vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is denoted and defined in the following way:

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^\top \boldsymbol{y} = \sum_{i=1}^p x_i y_i \tag{0.1}$$

Obviously, the relation $\|\boldsymbol{x}\|^2 = \langle \boldsymbol{x}, \boldsymbol{x} \rangle$ holds. It is well known that if θ is the angle between two p-dimensional vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ then it also holds

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \|\boldsymbol{x}\| \|\boldsymbol{y}\| \cos(\theta) \tag{0.2}$$

Since $|\cos(\theta)| \leq 1$, we have the inequality

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \leq \|\boldsymbol{x}\| \|\boldsymbol{y}\|$$

which is one variant of the *Cauchy–Bunyakovsky–Schwartz* Inequality. Further, if we want to *orthogonally project* the vector $\boldsymbol{x} \in \mathbb{R}^p$ on the vector $\boldsymbol{y} \in \mathbb{R}^p$ then (having in mind the geometric interpretation of orthogonal projection) the result will be: $\frac{\boldsymbol{x}^\top \boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{y}}\boldsymbol{y}$.

Finally, the rules for matrix multiplication are recalled: if $X \in \mathcal{M}_{p,k}$ and $Y \in \mathcal{M}_{k,n}$ (i.e. the number of columns in X is equal to the number of rows in Y) then the multiplication XY is possible and the result is a matrix $Z = XY \in \mathcal{M}_{p,n}$ with elements

$$z_{i,j},\ i = 1, 2, \ldots, p,\ j = 1, 2, \ldots, n:\ z_{i,j} = \sum_{m=1}^{k} x_{i,m} y_{m,j} \tag{0.3}$$

i.e. the element in the ith row and jth column of Z is a scalar product of the ith row of X and the jth column of Y. Note that the multiplication of matrices is **not commutative** and in general, it is not necessary for YX to even exist when XY exists. When the matrices are both square (quadratic) of the same dimension p (i.e. both $X \in \mathcal{M}_{p,p}$ and $Y \in \mathcal{M}_{p,p}$) then both XY and YX will be defined but would in general **not** give rise to the same result. The following transposition rule is important to be mentioned (and easy to check): if $X \in \mathcal{M}_{p,k}$ and $Y \in \mathcal{M}_{k,n}$ then the product XY exists and it holds:

$$(XY)^\top = Y^\top X^\top \tag{0.4}$$

One should be very careful with transposition though in order to avoid silly mistakes. If $X \in \mathbb{R}^p$, for example, both $X^\top X$ and $XX^\top$ exist. While the former is a scalar, the latter belongs to $\mathcal{M}_{p,p}$!

A square matrix $X \in \mathcal{M}_{p,p}$ is called *symmetric* if $x_{i,j} = x_{j,i}$ for $i = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, p$ holds. For such a matrix, we have $X^\top = X$. The square matrix $\overset{p \times p}{\boldsymbol{I}} = \delta_{ij}$ for $i = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, p$ holds (i.e., ones on the diagonal and zeros outside the diagonal) is called the *identity matrix* (of dimension p). Obviously, when the multiplication is possible then always $X\boldsymbol{I} = X$ and $\boldsymbol{I}X = X$ holds.

The trace of a square matrix $X \in \mathcal{M}_{p,p}$ is denoted by $\mathrm{tr}(X) = \sum_{i=1}^{p} x_{ii}$. The following properties of traces are easy to obtain:

1. $\mathrm{tr}(X + Y) = \mathrm{tr}(X) + \mathrm{tr}(Y)$
2. $\mathrm{tr}(XY) = \mathrm{tr}(YX)$
3. $\mathrm{tr}(X^{-1}YX) = \mathrm{tr}(Y)$
4. If $\boldsymbol{a} \in \mathbb{R}^p$ and $X \in \mathcal{M}_{p,p}$ then $\boldsymbol{a}^\top X \boldsymbol{a} = \mathrm{tr}(X\boldsymbol{a}\boldsymbol{a}^\top)$

Inverse matrices

To any **square** matrix $X \in \mathcal{M}_{p,p}$ one can attach a number $|X| \equiv \det(X)$ called a *determinant* of the matrix. It is defined as

$$|X| = \sum \pm x_{1i} x_{2j} \dots x_{pm}$$

where the summation is over **all** permutations $(i, j, \dots, m)$ of the numbers $(1, 2, \dots, p)$ by taking into account the **sign rule**: summands with an even permutation get a $(+)$ whereas the ones with an odd permutation get a $(-)$ sign.

It can be seen that this is equivalent to another recursive definition, namely:

- when $p = 1$ (scalar case) $X = a$ is just a number and $|X| = a$ in this case
- when $p = 2$ then $\begin{vmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{vmatrix} = x_{11}x_{22} - x_{12}x_{21}$
- when $p = 3$ then the following rule applies:

$$\begin{vmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{vmatrix} = x_{11}x_{22}x_{33} + x_{12}x_{23}x_{31} + x_{21}x_{32}x_{13} - \\ x_{31}x_{22}x_{13} - x_{11}x_{23}x_{32} - x_{12}x_{21}x_{33} \tag{0.5}$$

- recursively, for $X \in \mathcal{M}_{(p,p)}$ we can define

$$|X| = \sum_i (-1)^{i+j} x_{ij} |X_{ij}| = \sum_j (-1)^{i+j} x_{ij} |X_{ij}|$$

  where X_{ij} denotes the matrix we get by deleting the ith row and jth column of X.

Here we list some elementary properties of determinants that follow directly from the definition:

1. If one row or one column of the matrix contains zeros only then the value of the determinant is zero.
2. $|X^\top| = |X|$
3. If one row (or one column) of the matrix is modified by multiplying with a scalar c then so is the value of the determinant.
4. $|cX| = c^p |X|$
5. If $X, Y \in \mathcal{M}_{p,p}$ then $|XY| = |X||Y|$
6. If the matrix X is *diagonal* (i.e. all non-diagonal elements are zero) then $|X| = \prod_{i=1}^{p} x_{ii}$. In

particular, *the determinant of the identity matrix is always equal to one*.

Given that $|X| \neq 0$ (or equivalently, if the matrix $X \in \mathcal{M}_{p,p}$ is *nonsingular* then an **inverse** matrix $X^{-1} \in \mathcal{M}_{p,p}$ can be defined that has to satisfy $XX^{-1} = \boldsymbol{I}_{p,p}$. It is easy to check that the inverse X^{-1} has as its (j,i)th entry $\frac{|X_{-ij}|}{|X|}(-1)^{i+j}$, where $|X_{-ij}|$ is the (i,j)th *minor* of X, the determinant of the matrix constructed by removing ith row and jth column.

Some elementary properties of inverses follow:

1. $XX^{-1} = X^{-1}X = \boldsymbol{I}$
2. $(X^{-1})^\top = (X^\top)^{-1}$
3. $(XY)^{-1} = Y^{-1}X^{-1}$ when both X and Y are nonsingular square matrices of the same dimension.
4. $|X^{-1}| = |X|^{-1}$
5. If X is diagonal and nonsingular then all its diagonal elements are nonzero and X^{-1} is again diagonal with diagonal elements equal to $\frac{1}{x_{ii}}, i = 1, 2, \ldots, p$.

Rank & orthogonal matrices

A set of vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k \in \mathbb{R}^n$ is *linearly dependent* if there exist k numbers $a_1, a_2, \ldots, a_k$ **not all zero** such that

$$a_1\boldsymbol{x}_1 + a_2\boldsymbol{x}_2 + \cdots + a_k\boldsymbol{x}_k = \boldsymbol{0} \tag{0.6}$$

holds. Otherwise the vectors are *linearly independent*. In particular, for k linearly independent vectors the equality (0.6) would only be possible if **all** numbers $a_1, a_2, \ldots, a_k$ were zero.

The *row rank* of a matrix is the maximum number of linearly independent row vectors. The *column rank* is the rank of its set of column vectors. It turns out that the row rank and the column rank of a matrix are always equal. Thus the rank of a matrix X (denoted $\mathrm{rk}(X)$) is either the row or the column rank. If $X \in \mathcal{M}_{p,n}$ and $\mathrm{rk}(X) = \min(p, n)$ we say that the matrix is of full rank. In particular, a square matrix $A \in \mathcal{M}_{p,p}$ is of full rank if $\mathrm{rk}(A) = p$. As is well known from the basic theorem of linear algebra *Kronecker–Capelli* or *Rouché–Capelli Theorem* this means also that $|A| \neq 0$ when A is of full rank. Then the inverse of A will also exist. Let $\boldsymbol{b} \in \mathbb{R}^p$ be a given vector. Then the linear equation system $A\boldsymbol{x} = \boldsymbol{b}$ has a unique solution $\boldsymbol{x} = A^{-1}\boldsymbol{b} \in \mathbb{R}^p$.

Orthogonal matrices

A square matrix $X \in \mathcal{M}_{p,p}$ is *orthogonal* if $XX^\top = X^\top X = \boldsymbol{I}_{p,p}$ holds. The following properties of orthogonal matrices are obvious:

1. X is of full rank ($\mathrm{rk}(X) = p$) and $X^{-1} = X^\top$
2. The name *orthogonal* of the matrix originates from the fact that the scalar product of each two different column vectors equals zero. The same holds for the scalar product of each two different row vectors of the matrix. The norm of each column vector (or each row vector) is equal to one. These properties are equivalent to the definition.
3. $|X| = \pm 1$

Eigenvalues and eigenvectors

For **any** square matrix $X \in \mathcal{M}_{p,p}$ we can define the *characteristic polynomial* equation of degree p,

$$f(\lambda) = |X - \lambda \boldsymbol{I}| = 0. \tag{0.7}$$

Equation (0.7) is a polynomial equation of power p so it has exactly p roots. In general, some of them may be complex and some may coincide. Since the coefficients are real, if there is a complex root of (0.7) then also its complex conjugate must be a root of the same equation. Denote **any** such eigenvalue by λ^*. In addition, $\operatorname{tr}(X) = \sum_{i=1}^p \lambda_i$ and $|X| = \prod_{i=1}^p \lambda_i$.

Obviously, the matrix $X - \lambda^* \boldsymbol{I}$ is singular (its determinant is zero). Then, according to the Kronecker theorem, there exists a non-zero vector $\boldsymbol{y} \in \mathbb{R}^p$ such that $(X - \lambda^* \boldsymbol{I})\boldsymbol{y} = \boldsymbol{0}, \boldsymbol{0} \in \mathbb{R}^p$. We call $\boldsymbol{y}$ an *eigenvector* of X that corresponds to the eigenvalue λ^*. Note that the eigenvector is not uniquely defined: $\mu \boldsymbol{y}$ for any real non-zero μ would also be an eigenvector corresponding to the same eigenvalue.

Sparing some details of the derivation, we shall formulate the following basic result:

Theorem 0.1. *When the matrix X is real symmetric then **all** of its p eigenvalues are **real**. If the eigenvalues are all different then all the p eigenvectors that correspond to them, are orthogonal (and hence form a basis in $\mathbb{R}^p$). These eigenvectors are also unique (up to the norming constant μ above). If some of the eigenvalues coincide then the eigenvectors corresponding to them are not necessarily unique but even in this case they can be chosen to be mutually orthogonal.*

For each of the p eigenvalues λ_i, $i = 1, 2, \ldots, p$, of X, denote its corresponding set of mutually orthogonal eigenvectors of *unit length* by $\boldsymbol{e}_i$, $i = 1, 2, \ldots, p$, i.e.

$$X\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i,\ i = 1, 2, \ldots, p,\ \|\boldsymbol{e}_i\| = 1,\ \boldsymbol{e}_i^\top \boldsymbol{e}_j = 0,\ i \neq j$$

holds. Then is can be shown that the following decomposition *(spectral decomposition)* of any symmetric matrix $X \in \mathcal{M}_{p,p}$ holds:

$$X = \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1^\top + \lambda_2 \boldsymbol{e}_2 \boldsymbol{e}_2^\top + \ldots \lambda_p \boldsymbol{e}_p \boldsymbol{e}_p^\top. \tag{0.8}$$

Equivalently, $X = P \Lambda P^\top$ where $\Lambda = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p \end{pmatrix}$ is diagonal and $P \in \mathcal{M}_{p,p}$ is an *orthogonal matrix* containing the p orthogonal eigenvectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_p$.

The above decomposition is a very important analytical tool. One of its most widely used applications is for defining a square root of a symmetric positive definite matrix. A symmetric matrix $X \in \mathcal{M}_{p,p}$ is *positive definite* if all of its eigenvalues are positive (it is called *non-negative definite* if all eigenvalues are ≥ 0). For a symmetric positive definite matrix we have all λ_i, $i = 1, 2, \ldots, p$, to be positive in the spectral decomposition (0.8).

But then

$$X^{-1} = (P^\top)^{-1}\Lambda^{-1}P^{-1} = P\Lambda^{-1}P^\top = \sum_{i=1}^{p} \frac{1}{\lambda_i}\boldsymbol{e}_i\boldsymbol{e}_i^\top \tag{0.9}$$

(i.e. inverting X is very easy if the spectral decomposition of X is known).

Moreover we can define the *square root* of the symmetric non-negative definite matrix X in a natural way:

$$X^{\frac{1}{2}} = \sum_{i=1}^{p} \sqrt{\lambda_i}\boldsymbol{e}_i\boldsymbol{e}_i^\top \tag{0.10}$$

The definition (0.10) makes sense since $X^{\frac{1}{2}}X^{\frac{1}{2}} = X$ holds. Note that $X^{\frac{1}{2}}$ is also symmetric and non-negative definite. Also $X^{-\frac{1}{2}} = \sum_{i=1}^{p} \lambda_i^{-\frac{1}{2}} e_i e_i^\top = P\Lambda^{-\frac{1}{2}}P^\top$ can be defined where $\Lambda^{-\frac{1}{2}}$ is a diagonal matrix with $\lambda_i^{-1/2}, i = 1, 2, \ldots, p$ being its diagonal elements. These facts will be used essentially in the subsequent sections.

As an illustration of the usefulness of the spectral decomposition approach we shall show the following statement:

Example 0.1. Let $X \in \mathcal{M}_{p,p}$ be symmetric *positive definite* matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ and associated eigenvectors of unit length $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots \boldsymbol{e}_p$. Show that

- $\max_{\boldsymbol{y}\neq\boldsymbol{0}} \frac{\boldsymbol{y}^\top X\boldsymbol{y}}{\boldsymbol{y}^\top\boldsymbol{y}} = \lambda_1$ attained when $\boldsymbol{y} = \boldsymbol{e}_1$.
- $\min_{\boldsymbol{y}\neq\boldsymbol{0}} \frac{\boldsymbol{y}^\top X\boldsymbol{y}}{\boldsymbol{y}^\top\boldsymbol{y}} = \lambda_p$ attained when $\boldsymbol{y} = \boldsymbol{e}_p$.

Let $X = P\Lambda P^\top$ be the decomposition (0.8) for X. Denote $\boldsymbol{z} = P^\top\boldsymbol{y}$. Note that $\boldsymbol{y} \neq \boldsymbol{0}$ implies $\boldsymbol{z} \neq \boldsymbol{0}$. Thus

$$\frac{\boldsymbol{y}^\top X\boldsymbol{y}}{\boldsymbol{y}^\top\boldsymbol{y}} = \frac{\boldsymbol{y}^\top P\Lambda P^\top\boldsymbol{y}}{\boldsymbol{y}^\top\boldsymbol{y}} = \frac{\boldsymbol{z}^\top\Lambda\boldsymbol{z}}{\boldsymbol{z}^\top\boldsymbol{z}} = \frac{\sum_{i=1}^p \lambda_i z_i^2}{\sum_{i=1}^p z_i^2} \leq \lambda_1\frac{\sum_{i=1}^p z_i^2}{\sum_{i=1}^p z_i^2} = \lambda_1$$

If we take $\boldsymbol{y} = \boldsymbol{e}_1$ then having in mind the structure of the matrix P we have $\boldsymbol{z} = P^\top\boldsymbol{e}_1 = \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix}^\top$ and for this choice of $\boldsymbol{y}$ also $\frac{\boldsymbol{z}^\top\Lambda\boldsymbol{z}}{\boldsymbol{z}^\top\boldsymbol{z}} = \frac{\lambda_1}{1} = \lambda_1$. The first part of the exercise is shown. Similar arguments (just changing the sign of the inequality) apply to show the second part.

In addition, you can try to show that $\max_{\boldsymbol{y}\neq 0, \boldsymbol{y}\perp e_1} \frac{\boldsymbol{y}^\top X\boldsymbol{y}}{\boldsymbol{y}^\top\boldsymbol{y}} = \lambda_2$ holds. How?

Numerical stability and Cholesky decomposition

Computers perform arithmetic to a finite precision, typically around 16 decimal significant figures. Furthermore, the numbers are expressed internally in scientific notation, and so the absolute magnitude of the number typically has little effect on precision, but certain operations on numbers with very different magnitudes can sometimes produce severe rounding errors. For example, to a computer $1 \times 10^{18} + 1 \times 10^{0} = 1{,}000{,}000{,}000{,}000{,}000{,}000 + 1 = 1{,}000{,}000{,}000{,}000{,}000{,}000$: the 1 gets lost to a rounding error. When it comes to matrix inversion in particular, the key number is the *condition number*, $|\lambda_1/\lambda_p|$ of a positive definite matrix X, where λ_1 is the largest eigenvalue of X and λ_p is the smallest. (The definition for non-positive-definite matrices can be different.) The higher this number is, the less numerically stable the inversion is likely to be. (Notice that if the matrix is singular, this number is infinite.) We generally try to avoid asking the computer to invert matrices in ways that lose precision.

An alternative, more numerically stable definition of a "matrix square root" is the *Cholesky decomposition*. For a symmetric positive definite matrix $X \in \mathcal{M}_{p,p}$, there exists a unique upper-triangular matrix $U \in \mathcal{M}_{p,p}$ such that $U^\top U = X$ holds. Note that many sources use a lower-triangular matrix L such that $LL^\top = X$ instead. It is easy to see that $L \equiv U^\top$, and which definition is used is arbitrary, provided it is used consistently, since $UU^\top \neq X$ and neither do $L^\top L$. For example, the Wikipedia article uses L, whereas the R builtin function is `chol()` returns U. This decomposition is particularly useful for generating correlated variables.

Standard facts about multivariate distributions

Random samples in multivariate analysis

In order to study the sampling variability of statistics like $\bar{x}$ and S_n that we introduced in Lecture 1, with the ultimate goal of making inferences, one needs to make some assumptions about the random variables whose values constitute the dataset $X \in \mathcal{M}_{p,n}$ in (1.1). Suppose the data has not been observed yet but we *intend* to collect n sets of measurements on p variables. Since the actual observations can not be predicted before the measurements are made, we treat them as random variables. Each set of p measurements can be considered as a realisation of p-dimensional *random vector* and we have n independent realisations of such random vectors $\boldsymbol{X}_i, i = 1, 2, \ldots, n$, so we have the *random matrix* $\boldsymbol{X} \in \mathcal{M}_{p,n}$:

$$\boldsymbol{X} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1j} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2j} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{i1} & X_{i2} & \cdots & X_{ij} & \cdots & X_{in} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{p1} & X_{p2} & \cdots & X_{pj} & \cdots & X_{pn} \end{pmatrix} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n] \tag{0.11}$$

The vectors $\boldsymbol{X}_i, i = 1, 2, \ldots, n$ are considered as independent observations of a p-dimensional random vector. We start discussing the distribution of such a vector.

Further reading

Johnson and Wichern:

- 2.1–2.5

Härdle and Simar:

- 2.1–2.2