Optimization and Computational Linear Algebra for Data Science Lecture 6: Eigenvalues, eigenvectors and Markov chains

Léo MIOLANE · leo.miolane@gmail.com
June 19, 2019

Warning: This material is not meant to be lecture notes. It only gathers the main concepts and results from the lecture, without any additional explanation, motivation, examples, figures...

1 Eigenvalues and eigenvectors

Definition 1.1

Let $A \in \mathbb{R}^{n \times n}$. A **non-zero** vector $v \in \mathbb{R}^n$ is said to be an eigenvector of A is there exists $\lambda \in \mathbb{R}$ such that

$$Av = \lambda v$$
.

The scalar λ is called the eigenvalue (of A) associated to v. The set

$$E_{\lambda}(A) = \{x \in \mathbb{R}^n \mid Ax = \lambda x\} = \text{Ker}(A - \lambda \text{Id})$$

is called the eigenspace of A associated to λ . The dimension of $E_{\lambda}(A)$ is called the multiplicity of the eigenvalue λ .

Remark 1.1. Notice that $E_{\lambda}(A)$ is a subspace of \mathbb{R}^n : any (non-zero) linear combination of eigenvectors associated with the eigenvalue λ is also an eigenvector of A associated with λ .

Proposition 1.1

Let $A \in \mathbb{R}^{n \times n}$. Suppose that A has an eigenvalue $\lambda \in \mathbb{R}$ and let $x \in \mathbb{R}^n$ be an eigenvector associated to λ . The following holds:

- For all $\alpha \in \mathbb{R}$, $\alpha \lambda$ is an eigenvalue of the matrix αA and x is an associated eigenvector.
- For all $\alpha \in \mathbb{R}$, $\lambda + \alpha$ is an eigenvalue of the matrix $A + \alpha \operatorname{Id}$ and x is an associated eigenvector.
- For all $k \in \mathbb{N}$, λ^k is an eigenvalue of the matrix A^k and x is an associated eigenvector.
- If A is invertible then $1/\lambda$ is an eigenvalue of the matrix inverse A^{-1} and x is an associated eigenvector.

Definition 1.2

The set of all eigenvalues of A is called the spectrum of A and denoted by Sp(A).

Proposition 1.2

 $A \ n \times n \ \text{matrix} \ A \ \text{admits} \ \text{at most} \ n \ \text{eigenvalues:} \ \#\mathrm{Sp}(A) \leq n.$

2 Diagonalizable matrices

Definition 2.1

A matrix $A \in \mathbb{R}^{n \times n}$ is said to be diagonalizable if there exists a basis (v_1, \ldots, v_n) of \mathbb{R}^n consisting of eigenvectors of A, i.e. such that there exists $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$, $Av_i = \lambda_i v_i$.

Proposition 2.1

A matrix $A \in \mathbb{R}^{n \times n}$ is diagonalizable if and only if there exists an invertible $n \times n$ matrix Pand a diagonal matrix $D = \text{Diag}(\lambda_1, \dots, \lambda_n)$ such that

$$A = PDP^{-1}.$$

In this case, the i^{th} column of P is an eigenvector of A associated with the eigenvalue λ_i .

Proposition 2.2

Let $A = P \operatorname{Diag}(\lambda_1, \dots, \lambda_n) P^{-1}$ (where $P \in \mathbb{R}^{n \times n}$ is invertible) be a diagonalizable matrix. Then

$$\operatorname{Tr}(A) = \sum_{i=1}^{n} \lambda_i$$
 and $\operatorname{rank}(A) = \#\{i \mid \lambda_i \neq 0\}.$

Consequently, A is invertible if and only if $\lambda_i \neq 0$ for all i. In such case, $A^{-1} = P \operatorname{Diag}(\lambda_1^{-1}, \dots, \lambda_n^{-1}) P^{-1}$.

Application to Markov chains

3.1 First definitions and properties

A finite Markov chain is a process which moves among the elements of a finite set E in the following manner: when at $x \in E$, the next position is chosen according to a fixed probability distribution $P(x,\cdot)$. More formally:

Definition 3.1

A sequence of random variables $(X_0, X_1, ...)$ is a Markov chain with state space E and transition matrix P if for all t > 0,

$$\mathbb{P}(X_{t+1} = y \mid X_0 = x_0, \dots, X_t = x_t) = P(x_t, y)$$

for all x_0, \ldots, x_t such that $\mathbb{P}(X_0 = x_0, \ldots, X_t = x_t) > 0$.

The transition matrix P verifies therefore, for all $x \in E$,

$$\sum_{y \in E} P(x, y) = 1. \tag{1}$$

In order to simplify the notations, we will assume that $E = \{1, 2, ..., n\}$ and write for all $i,j \in E, P_{i,j} = P(j,i)$. Note that we switched here the order of i and j. This is not what is usually done in the literature, but this will allow us to be more coherent. Such matrix is said to be stochastic:

Definition 3.2 (Stochastic matrix)

A matrix $P \in \mathbb{R}^{n \times n}$ is said to be stochastic if:

(i)
$$P_{i,j} \geq 0$$
 for all $1 \leq i, j \leq n$.

(i)
$$P_{i,j} \ge 0$$
 for all $1 \le i, j \le n$.
(ii) $\sum_{i=1}^{n} P_{i,j} = 1$, for all $1 \le j \le n$.

Let $(X_0, X_1, ...)$ be a Markov chain on $\{1, ..., n\}$ with transition matrix P. For $t \ge 0$ we will encode the distribution of X_t in the $1 \times n$ vector

$$x^{(t)} = (x_1^{(t)}, \dots, x_n^{(t)}) = (\mathbb{P}(X_t = 1), \dots, \mathbb{P}(X_t = n)) \in \Delta_n$$

where Δ_n is the "n-simplex"

$$\Delta_n \stackrel{\text{def}}{=} \left\{ x \in \mathbb{R}^n \, \Big| \, \sum_{i=1}^n x_i = 1 \text{ and } x_i \ge 0 \text{ for all } i \right\}.$$

Proposition 3.1

For all $t \geq 0$

$$x^{(t+1)} = Px^{(t)}$$
 and consequently, $x^{(t)} = P^t x^{(0)}$.

Proof. Let $i \in \{1, ..., n\}$.

$$x_i^{(t+1)} = \mathbb{P}(X_{t+1} = i) = \sum_{j=1}^n \mathbb{P}(X_{t+1} = i | X_t = j) \mathbb{P}(X_t = j) = \sum_{i=1}^n P_{i,j} x_j^{(t)} = (x^{(t)} P)_i.$$

Corollary 3.1

Let P be a stochastic matrix. Then

- For all $x \in \Delta_n$, $Px \in \Delta_n$.
- For all $t \ge 1$, P^t is stochastic.

3.2 Invariant measures and the Perron-Frobenius Theorem

We will be interested in the distribution of X_t for t large, that is the limit of $x^{(t)} = x^{(0)}P^t$. As we will see, under suitable conditions on the matrix A, this

Definition 3.3

A vector $\mu \in \Delta_n$ is an invariant measure for the transition matrix P if $\mu = P\mu$, i.e.

for all
$$j \in \{1, ..., n\}$$
, $\mu_i = \sum_{j=1}^n P_{i,j} \mu_j$.

Remark 3.1. An invariant measure is an eigenvector of P with associated eigenvalue 1.

Theorem 3.1 (Perron-Frobenius, stochastic case)

Let P be a stochastic matrix such that there exists $k \ge 1$ such that all the entries of P^k are strictly positive. Then the following holds:

- (i) 1 is an eigenvalue of P and there exists an eigenvector $\mu \in \Delta_n$ associated to 1.
- (ii) The eigenvectors associated to 1 are unique up to scalar multiple (i.e. $Ker(P Id) = Span(\mu)$).
- (iii) For all $x \in \Delta_n$, $P^t x \xrightarrow[t \to \infty]{} \mu$.

Theorem 3.1 is proved in the next section.

Corollary 3.2

Let P be a stochastic matrix such that there exists $k \ge 1$ such that all the entries of P^k are strictly positive. Then there exists a unique invariant measure μ and for all initial condition $x^{(0)} \in \Delta_n$,

$$x^{(t)} \xrightarrow[t \to \infty]{} \mu.$$

3.3 Proof of Theorem 3.1

We first prove the theorem in the case k = 1, when $P_{i,j} > 0$ for all i, j.

Lemma 3.1

The mapping

$$\varphi: \Delta_n \to \Delta_n$$

$$x \mapsto Px$$

is contracting for the ℓ_1 -norm: there exists $c \in (0,1)$ such that for all $x,y \in \Delta_n$:

$$||Px - Py||_1 \le c||x - y||_1.$$

Proof. First notice that φ is well-defined by Corollary 3.1. Let us write $\alpha \stackrel{\text{def}}{=} \min_{i,j} P_{i,j} \in (0,1)$. Let $x,y \in \Delta_n$. We will show that $||Px - Py||_1 \le (1-\alpha)||x-y||_1$, i.e. $||Pz||_1 \le \alpha ||z||_1$ where z = x - y. Compute

$$||Pz||_1 = \sum_{i=1}^n |(Pz)_i| = \sum_{i=1}^n |\sum_{j=1}^n P_{i,j}z_j|.$$

Since $\sum_{j} z_{j} = 0$ we have $\sum_{j} (P_{i,j} - \alpha/n) z_{j} = \sum_{j} P_{i,j} z_{j}$. Hence

$$||Pz||_1 = \sum_{i=1}^n \left| \sum_{j=1}^n (P_{i,j} - \alpha/n) z_j \right| \le \sum_{i=1}^n \sum_{j=1}^n (P_{i,j} - \alpha/n) |z_j| = \sum_{j=1}^n (1 - \alpha) |z_j| = (1 - \alpha) ||z||_1.$$

Using Lemma 3.1, Banach fixed point Theorem tells us that φ admits a unique fixed point μ on Δ_n (i.e. a unique $\mu \in \Delta_n$ such that $P\mu = \mu$) and that for all $x \in \Delta_n$, $P^t x \xrightarrow[t \to \infty]{} \mu$. This proves Theorem 3.1 in the case k = 1.

In the case k > 1 we simply apply the result for k = 1 to P^k .

This gives that there exists a unique $\mu \in \Delta_n$ such that $P^k \mu = \mu$. Multiplying by P on both sides leads to $P^k(P\mu) = P\mu$. Since $P\mu \in \Delta_n$ we obtain that $P\mu = \mu$ by uniqueness of μ . This proves (i). To prove (ii) we consider $x \in \mathbb{R}^n$ such that Px = x. By iteration we get $P^k x = x$ which implies (using the result on P^k) that $x \in (\mu)$. To prove (iii) we fix $\ell \in \{0, \dots, k-1\}$. Let $x \in \Delta_n$. By applying the point (iii) to P^k , we have

$$P^{kt}P^{\ell}x \xrightarrow[t\to\infty]{} \mu.$$

Since this holds for all $\ell \leq k-1$ we obtain that $P^T x \xrightarrow[T \to \infty]{} \mu$ using the Euclidean division of T by k.

4 Example: Google's PageRank algorithm

4.1 The PageRank algorithm

4.2 Ranking tennis players

We aim at ranking the following n = 54 tennis players:

Federer, Nadal, Djokovic, Murray, Del Potro, Roddick, Coria, Zverev, Ferrer, Soderling, Tsonga,
Nishikori, Raonic, Nalbandian, Wawrinka, Berdych, Hewitt, Tsitsipas, Monfils, Gonzalez,
Thiem, Ljubicic, Davydenko, Cilic, Pouille, Safin, Isner, Dimitrov, Medvedev, Ferrero, Goffin,
Bautista Agut, Sock, Gasquet, Simon, Blake, Monaco, Coric, Stepanek, Khachanov, Almagro,
Robredo, Verdasco, Anderson, Youzhny, Baghdatis, Dolgopolov, Kohlschreiber, Fognini, Melzer,
Paire, Querrey, Tomic, Basilashvili.

To do so, we have access to the "head to head" record between them (see Figure 1) in the from of the matrix $R \in \mathbb{R}^{n \times n}$:

$$R_{i,j} =$$
 « number of wins of player i against player j ». (2)

We will use the PageRank strategy of the previous section in order to rank the players. In our case, instead of a "drunk surfer" we will consider a "drunk spectator". At time t the value $X_t \in \{1, ..., n\}$ indicates which tennis player our spectator believes to be the best. At time t+1, the spectator picks uniformly at random a game played by its favorite player X_t against one of the other players, x. If X_t wins the game, then the spectator still believes that X_t is the best: $X_{t+1} = X_t$. Otherwise the spectator changes his mind: $X_{t+1} = x$.

This corresponds to a Markov chain with transition matrix

$$P_{i,j} = \begin{cases} V_j/G_j & \text{if } i = j\\ R_{i,j}/G_j & \text{otherwise,} \end{cases}$$

where V_j denotes the total number of victories of player i and where G_j denotes the total number of game played by j:

$$V_j = \sum_{i=1}^n R_{j,i}$$
 and $G_j = \sum_{i=1}^n R_{i,j} + R_{j,i}$.

Let μ be the "Perron-Frobenius" eigenvector of P. The vector μ is displayed on Figure 2. Applying Corollary 3.2 to the matrix P we get that the "drunk spectator" will (in the $t \to \infty$ limit) spend a fraction μ_i of its time thinking that the player i is the best. The values (μ_1, \ldots, μ_n) can therefore be used to rank the players. We obtained the following order:

Federer (14.4%), Djokovic (13.7%), Nadal (13.6%), Murray (5.8%), Ferrer (3.0%), Del Potro (2.8%), Berdych (2.5%), Roddick (2.4%), Wawrinka (2.3%), Tsonga (2.1%), Nishikori (1.6%), Nalbandian (1.6%), Hewitt (1.5%), Monfils (1.5%), Davydenko (1.5%), Cilic (1.4%), Soderling (1.4%), Verdasco (1.2%), Gonzalez (1.2%), Raonic (1.2%), Ljubicic (1.2%), Gasquet (1.2%), Simon (1.1%), Thiem (1.1%), Isner (1.0%), Zverev (1.0%), Youzhny (1.0%), Robredo (0.9%), Kohlschreiber (0.9%), Ferrero (0.9%), Stepanek (0.8%), Safin (0.8%), Dimitrov (0.8%), Almagro (0.7%), Baghdatis (0.7%), Blake (0.7%), Anderson (0.7%), Goffin (0.7%), Coria (0.7%), Bautista Agut (0.6%), Monaco (0.6%), Fognini (0.6%), Querrey (0.6%), Melzer (0.6%), Dolgopolov (0.5%), Coric (0.5%), Pouille (0.4%), Tsitsipas (0.4%), Sock (0.4%), Paire (0.3%), Medvedev (0.3%), Khachanov (0.3%), Tomic (0.2%), Basilashvili (0.1%).



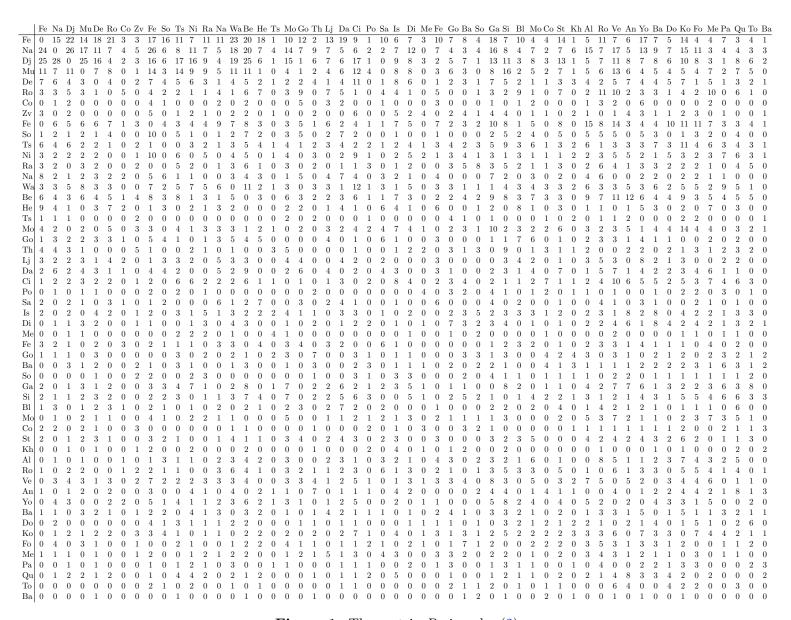


Figure 1: The matrix R given by (2)

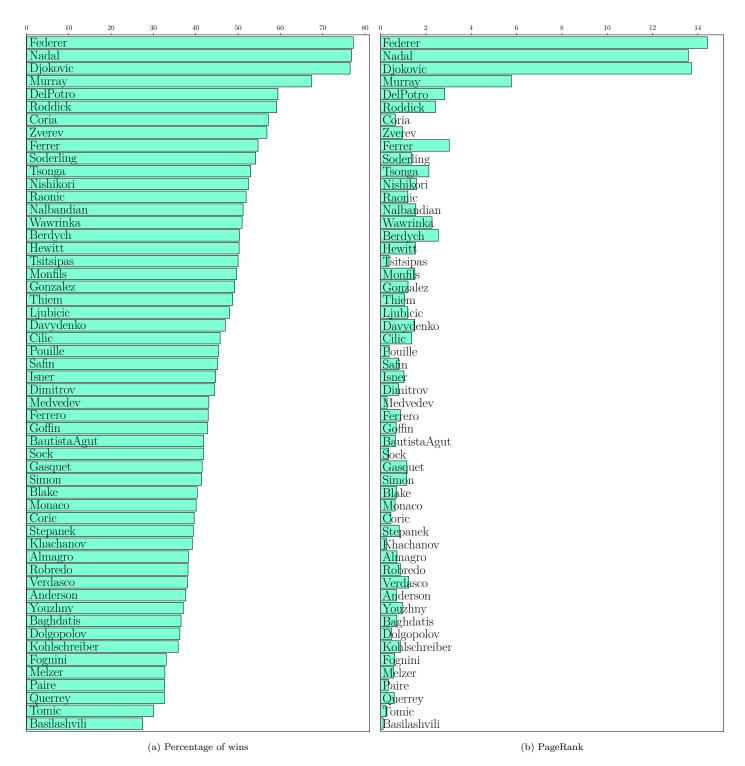


Figure 2: Comparison of the ranking by the percentage of wins (on the left) and the ranking using PageRank.