Linear algebra warmup and basics of random vectors

Let R^n denote the n-dimensional Euclidean space, which is a vector space endowed with the usual inner product. Vectors are column-vectors. The inner product of the vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ is therefore written with matrix multiplication, like $\mathbf{x}^T\mathbf{y}$, where ^{T} stands for the transposition, hence $\mathbf{x}^T$ is a row-vector. Matrices will be denoted by bold-face upper-case letters. An $m \times n$ matrix $\mathbf{A} = (a_{ij})$ of real entries a_{ij}'s corresponds to an $\mathbb{R}^n \to \mathbb{R}^m$ linear transformation (operator). Its transpose, $\mathbf{A}^T$, is an $n \times m$ matrix. An $n \times n$ matrix is called quadratic and it maps $\mathbb{R}^n$ into itself. The identity matrix is denoted by $\mathbf{I}$ or $\mathbf{I}_n$ if we want to refer to its size.

The quadratic matrix $\mathbf{A}$ is *symmetric* if $\mathbf{A} = \mathbf{A}^T$ and *orthogonal* (rotation) if $\mathbf{A}\mathbf{A}^T = \mathbf{I}$. The quadratic matrix $\mathbf{P}$ corresponds to an *orthogonal projection* if it is symmetric and idempotent: $\mathbf{P}^2 = \mathbf{P}$.

The $n \times n$ matrix $\mathbf{A}$ has an inverse if and only if its determinant, $|\mathbf{A}| \neq 0$, and its inverse is denoted by $\mathbf{A}^{-1}$. In this case, the linear transformation corresponding to $\mathbf{A}^{-1}$ undoes the effect of the $\mathbb{R}^n \to \mathbb{R}^n$ transformation corresponding to $\mathbf{A}$, i.e. $\mathbf{A}^{-1}\mathbf{y} = \mathbf{x}$ if and only if $\mathbf{A}\mathbf{x} = \mathbf{y}$ for any $\mathbf{y} \in \mathbb{R}^n$. It is important that in case of an invertible (*regular*) matrix $\mathbf{A}$, the *range* (or image space) of $\mathbf{A}$ – denoted by $\mathcal{R}(\mathbf{A})$ – is the whole $\mathbb{R}^n$, and in exchange, the kernel of $\mathbf{A}$ (the subspace of vectors that are mapped into the zero vector by $\mathbf{A}$) consists of the only $\mathbf{0}$.

Note that for an $m \times n$ matrix $\mathbf{A}$, its range is

$$\mathcal{R}(\mathbf{A}) = \mathrm{Span}\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$$

where $\mathbf{a}_1, \ldots, \mathbf{a}_n$ are the column vectors of $\mathbf{A}$ for which fact the notation $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)$ will be used; further, $\mathrm{Span}\{\ldots\}$ is the subspace spanned by the vectors in its argument. The *rank* of $\mathbf{A}$ is the dimension of its range:

$$\mathrm{rank}(\mathbf{A}) = \mathrm{dim}\mathcal{R}(\mathbf{A}),$$

and it is also equal to the maximum number of linearly independent rows of $\mathbf{A}$; trivially, $\mathrm{rank}(\mathbf{A}) \leq \min\{m, n\}$. In case of $m = n$, $\mathbf{A}$ is regular if and only if $\mathrm{rank}(\mathbf{A}) = n$, and *singular*, otherwise.

An orthogonal matrix $\mathbf{A}$ is always regular and $\mathbf{A}^{-1} = \mathbf{A}^T$; further its rows (or columns) constitute a complete orthonormal set in $\mathbb{R}^n$. Let k $(1 \leq k < n)$ be an integer; an $n \times k$ matrix $\mathbf{A}$ is called *suborthogonal* if its columns form (a not complete) orthonormal set in $\mathbb{R}^n$. For such an $\mathbf{A}$, the relation $\mathbf{A}^T\mathbf{A} = \mathbf{I}_k$ holds, but $\mathbf{A}\mathbf{A}^T \neq \mathbf{I}_n$. In fact, the $n \times n$ matrix $\mathbf{P} = \mathbf{A}\mathbf{A}^T$ is symmetric and idempotent ($\mathbf{P}^2 = \mathbf{P}$), hence, it corresponds to the orthogonal projection onto $\mathcal{R}(\mathbf{A})$. The *trace* of the $n \times n$ matrix $\mathbf{A}$ is

$$\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}.$$

How the above matrix–matrix and matrix–scalar functions will look like if the underlying matrix is a product? If $\mathbf{A}$ and $\mathbf{B}$ can be multiplied together ($\mathbf{A}$ is $m \times n$ and $\mathbf{B}$ is $n \times k$ type), then their product corresponds to the succession of linear operations $\mathbf{B}$ and $\mathbf{A}$ in this order, therefore

$$(\mathbf{A}\mathbf{B})^T = \mathbf{B}^T\mathbf{A}^T$$

and if $\mathbf{A}$ and $\mathbf{B}$ are regular $n \times n$ matrices, then so is $\mathbf{AB}$, and

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}.$$

Further, $(\mathbf{A}^{-1})^T = (\mathbf{A}^T)^{-1}$, and vice versa. If $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ matrices, then

$$|\mathbf{AB}| = |\mathbf{A}| \cdot |\mathbf{B}|.$$

Therefore, the determinant of the product of several matrices of the same size does not depend on the succession of the matrices, however, the matrix multiplication is usually not commutative. The trace is commutative in the following sense: if $\mathbf{A}$ is an $n \times k$ and $\mathbf{B}$ is a $k \times n$ matrix, then

$$\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA}).$$

For several factors, the trace is accordingly, cyclically commutative:

$$\operatorname{tr}(\mathbf{A}_1\mathbf{A}_2 \dots \mathbf{A}_n) = \operatorname{tr}(\mathbf{A}_2 \dots \mathbf{A}_n\mathbf{A}_1) = \dots = \operatorname{tr}(\mathbf{A}_n\mathbf{A}_1 \dots \mathbf{A}_{n-1})$$

when, of course, the sizes of the factors are such that the successive multiplications in $\mathbf{A}_1 \dots \mathbf{A}_n$ can be performed and the number of rows in $\mathbf{A}_1$ is equal to the number of columns in $\mathbf{A}_n$. Further,

$$\operatorname{rank}(\mathbf{AB}) \leq \min\{\operatorname{rank}(\mathbf{A}), \operatorname{rank}(\mathbf{B})\},$$

consequently, the rank cannot be increased in course of matrix multiplications.

Given an $n \times n$ symmetric real matrix $\mathbf{A}$, the quadratic form in the variables $x_1, \dots, x_n$ is the homogeneous quadratic function of these variables:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j = \mathbf{x}^T\mathbf{A}\mathbf{x},$$

where $\mathbf{x} = (x_1, \dots, x_n)^T$, hence the matrix multiplication results in a scalar. The possible signs of a quadratic form (with different $\mathbf{x}$'s) characterize the underlying matrix. Accordingly, they fall into exactly one of the following categories.

Definition 1 *Let* $\mathbf{A}$ *be* $n \times n$ *symmetric real matrix.*

- $\mathbf{A}$ *is positive (negative) definite if* $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$ *(*$\mathbf{x}^T\mathbf{A}\mathbf{x} < 0$*),* $\forall \mathbf{x} \neq \mathbf{0}$*.*
- $\mathbf{A}$ *is positive (negative) semidefinite if* $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$ *(*$\mathbf{x}^T\mathbf{A}\mathbf{x} \leq 0$*),* $\forall \mathbf{x} \in \mathbb{R}^n$*, and* $\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{0}$ *for at least one* $\mathbf{x} \neq \mathbf{0}$*.*
- $\mathbf{A}$ *is indefinite if* $\mathbf{x}^T\mathbf{A}\mathbf{x}$ *takes on both positive and negative values (with different, non-zero* $\mathbf{x}$*'s).*

The positive and negative definite matrices are all regular, whereas the positive and negative semidefinite ones are singular. The indefinite matrices can be either regular or singular. To more easily characterize the definiteness of symmetric matrices, we will use their eigenvalues.

The notion of an eigenvalue and eigenvector is introduced: λ is an eigenvalue of the $n \times n$ real matrix $\mathbf{A}$ with corresponding eigenvector $\mathbf{u} \neq \mathbf{0}$ if $\mathbf{A}\mathbf{u} = \lambda\mathbf{u}$. If $\mathbf{u}$ is an eigenvector of $\mathbf{A}$, it is easy to see that for $c \neq 0$, $c\mathbf{u}$ is also an eigenvector

with the same eigenvalue. Therefore, it is better to speak about *eigen-directions* instead of eigenvectors; or else, we will consider specially normalized, e.g. unit-norm eigenvectors, when only the orientation is divalent. It is well known that an $n \times n$ matrix $\mathbf{A}$ has exactly n eigenvalues (with multiplicities) which are (possibly complex) roots of the characteristic polynomial $|\mathbf{A} - \lambda\mathbf{I}|$. Knowing the eigenvalues, the corresponding eigenvectors are obtained by solving the system of linear equations $(\mathbf{A} - \lambda\mathbf{I})\mathbf{u} = \mathbf{0}$ which must have a non-trivial solution due to the choice of λ. In fact, there are infinitely many solutions (in case of single eigenvalues they are constant multiples of each other). An eigenvector corresponding to a complex eigenvalue must also have complex coordinates, but in case of our main interest (the symmetric matrices) this cannot occur.

The notion of an eigenvalue and eigenvector extends to matrices of complex entries in the same way. As for the allocation of the eigenvalues of a quadratic matrix (even of complex entries), the following result is known.

Theorem 1 (Gersgorin disc theorem) *Let $\mathbf{A}$ be an $n \times n$ matrix of entries $a_{ij} \in \mathbb{C}$. The Gersgorin disks of $\mathbf{A}$ are the following regions of the complex plane:*

$$D_i = \{z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j \neq i} |a_{ij}|\}, \quad i = 1, \ldots n.$$

Let $\lambda_1, \ldots, \lambda_n$ denote the (possibly complex) eigenvalues of $\mathbf{A}$. Then

$$\{\lambda_1, \ldots, \lambda_n\} \subset \cup_{i=1}^{n} D_i.$$

Furthermore, any connected component of the set $\cup_{i=1}^{n} D_i$ contains as many eigenvalues of $\mathbf{A}$ as the number of discs that form this component.

We will introduce the notion of *normal matrices* which admit a spectral decomposition (briefly, SD) similar to that of compact operators. The real matrix $\mathbf{A}$ is called normal if $\mathbf{A}\mathbf{A}^T = \mathbf{A}^T\mathbf{A}$. Among real matrices, only the symmetric, anti-symmetric ($\mathbf{A}^T = -\mathbf{A}$), and orthogonal matrices are normal. Normal matrices have the following important spectral property: to their eigenvalues there corresponds an orthonormal set of eigenvectors; choosing this as a new basis, the matrix becomes *diagonal* (all the off-diagonal entries are zeros). Here we state the Hilbert–Schmidt theorem for symmetric matrices which, in addition, have all real eigenvalues, and consequently, eigenvectors of real coordinates.

Theorem 2 *The $n \times n$ symmetric, real matrix $\mathbf{A}$ has real eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ (with multiplicities), and the corresponding eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ can be chosen such that they constitute a complete orthonormal set in $\mathbb{R}^n$.*

This so-called Spectral Decomposition theorem implies the following SD of the $n \times n$ symmetric matrix $\mathbf{A}$:

$$\mathbf{A} = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i^T = \mathbf{U}\underline{\Lambda}\mathbf{U}^T, \tag{1}$$

where $\underline{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ is the diagonal matrix containing the eigenvalues – called *spectrum* – in its main diagonal, while $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_n)$ is the orthogonal matrix containing the corresponding eigenvectors of $\mathbf{A}$ in its columns in the order of the eigenvalues. Of course, permuting the eigenvalues in the main

diagonal of $\underline{\Lambda}$, and the columns of $\mathbf{U}$ accordingly, will lead to the same SD, however – if not otherwise stated – we will enumerate the real eigenvalues in non-increasing order. About the uniqueness of the above SD we can state the following: the unit-norm eigenvector corresponding to a single eigenvalue is unique (up to orientation), whereas to an eigenvalue with multiplicity m there corresponds a unique m-dimensional so-called *eigen-subspace* within which any orthonormal set can be chosen for the corresponding eigenvectors.

It is easy to verify that for the eigenvalues of the symmetric matrix $\mathbf{A}$

$$\sum_{i=1}^{n} \lambda_i = \mathrm{tr}(\mathbf{A}) \quad \text{and} \quad \prod_{i=1}^{n} \lambda_i = |\mathbf{A}|$$

hold. Therefore $\mathbf{A}$ is singular if and only if it has a 0 eigenvalue, and

$$r = \mathrm{rank}(\mathbf{A}) = \mathrm{rank}(\underline{\Lambda}) = |\{i \,:\, \lambda_i \neq 0\}|;$$

moreover, $\mathcal{R}(\mathbf{A}) = \mathrm{Span}\{\mathbf{u}_i \,:\, \lambda_i \neq 0\}$. Therefore, the SD of $\mathbf{A}$ simplifies to

$$\sum_{\lambda_i \neq 0} \lambda_i \mathbf{u}_i \mathbf{u}_i^T.$$

Its spectrum also determines the definiteness of $\mathbf{A}$ in the following manner.

Proposition 1 *Let $\mathbf{A}$ be $n \times n$ symmetric real matrix.*

- $\mathbf{A}$ *is positive (negative) definite if and only if all of its eigenvalues are positive (negative).*
- $\mathbf{A}$ *is positive (negative) semidefinite if and only if all of its eigenvalues are nonnegative (nonpositive), and its spectrum includes the zero.*
- $\mathbf{A}$ *is indefinite if its spectrum contains at least one positive and one negative eigenvalue.*

The matrix of an orthogonal projection $\mathbf{P}_F$ onto the r-dimensional subspace $F \subset \mathbb{R}^n$ has the following SD (only the $r < n$ case is of importance, since in the $r = n$ case $\mathbf{P}_F = \mathbf{I}_n$):

$$\mathbf{P}_F = \sum_{i=1}^{r} \mathbf{u}_i \mathbf{u}_i^T = \mathbf{A}\mathbf{A}^T,$$

where $\mathbf{u}_1, \ldots, \mathbf{u}_r$ is any orthonormal set in F which is the eigen-subspace corresponding to the eigenvalue 1 of multiplicity r. Note that the eigenspace corresponding to the other eigenvalue 0 of multiplicity $n - r$ is the orthogonal complementary subspace $F^\perp$ of F in $\mathbb{R}^n$, but it has no importance, as only the eigenvectors in the first r columns of $\mathbf{U}$ enter into the above SD of $\mathbf{P}_F$. With the notation $\mathbf{A} = (\mathbf{u}_1, \ldots, \mathbf{u}_r)$, the SD of $\mathbf{P}_F$ simplifies to $\mathbf{A}\mathbf{A}^T$, indicating that $\mathbf{A}$ is a suborthogonal matrix.

For rectangular matrices the following can be stated.

Theorem 3 *Let $\mathbf{A}$ be an $m \times n$ rectangular matrix of real entries,* $\mathrm{rank}(\mathbf{A}) = r \leq \min\{m, n\}$. *Then there exist an orthonormal set $(\mathbf{v}_1, \ldots, \mathbf{v}_r) \subset \mathbb{R}^m$ and $(\mathbf{u}_1, \ldots, \mathbf{u}_r) \subset \mathbb{R}^n$ together with the positive real numbers $s_1 \geq s_2 \geq \cdots \geq s_r > 0$ such that*

$$\mathbf{A}\mathbf{u}_i = s_i \mathbf{v}_i, \quad \mathbf{A}^T \mathbf{v}_i = s_i \mathbf{u}_i, \quad i = 1, 2, \ldots, r. \tag{2}$$

The elements $\mathbf{v}_i \in \mathbb{R}^m$ and $\mathbf{u}_i \in \mathbb{R}^n$ $(i = 1, \ldots, r)$ in (2) are called *relevant singular vector pairs* (or left and right singular vectors) corresponding to the *singular value* s_i $(i = 1, 2, \ldots, r)$. The transformations in (2) give a one-to one mapping between $\mathcal{R}(\mathbf{A})$ and $\mathcal{R}(\mathbf{A}^T)$, all the other vectors of $\mathbb{R}^n$ and $\mathbb{R}^m$ are mapped into the zero vector of $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively. However, the left and right singular vectors can appropriately be completed into a complete orthonormal set $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\} \subset \mathbb{R}^m$ and $\{\mathbf{u}_1, \ldots \mathbf{u}_n\} \subset \mathbb{R}^n$, respectively, such that, the so introduced extra vectors in the kernel subspaces in $\mathbb{R}^m$ and $\mathbb{R}^n$ are mapped into the zero vector of $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. With the orthogonal matrices $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_m)$ and $\mathbf{U} = (\mathbf{u}_1, \ldots \mathbf{u}_n)$, the following SVD of $\mathbf{A}$ and $\mathbf{A}^T$ holds:

$$\mathbf{A} = \mathbf{V}\mathbf{S}\mathbf{U}^T = \sum_{i=1}^{r} s_i \mathbf{v}_i \mathbf{u}_i^T \quad \text{and} \quad \mathbf{A}^T = \mathbf{U}\mathbf{S}^T\mathbf{V}^T = \sum_{i=1}^{r} s_i \mathbf{u}_i \mathbf{v}_i^T, \tag{3}$$

where $\mathbf{S}$ is an $m \times n$ so-called *generalized diagonal matrix* which contains the singular values $s_1, \ldots, s_r$ in the first r positions of its main diagonal (starting from the upper left corner) and zeros otherwise. We remark that there are other equivalent forms of the above SVD depending on, whether $m < n$ or $m \geq n$. For example, in the $m < n$ case, $\mathbf{V}$ can be an $m \times m$ orthogonal, $\mathbf{S}$ an $m \times m$ diagonal, and $\mathbf{U}$ an $n \times m$ suborthogonal matrix with the same relevant entries. About the uniqueness of the SVD the following can be stated: to a single positive singular value there corresponds a unique singular vector pair (of course, the orientation of the left and right singular vectors can be changed at the same time). To a positive singular value of multiplicity say $k > 1$ a k-dimensional left and right so-called *isotropic subspace* corresponds, within which, any k-element orthonormal sets can embody the left and right singular vectors with orientation such that the requirements in (2) are met.

We also remark that the singular values of a symmetric matrix are the absolute values of its eigenvalues. In case of a positive eigenvalue, the left and right singular vectors are the same (they coincide with the corresponding eigenvector with any, but the same orientation). In case of a negative eigenvalue, the left and right side singular vectors are opposite (any of them is the corresponding eigenvector which have a divalent orientation). In case of a zero singular value the orientation is immaterial, as it does not contribute to the SVD of the underlying matrix.

Assume that the $m \times n$ matrix $\mathbf{A}$ of rank r has SVD (3). It is easy to see that the matrices $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$ are positive semidefinite (possibly, positive definite) matrices of rank r, and their SD is

$$\mathbf{A}\mathbf{A}^T = \mathbf{V}(\mathbf{S}\mathbf{S}^T)\mathbf{V}^T = \sum_{i=1}^{r} s_i^2 \mathbf{v}_i \mathbf{v}_i^T \quad \text{and} \quad \mathbf{A}^T\mathbf{A} = \mathbf{U}(\mathbf{S}^T\mathbf{S})\mathbf{U}^T = \sum_{i=1}^{r} s_i^2 \mathbf{u}_i \mathbf{u}_i^T$$

where the diagonal matrices $\mathbf{S}\mathbf{S}^T$ and $\mathbf{S}^T\mathbf{S}$ both contain the numbers $s_1^2, \ldots, s_r^2$ in the leading positions of their main diagonals as non-zero eigenvalues.

These facts together also imply that the only positive singular value of a suborthogonal matrix is the 1 with multiplicity of its rank.

Definition 2 *We say that the $n \times n$ symmetric matrix $\mathbf{G} = (g_{ij})$ is a Gram-matrix if its entries are inner products: there is a dimension $d > 0$ and vectors*

$\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ *such that*

$$g_{ij} = \mathbf{x}_i^T \mathbf{x}_j, \quad i, j = 1, \ldots n.$$

Proposition 2 *The symmetric matrix* $\mathbf{G}$ *is a Gram-matrix if and only if it is positive semidefinite or positive definite.*

Proof (we give the proof, since its construction will later be used in some multivariate methods). If $\mathbf{G}$ is a Gram-matrix, then it can be decomposed as $\mathbf{G} = \mathbf{A}\mathbf{A}^T$, where $\mathbf{A}^T = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$. With this,

$$\mathbf{x}^T \mathbf{G} \mathbf{x} = \mathbf{x}^T \mathbf{A}\mathbf{A}^T \mathbf{x} = (\mathbf{A}^T \mathbf{x})^T (\mathbf{A}^T \mathbf{x}) = \|\mathbf{A}^T \mathbf{x}\|^2 \geq 0, \quad \forall \mathbf{x} \in \mathbb{R}^n.$$

Conversely, if $\mathbf{G}$ is positive semidefinite (or positive definite) with rank $r \leq n$, then its SD – using (1) – can be written as

$$\mathbf{G} = \sum_{i=1}^{r} \lambda_i \mathbf{u}_i \mathbf{u}_i^T.$$

Let the $n \times r$ matrix $\mathbf{A}$ be defined as

$$\mathbf{A} = (\sqrt{\lambda_1}\mathbf{u}_1, \ldots, \sqrt{\lambda_r}\mathbf{u}_r). \tag{4}$$

Then the row vectors of the matrix $\mathbf{A}$ will be r-dimensional vectors reproducing $\mathbf{G}$. Of course, such a decomposition is not unique: first of all, instead of $\mathbf{A}$ the matrix $\mathbf{A}\mathbf{Q}$ will also do, where $\mathbf{Q}$ is an arbitrary $r \times r$ orthogonal matrix (obviously, $\mathbf{x}_i$'s can be rotated); and $\mathbf{x}_i$'s can also be put in a higher ($d > r$) dimension with attaching any (but the same) number of zero coordinates to them.

The spectral norm (operator norm) of an $m \times n$ real matrix $\mathbf{A}$ of rank r, with positive singular values $s_1 \geq \cdots \geq s_r > 0$, is

$$\|\mathbf{A}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\| = s_1,$$

and its *Frobenius norm*, denoted by $\|.\|_2$, is

$$\|\mathbf{A}\|_2 = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2 \right)^{1/2} = \sqrt{\mathrm{tr}(\mathbf{A}\mathbf{A}^T)} = \sqrt{\mathrm{tr}(\mathbf{A}^T\mathbf{A})} = \left(\sum_{i=1}^{r} s_i^2 \right)^{1/2}.$$

The Frobenius norm is sometimes called Euclidean norm and corresponds to the Hilbert–Schmidt norm of operators between separable Hilbert spaces. For a symmetric real matrix $\mathbf{A}$,

$$\|\mathbf{A}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\| = \max_i |\lambda_i| \quad \text{and} \quad \|\mathbf{A}\|_2 = \left(\sum_{i=1}^{r} \lambda_i^2 \right)^{1/2}.$$

Obviously, for a real matrix $\mathbf{A}$ of rank r,

$$\|\mathbf{A}\| \leq \|\mathbf{A}\|_2 \leq \sqrt{r}\|\mathbf{A}\|. \tag{5}$$

More generally, a matrix norm is called *unitary invariant* if

$$\|\mathbf{A}\|_{\text{un}} = \|\mathbf{QAR}\|_{\text{un}}$$

with any $m \times m$ and $n \times n$ orthogonal matrices $\mathbf{Q}$ and $\mathbf{R}$, respectively. It is easy to see that a unitary invariant norm of a real matrix merely depends on its singular values (or eigenvalues if it is symmetric). The spectral and Frobenius norms are unitary invariant.

By means of SD or SVD we are able to define so-called *generalized inverses* of singular square or rectangular matrices: in fact, any matrix that undoes the effect of the underlying linear transformation between the ranges of $\mathbf{A}^T$ and $\mathbf{A}$ will do. A generalized inverse is far not unique as any transformation operating on the kernels can be added. However, the following *Moore–Penrose inverse* is uniquely defined and it coincides with the usual inverse if exists.

Definition 3 *The $m \times n$ matrix $\mathbf{X}$ is a generalized inverse of the $n \times m$ matrix $\mathbf{A}$ if $\mathbf{AXA} = \mathbf{A}$.*

A generalized inverse $\mathbf{A}$ satisfying $\mathbf{AXA} = \mathbf{A}$ is denoted by $\mathbf{A}^-$. In fact, any matrix that undoes the effect of the underlying linear transformation between the ranges of $\mathbf{A}^T$ and $\mathbf{A}$ will do. A generalized inverse is far not unique as any transformation operating on the kernels can be added. However, the following *pseudoinverse (Moore–Penrose inverse)* is unique and, in case of a quadratic matrix, it coincides with the usual inverse if exists.

Definition 4 *The $m \times n$ matrix $\mathbf{X}$ is the pseudoinverse (in other words, the Moore–Penrose inverse) of the $n \times m$ matrix $\mathbf{A}$ if it satisfies all of the following conditions:*

$$\begin{aligned}
\mathbf{AXA} &= \mathbf{A}, \\
\mathbf{XAX} &= \mathbf{X}, \\
(\mathbf{AX})^T &= \mathbf{AX}, \\
(\mathbf{XA})^T &= \mathbf{XA}.
\end{aligned}$$

It can be proved that there uniquely exists a pseudoinverse satisfying the conditions in the above definition, and it is denoted by $\mathbf{A}^+$.

Definition 5 *The Moore–Penrose inverse of the $n \times n$ symmetric matrix with SD (1) is*

$$\mathbf{A}^+ = \sum_{i=1}^{r} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T = \mathbf{U} \underline{\Lambda}^+ \mathbf{U}^T,$$

where $\underline{\Lambda}^+ = \text{diag}(\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_r}, 0, \ldots, 0)$ is the diagonal matrix containing the reciprocals of the non-zero eigenvalues, otherwise zeros, in its main diagonal.

The Moore–Penrose inverse of the $m \times n$ real matrix is the $n \times m$ matrix $\mathbf{A}^+$ with SVD (3)

$$\mathbf{A}^+ = \sum_{i=1}^{r} \frac{1}{s_i} \mathbf{u}_i \mathbf{v}_i^T = \mathbf{U} \mathbf{S}^+ \mathbf{V}^T,$$

where $\mathbf{S}^+$ is $n \times m$ generalized diagonal matrix containing the reciprocals of the non-zero singular values of $\mathbf{A}$ in the leading positions, otherwise zeros, in its main diagonal.

Note that, analogously, any analytic function f of the symmetric real matrix $\mathbf{A}$ can be defined by its SD, $\mathbf{A} = \mathbf{U}\underline{\Lambda}\mathbf{U}^T$, in the following way:

$$f(\mathbf{A}) := \mathbf{U} f(\underline{\Lambda}) \mathbf{U}^T \tag{6}$$

where $f(\underline{\Lambda}) = \mathrm{diag}(f(\lambda_1), \ldots, f(\lambda_n))$, of course, only if every eigenvalue is in the domain of f. In this way, for a positive semidefinite (or positive definite) $\mathbf{A}$, its squareroot is

$$\mathbf{A}^{1/2} = \mathbf{U}\underline{\Lambda}^{1/2}\mathbf{U}^T, \tag{7}$$

and for a regular $\mathbf{A}$ its inverse is obtained by applying the $f(x) = x^{-1}$ function to it:

$$\mathbf{A}^{-1} = \mathbf{U}\underline{\Lambda}^{-1}\mathbf{U}^T.$$

For a singular $\mathbf{A}$, the Moore–Penrose inverse is obtained by using $\underline{\Lambda}^+$ instead of $\underline{\Lambda}^{-1}$. Accordingly, for a positive semidefinite matrix, its $-1/2$ power is defined as the squareroot of $\mathbf{A}^+$.

We will frequently use the following propositions, called *separation theorems* for singular values and eigenvalues.

Proposition 3 *Let $\mathbf{A}$ be an $m \times n$ real matrix with SVD in (3). Assume that its non-zero singular values are enumerated in non-increasing order ($s_1 \geq s_2 \geq \ldots s_r > 0$). Then*

$$\max_{\substack{\mathbf{x}\in\mathbb{R}^n,\, \mathbf{y}\in\mathbb{R}^m \\ \|\mathbf{x}\|=1,\, \|\mathbf{y}\|=1}} \mathbf{y}^T\mathbf{A}\mathbf{x} = s_1$$

and it is attained with the choice $x = \mathbf{u}_1$ and $y = \mathbf{v}_1$ (uniquely if $s_1 > s_2$). This was the $k = 1$ case. Further, for $k = 2, 3, \ldots, r$

$$\max_{\substack{\mathbf{x}\in\mathbb{R}^n,\, \mathbf{y}\in\mathbb{R}^m \\ \|x\|=1,\, \|y\|=1 \\ \mathbf{x}^T\mathbf{u}_i=0\,(i=1,\ldots,k-1) \\ \mathbf{y}^T\mathbf{v}_i=0\,(i=1,\ldots,k-1)}} \mathbf{y}^T\mathbf{A}\mathbf{x} = s_k$$

and it is attained with the choice $\mathbf{x} = \mathbf{u}_k$ and $y = \mathbf{v}_k$ (uniquely if $s_k > s_{k+1}$).

Proposition 4 *Let $\mathbf{A}$ be $n \times n$ real symmetric matrix with SD in (1). Assume that its eigenvalues are enumerated in non-increasing order ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$). Then*

$$\max_{\|\mathbf{x}\|=1} \mathbf{x}^T\mathbf{A}\mathbf{x} = \lambda_1$$

and it is attained with the choice $\mathbf{x} = \mathbf{u}_1$ (uniquely if $\lambda_1 > \lambda_2$). This was the $k = 1$ case. Further, for $k = 2, 3, \ldots, n$

$$\max_{\substack{\mathbf{x}\in\mathbb{R}^n,\, \|\mathbf{x}\|=1 \\ \mathbf{x}^T\mathbf{u}_i=0\,(i=1,\ldots,k-1)}} \mathbf{x}^T\mathbf{A}\mathbf{x} = \lambda_k$$

and it is attained with the choice $\mathbf{x} = \mathbf{u}_k$ (uniquely if $\lambda_k > \lambda_{k+1}$).

Many of the above propositions follow from the forthcoming so-called *minimax principle*.

Theorem 4 (Courant–Fischer–Weyl theorem) *Let* $\mathbf{A}$ *be an* $n \times n$ *symmetric real matrix with eigenvalues* $\lambda_1 \geq \cdots \geq \lambda_n$*. Then*

$$\lambda_k = \max_{\substack{F \subset \mathbb{R}^n \\ \dim(F)=k}} \min_{\substack{\mathbf{x} \in F \\ \|\mathbf{x}\|=1}} \mathbf{x}^T \mathbf{A} \mathbf{x} = \min_{\substack{F \subset \mathbb{R}^n \\ \dim(F)=n-k+1}} \max_{\substack{\mathbf{x} \in F \\ \|\mathbf{x}\|=1}} \mathbf{x}^T \mathbf{A} \mathbf{x} \quad (k = 1, \ldots, n).$$

The statement naturally extends to singular values of rectangular matrices.

Theorem 5 *Let* $\mathbf{A}$ *be an* $m \times n$ *real matrix with positive singular values* $s_1 \geq \cdots \geq s_r$*, where* $r = \operatorname{rank}(\mathbf{A})$*. Then*

$$s_k = \max_{\substack{F \subset \mathbb{R}^n \\ \dim(F)=k}} \min_{\mathbf{x} \in F} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} \quad (k = 1, \ldots, r).$$

Theorem 4 is, in turn, implied by the upcoming separation theorem. In the sequel, we will denote by $\lambda_i(.)$ the ith largest eigenvalue of the symmetric matrix in the argument (they are enumerated in non-increasing order).

Theorem 6 (Cauchy–Poincaré separation theorem) *Let* $\mathbf{A}$ *be an* $n \times n$ *symmetric real matrix and* $\mathbf{B}$ *be an* $n \times k$ *suborthogonal matrix* ($k \leq n$)*. Then*

$$\lambda_i(\mathbf{A}) \geq \lambda_i(\mathbf{B}^T \mathbf{A} \mathbf{B}) \geq \lambda_{i+n-k}(\mathbf{A}), \quad i = 1, \ldots, k.$$

The first inequality is attained with equality if $\mathbf{B}$ *contains the eigenvectors corresponding to the* k *largest eigenvalues of* $\mathbf{A}$ *in its columns; whereas, the second inequality is attained with equality if* $\mathbf{B}$ *contains the eigenvectors corresponding to the* k *smallest eigenvalues of* $\mathbf{A}$ *in its columns.*

Note that the first inequality makes sense for a k such that $\lambda_k > \lambda_{k+1}$, whereas the second inequality makes sense for a k such that $\lambda_{n-k+1} < \lambda_{n-k}$.

The Cauchy–Poincaré theorem implies the following important inequalities due to H. Weyl.

Theorem 7 (Weyl's perturbation theorem) *Let* $\mathbf{A}$ *and* $\mathbf{C}$ *be* $n \times n$ *symmetric matrices. Then*

$$\begin{aligned} \lambda_j(\mathbf{A} + \mathbf{C}) &\leq \lambda_i(\mathbf{A}) + \lambda_{j-i+1}(\mathbf{C}) \quad \text{if} \quad i \leq j, \\ \lambda_j(\mathbf{A} + \mathbf{C}) &\geq \lambda_i(\mathbf{A}) + \lambda_{j-i+n}(\mathbf{C}) \quad \text{if} \quad i \geq j. \end{aligned}$$

The above inequalities give rise to the following perturbation result for symmetric matrices. Here we consider symmetric matrices such that $\mathbf{A} = \mathbf{B} + \mathbf{C}$, where $\mathbf{C}$ is a 'small' perturbation on $\mathbf{B}$.

Theorem 8 *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be* $n \times n$ *symmetric matrices. Then*

$$|\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})| \leq \|\mathbf{A} - \mathbf{B}\| = \|\mathbf{C}\|, \quad i = 1, \ldots, n.$$

A similar statement is valid for rectangular matrices.

Theorem 9 *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be* $m \times n$ *real matrices with singular values* $s_1(\mathbf{A}) \geq \cdots \geq s_{\min\{m,n\}}(\mathbf{A})$ *and* $s_1(\mathbf{B}) \geq \cdots \geq s_{\min\{m,n\}}(\mathbf{B})$*. Then*

$$|s_i(\mathbf{A}) - s_i(\mathbf{B})| \leq \|\mathbf{A} - \mathbf{B}\|, \quad i = 1, \ldots, \min\{m, n\}.$$

Applying the above theorems for rank k matrices $\mathbf{B}$ we can solve the following optimization problems stated in a more general form, for rectangular matrices.

Theorem 10 *Let $\mathbf{A}$ be an arbitrary $m \times n$ real matrix with SVD $\sum_{i=1}^{r} s_i \mathbf{u}_i \mathbf{v}_i^T$, where r is the rank of $\mathbf{A}$. Then for any positive integer $k \leq r$ such that $s_k > s_{k+1}$,*

$$\min_{\substack{\mathbf{B} \text{ is } m \times n \\ \operatorname{rank}(\mathbf{B})=k}} \|\mathbf{A} - \mathbf{B}\| = s_{k+1} \quad \textit{and} \quad \min_{\substack{\mathbf{B} \text{ is } m \times n \\ \operatorname{rank}(\mathbf{B})=k}} \|\mathbf{A} - \mathbf{B}\|_2 = \left(\sum_{i=k+1}^{r} s_i^2 \right)^{1/2}$$

hold, and both minima are attained with the matrix $\mathbf{B}_k = \sum_{i=1}^{k} s_i \mathbf{u}_i \mathbf{v}_i^T$.

Note that $\mathbf{B}_k$ is called the *best rank k approximation* of $\mathbf{A}$, and the aforementioned theorem guarantees that it is the best approximation both in spectral and Frobenius norm. In fact, it is true for any unitary invariant norm:

$$\min_{\substack{\mathbf{B} \text{ is } m \times n \\ \operatorname{rank}(\mathbf{B})=k}} \|\mathbf{A} - \mathbf{B}\|_{\text{un}} = \|\mathbf{A} - \mathbf{B}_k\|_{\text{un}}.$$

Proposition 5 *Let $\mathbf{A}$ and $\mathbf{B}$ be $n \times n$ symmetric, positive semidefinite matrices with eigenvalues $\lambda_i(\mathbf{A})$'s and $\lambda_i(\mathbf{B})$'s. Then*

$$\operatorname{tr}(\mathbf{AB}) \leq \sum_{i=1}^{n} \lambda_i(\mathbf{A}) \cdot \lambda_i(\mathbf{B}),$$

with equality if and only if $\mathbf{A}$ and $\mathbf{B}$ commute, i.e. $\mathbf{AB} = \mathbf{BA}$.

Note that a necessary and sufficient condition for $\mathbf{A}$ and $\mathbf{B}$ commute is that they have the same system of eigenvectors (possibly, eigenspaces).

Proposition 6 *Let $\mathbf{A}$ and $\mathbf{B}$ be $n \times n$ real matrices with singular values $s_1(\mathbf{A}) \geq \cdots \geq s_n(\mathbf{A}) \geq 0$ and $s_1(\mathbf{B}) \geq \cdots \geq s_n(\mathbf{B}) \geq 0$. Then*

$$\prod_{i=1}^{k} s_i(\mathbf{AB}) \leq \prod_{i=1}^{k} [s_i(\mathbf{A}) \cdot s_i(\mathbf{B})], \quad k = 1, \ldots, n.$$

Especially, for $k = 1$, this implies that

$$s_{\max}(\mathbf{AB}) \leq s_{\max}(\mathbf{A}) \cdot s_{\max}(\mathbf{B}),$$

which is not surprising, since the maximal singular value is the operator norm of the matrix.

The next part will be devoted to the Perron–Frobenius theory of matrices with nonnegative entries. First we define the notion of the irreducibility for a quadratic matrix, and a similar notion for rectangular matrices.

Definition 6 *A quadratic matrix $\mathbf{A}$ is called reducible if there exists an appropriate permutation of its rows and columns, or equivalently, the exists a permutation matrix $\mathbf{P}$ such that, with it, $\mathbf{A}$ can be transformed into the following block-matrix form:*

$$\mathbf{PAP}^T = \begin{pmatrix} \mathbf{B} & \mathbf{O} \\ \mathbf{D} & \mathbf{C} \end{pmatrix} \quad \textit{or} \quad \mathbf{PAP}^T = \begin{pmatrix} \mathbf{B} & \mathbf{D} \\ \mathbf{O} & \mathbf{C} \end{pmatrix},$$

where $\mathbf{A}$ and $\mathbf{B}$ are quadratic matrices, whereas $\mathbf{O}$ is the zero matrix of appropriate size. A quadratic matrix is called irreducible if it is not reducible.

Note that the eigenvalues of a quadratic matrix are unaffected under the same permutation of its rows and columns, while the coordinates of the corresponding eigenvectors are subject to the same permutation. Since in Definition 6, the same permutation is applied to the rows and columns, and the spectrum of the involved block-matrix consists of the spectra of $\mathbf{B}$ and $\mathbf{C}$, the SD of a reducible matrix can be traced back to the SD of some smaller matrices.

The subsequent theorems apply to matrices of nonnegative entries.

Theorem 11 (Frobenius theorem) *Any irreducible, quadratic real matrix of nonnegative entries has a single positive eigenvalue among its maximum absolute value ones with corresponding eigenvector of all positive coordinates.*

Remark 1 *More precisely, there may be $k \geq 1$ complex eigenvalues of maximum absolute value r, allocated along the circle of radius r in the complex plane. In fact, those complex numbers are vertices of a regular k-gone, but the point is that exactly one of these vertices is allocated on the positive part of the real axis.*

The Perron theorem is the specialized version of the Frobenius theorem, applicable to matrices of strictly positive entries.

Theorem 12 (Perron theorem) *Any irreducible, quadratic real matrix of positive entries has only one maximum absolute value eigenvalue which is positive with multiplicity one, and the corresponding eigenvector has all positive coordinates.*

As a byproduct of the proof of the above theorems, the following useful bounds for the maximum absolute value positive eigenvalue – guaranteed by the Frobenius theorem – can be obtained.

Proposition 7 *Let $\mathbf{A}$ be an irreducible $n \times n$ real matrix of nonnegative entries and introduce the following notation for the maxima and minima of the row-sums of $\mathbf{A}$:*

$$m := \min_{i \in \{1,\ldots,n\}} \sum_{j=1}^{n} a_{ij} \quad \text{and} \quad M := \max_{i \in \{1,\ldots,n\}} \sum_{j=1}^{n} a_{ij}.$$

Then the single positive eigenvalue λ with maximum absolute value admits the following lower and upper bound:

$$m \leq \lambda \leq M,$$

where either the lower or the upper bound is attained if and only if $m = M$, i.e. the row-sums of $\mathbf{A}$ have a constant value.

Finally, we introduce the *Kronecker-sum* and *Kronecker-product* of matrices.

Definition 7 *Let $\mathbf{A}_i$ be $n_i \times n_i$ matrix $(i = 1, \ldots, k)$, $n := \sum_{i=1}^{k} n_i$. The Kronecker-sum of $\mathbf{A}_1, \ldots, \mathbf{A}_k$ is the $n \times n$ block-diagonal matrix $\mathbf{A}$ the diagonal blocks of which are the matrices $\mathbf{A}_1, \ldots, \mathbf{A}_k$ in this order. We use the notation $\mathbf{A} = \mathbf{A}_1 \oplus \cdots \oplus \mathbf{A}_k$ for it.*

Definition 8 *Let* $\mathbf{A}$ *be* $p \times n$ *and* $\mathbf{B}$ *be* $q \times m$ *real matrix. Their Kronecker-product, denoted by* $\mathbf{A} \otimes \mathbf{B}$*, is the following* $pq \times nm$ *block-matrix: it has* pn *blocks each of which is a* $q \times m$ *matrix such that the block indexed by* (i, j) *is the matrix* $a_{ij}\mathbf{B}$ $(i = 1, \ldots, p;\ j = 1, \ldots, n)$.

This product is associative, for the addition distributive, but usually not commutative. If $\mathbf{A}$ is $n \times n$ and $\mathbf{B}$ is $m \times m$ quadratic matrix, then

$$|\mathbf{A} \otimes \mathbf{B}| = |\mathbf{A}|^m \cdot |\mathbf{B}|^n;$$

further, if both are regular, then so is their Kronecker-product. Namely,

$$(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}.$$

It is also useful to know that – provided $\mathbf{A}$ and $\mathbf{B}$ are symmetric – the spectrum of $\mathbf{A} \otimes \mathbf{B}$ consists of the real numbers

$$\alpha_i \beta_j \quad (i = 1, \ldots, n;\ j = 1, \ldots, m),$$

where α_i's and β_j's are the eigenvalues of $\mathbf{A}$ and $\mathbf{B}$, respectively.

Random vectors

Random vectors are vector valued random variables with distribution characterized by the joint distribution of their coordinates. Scalar valued random variables will be denoted by upper-case letters, whereas random vectors by boldface upper-case ones (usually with letters at the end of the alphabet).

Consider first a two-variate joint distribution for introducing the notion of conditional expectation. We will distinguish between the following two cases depending on the coordinates of the random vector (X, Y).

(a) Both X and Y are so-called categorical variables (they have finitely many values which cannot be compared on any scale, like hair-color and eye-color, medical diagnoses or possible answers to a questionnaire). Say, X takes on m possible values $x_1, \ldots, x_m$, while Y takes on n possible ones $y_1, \ldots, y_n$. The joint distribution of X and Y is defined by the probabilities $p_{ij} = \mathbb{P}(X = x_i, Y = y_j)$ such that $\sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} = 1$. These are usually estimated from the frequency counts collected in an $m \times n$ rectangular array, called *contingency table*. The marginal distributions of X and Y are given by the probabilities

$$p_{i\cdot} = \sum_{j=1}^{n} p_{ij} \quad \text{and} \quad p_{\cdot j} = \sum_{i=1}^{m} p_{ij},$$

respectively. The conditional distribution of Y given $X = x_i$ is defined by the conditional probabilities $\frac{p_{ij}}{p_{i\cdot}}$ for $j = 1, \ldots n$, and the *conditional expectation* of Y under the same condition is

$$\mathbb{E}(Y|X = x_i) = \sum_{j=1}^{n} y_j \frac{p_{ij}}{p_{i\cdot}} = \frac{1}{p_{i\cdot}} \sum_{j=1}^{n} y_j p_{ij}, \quad i = 1, \ldots, m.$$

Note that neither the conditional distribution nor the conditional expectation of Y depends on the actual value of X. Making use of this property, we can define the $\mathbb{E}(Y|X)$ random variable which takes on the possible value $\mathbb{E}(Y|X = x_i)$ with probability $p_{i\cdot}$ for $i = 1, \ldots, m$. (We may say that the random variable $\mathbb{E}(Y|X)$ is measurable with respect to the σ-algebra generated by X, but we do not want to go into measure theoretical considerations.) The most important fact is that $\mathbb{E}(Y|X)$ is a measurable function of X.

(b) Both X and Y are absolutely continuous random variables (e.g. body-hight and body-weight, or two clinical measurements), then we use the joint density $f(x, y)$ in the calculations. By means of the marginal density $f_X(x) = \int f(x, y)\, dy$ we define the conditional distribution of Y given that X takes on the value x by means of the conditional density $\frac{f(x,y)}{f_X(x)}$, and the *conditional expectation* of Y under the same condition is

$$\mathbb{E}(Y|X = x) = \int y \frac{f(x, y)}{f_X(x)}\, dy = \frac{1}{f_X(x)} \int y f(x, y)\, dy.$$

The conditional expectation of Y given X is the random variable $\mathbb{E}(Y|X)$ which is again a measurable function of X.

In both cases the conditional expectation $\mathbb{E}(Y|X)$ provides the best least-square approximation of Y in terms of measurable functions of X in the following sense:

$$\min_{t=t(X)} \mathbb{E}(Y - t(X))^2 = \mathbb{E}(Y - \mathbb{E}(Y|X))^2.$$

The conditional expectation $E(X|Y)$ can be defined likewise, akin to the conditional expectation of a subset of coordinates on another subset of a random vector with several coordinates.

In fact, we take conditional expectations in the everyday life. For example, if we have recorded students' grades in two subjects which measure similar abilities, and we have lost the grade of a student in subject Y, then we can conclude for it, based on his or her grade in subject X in the following way. We take the average Y-grade of other students who have the same X-grade as the student in question. In this way, we take the conditional expectation of the (unknown) Y-grade given the (known) X-grade. (Of course, grades are coded with integers and the conditional expectation is rounded). Or, in other situation, we conclude for the (unknown) age of a person through the average (known) age of those who are similar to him/her in other respects.

Definition 9 *The expectation (vector) of the random vector* $\mathbf{X} = (X_1, \ldots, X_n)^T$ *is* $\mathbb{E}\mathbf{X} = (\mathbb{E}X_1, \ldots, \mathbb{E}X_n)^T$. *The covariance matrix of* $\mathbf{X}$ *is the* $n \times n$ *symmetric matrix* $\mathbf{C}$ *of entries:*

$$c_{ij} = \mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i) = \mathbb{E}[(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)], \quad i, j = 1, \ldots, n.$$

Sometimes the covariance matrix of $\mathbf{X}$ is denoted by $\mathrm{Var}(\mathbf{X})$, and its diagonal entries are the variances of the components of $\mathbf{X}$. With matrix notation

$$\mathrm{Var}(\mathbf{X}) = \mathbb{E}[(\mathbf{X} - \mathbb{E}\mathbf{X})(\mathbf{X} - \mathbb{E}\mathbf{X})^T],$$

where the expectation of a matrix is the matrix of the expectations of its entries.

The covariance matrix is always positive semidefinite ($\mathbf{C} \geq 0$) and it is positive definite ($\mathbf{C} > 0$) if and only if there are no linear relations between the components of $\mathbf{X}$. Note that the independence of X_i and X_j always implies $c_{ij} = 0$ ($i \neq j$); however, the converse is usually not true, except if the components are normally distributed.

In multivariate statistics, the most frequently used multivariate distribution is the *multivariate normal (Gaussian) distribution.*

Definition 10 *We say that* $\mathbf{Y}$ *is a* p*-dimensional standard normal vector if its components are independent standard normal variables. Let* $\mathbf{A}$ *be a* $p \times p$ *regular real matrix and* $\mathbf{m} \in \mathbb{R}^p$ *be a vector. Then the linear transformation* $\mathbf{X} = \mathbf{AY} + \mathbf{m}$ *defines a* p*-dimensional random normal vector.*