

*Comment établir le classement des équipes d'une ligue
à partir du résultat des parties jouées dans la saison ?
L'algèbre linéaire peut nous aider.*

Classement des équipes sportives : l'algèbre linéaire à la rescousse

Paul Guertin
Collège André-Grasset



Wesley Colley

À la fin d'une saison, une ligue sportive souhaite établir un classement des équipes de la ligue à partir du résultat des parties jouées durant la saison. On cherche donc une méthode qui permet, à partir d'une liste de toutes les parties et de leur pointage final, de donner un rang à chaque équipe (le rang 1 correspondant à la meilleure équipe).

Il est important de noter qu'on ne suppose pas que chaque équipe a joué une seule fois contre chaque autre équipe : les deux équipes peuvent s'être rencontrées plus d'une fois, ou ne jamais avoir joué l'une contre l'autre. On ne suppose pas non plus que toutes les équipes ont joué le même nombre de parties.

Dans un premier temps, on va supposer que les parties nulles ne se produisent pas : chaque partie se solde par la victoire d'une équipe et la défaite de son adversaire. Cette condition sera levée un peu plus tard.

La première idée qui peut nous venir est de classer les équipes selon leur nombre de victoires : une équipe qui a gagné plus de parties est considérée supérieure à une équipe qui en a gagné moins. Pour tenir compte du fait que les équipes n'ont pas forcément toutes joué le même nombre de parties, on peut utiliser plutôt la proportion de victoires, c'est-à-dire le nombre de victoires d'une équipe divisé par le nombre de parties jouées par celle-ci.

Cependant, cette première idée ne paraît pas entièrement équitable. Un premier défaut est qu'elle semble favoriser les équipes qui ont joué moins de parties. Pour prendre un exemple extrême, si une équipe A a joué une seule partie et l'a gagnée, elle sera classée meilleure qu'une équipe B qui a joué 100 parties et en a gagné 99.

Un deuxième défaut est qu'elle ne tient pas compte de la force des adversaires et donc qu'elle avantage les équipes qui ont joué contre des équipes faibles. Une équipe qui a gagné 15 parties sur 20 en jouant contre des adversaires faibles sera classée meilleure qu'une équipe qui a gagné 14 parties sur 20 contre des adversaires forts. Mais comment savoir si une équipe est forte ou faible avant d'avoir établi le classement, ce qui est précisément ce que nous tentons de faire ? Il semble que nous soyons en train de faire un raisonnement circulaire.

Le mathématicien Wesley Colley de l'université Princeton s'est penché sur ce problème en 2002, dans le but d'établir un classement pour une ligue de football universitaire aux États-Unis, et a proposé la première méthode que nous allons présenter. Cette méthode utilise des notions de probabilité bayésienne et d'algèbre linéaire pour obtenir un classement des équipes.

Dans un deuxième temps, nous verrons une autre méthode, due à Kenneth Massey, qui fait aussi appel à l'algèbre linéaire, spécifiquement à la méthode des moindres carrés, pour résoudre un système d'équations linéaires incompatibles.

La méthode de Colley

Soit n le nombre d'équipes dans la ligue. Numérotions les équipes de 1 à n . Nous allons noter v_i le nombre de victoires de l'équipe i et p_i le nombre de parties jouées par cette équipe. Le nombre de défaites de l'équipe i est donc $d_i = p_i - v_i$.

La force f_i d'une équipe dépend de sa proportion de victoires, donc on serait tenté de la définir comme étant égale à v_i/p_i . Cependant, nous allons plutôt définir

$$f_i = \frac{v_i + 1}{p_i + 2}. \quad (1)$$

Les nombres 1 et 2 dans cette formule ne sont pas arbitraires mais découlent de la règle de succession de Laplace. Avec cette définition, une équipe qui n'a joué aucune partie a une force de $1/2$, une équipe qui a joué une seule partie et l'a gagnée a une force de $2/3$ et une équipe qui a gagné 99 de ses 100 parties a une force de $100/102 \approx 0,98$. La force f_i d'une équipe est donc un nombre entre 0 et 1 correspondant à la probabilité qu'elle gagne sa prochaine partie, étant donné ses résultats jusqu'à maintenant.

Si on établissait le classement des équipes en les triant selon les valeurs des f_i , cela ne tiendrait pas compte de la force de leurs adversaires. Au lieu de cela, Colley fait un tour de passe-passe algébrique en manipulant l'équation (1) qui définit la force d'une équipe. On la réécrit d'abord sous la forme

$$f_i(p_i + 2) = v_i + 1,$$

et on réécrit le membre de droite sous une forme plus compliquée en ajoutant et en retranchant la même quantité.

$$f_i(p_i + 2) = \frac{v_i + d_i}{2} + \frac{v_i - d_i}{2} + 1.$$

Dans cette équation, $v_i + d_i$ est simplement le nombre de parties totales jouées par l'équipe i , donc on peut écrire

$$f_i(p_i + 2) = \frac{p_i}{2} + \frac{v_i - d_i}{2} + 1.$$

Si on ne savait rien sur la force des adversaires, on prendrait leur force égale à $1/2$ et on aurait

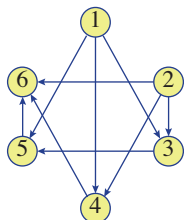
$$f_i(p_i + 2) = \sum_{k=1}^{p_i} \frac{1}{2} + \frac{v_i - d_i}{2} + 1.$$

chaque terme de la somme correspondant à une partie jouée par l'équipe i . Mais pour tenir compte de la force des adversaires, on remplace le terme constant $1/2$ dans la somme par la force de chaque adversaire telle que calculée par la formule (1). Si on note $a_1, a_2, \dots, a_{p_i}$ les numéros des équipes que l'équipe i a affrontées, on obtient

$$f_i(p_i + 2) = \sum_{k=1}^{p_i} f_{a_k} + \frac{v_i - d_i}{2} + 1.$$

Chacune des n équipes donne lieu à une telle équation. On obtient donc un système de n équations en n inconnues f_1 à f_n . Comme chaque équation est linéaire en les f_i , on peut utiliser l'algèbre linéaire pour résoudre le système, déterminer la valeur des f_i et obtenir le classement cherché (l'équipe dont le f_i est maximal finissant au premier rang).





Un exemple

Considérons une ligue avec six équipes numérotées de 1 à 6. Le tableau ci-dessous donne les neuf parties qui ont été jouées et le résultat de chacune.

Numéro de la partie	Gagnant	Perdant
1	1	3
2	1	4
3	1	5
4	2	3
5	2	4
6	2	6
7	3	5
8	4	6
9	5	6

On voit que chaque équipe a joué trois parties :

- l'équipe 1 a gagné ses trois parties contre les équipes 3, 4 et 5;
- l'équipe 2 a gagné ses trois parties contre les équipes 3, 4 et 6;
- l'équipe 3 a perdu contre 1 et 2 et gagné contre 5;
- l'équipe 4 a perdu contre 1 et 2 et gagné contre 6;
- l'équipe 5 a perdu contre 1 et 3 et gagné contre 6;
- l'équipe 6 a perdu ses trois parties contre 2, 4 et 5.

On peut aussi représenter cette situation au moyen d'un graphe orienté. Les sommets représentent les six équipes et il y a un arc du sommet X au sommet Y si l'équipe X a gagné sa partie contre l'équipe Y .

Équipe	v_i	d_i	$\frac{v_i - d_i}{2} + 1$
1	3	0	$\frac{3-0}{2} + 1 = 2,5$
2	3	0	$\frac{3-0}{2} + 1 = 2,5$
3	1	2	$\frac{1-2}{2} + 1 = 0,5$
4	1	2	$\frac{1-2}{2} + 1 = 0,5$
5	1	2	$\frac{1-2}{2} + 1 = 0,5$
6	0	3	$\frac{0-3}{2} + 1 = -0,5$

Le système d'équations linéaires que nous allons établir comporte six équations, chacune de la forme

$$f_i(p_i + 2) = \sum_{k=1}^{p_i} f_{a_k} + \frac{v_i - d_i}{2} + 1.$$

Comme chaque équipe a joué trois parties, le membre de gauche est simplement $5f_i$ pour chaque équipe. On peut aussi calculer la valeur de l'expression $\frac{v_i - d_i}{2} + 1$ pour chaque équipe.

Le système de Colley est donc donné par

$$\begin{cases} 5f_1 = f_3 + f_4 + f_5 + 2,5 \\ 5f_2 = f_3 + f_4 + f_6 + 2,5 \\ 5f_3 = f_1 + f_2 + f_5 + 0,5 \\ 5f_4 = f_1 + f_2 + f_6 + 0,5 \\ 5f_5 = f_1 + f_3 + f_6 + 0,5 \\ 5f_6 = f_2 + f_4 + f_5 - 0,5 \end{cases}$$

qu'on peut écrire sous forme matricielle $A\vec{x} + \vec{b}$ avec $\vec{x}$ le vecteur des forces:

$$\begin{pmatrix} 5 & 0 & -1 & -1 & -1 & 0 \\ 0 & 5 & -1 & -1 & 0 & -1 \\ -1 & -1 & 5 & 0 & -1 & 0 \\ -1 & -1 & 0 & 5 & 0 & -1 \\ -1 & 0 & -1 & 0 & 5 & -1 \\ 0 & -1 & 0 & -1 & -1 & 5 \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \\ f_6 \end{pmatrix} = \begin{pmatrix} 2,5 \\ 2,5 \\ 0,5 \\ 0,5 \\ 0,5 \\ -0,5 \end{pmatrix}$$

La résolution de ce système donne, à trois décimales :

$$\begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \\ f_6 \end{pmatrix} = \begin{pmatrix} 0,761 \\ 0,725 \\ 0,475 \\ 0,439 \\ 0,389 \\ 0,211 \end{pmatrix}$$

L'équipe 1 finit donc en tête, suivie de l'équipe 2. Même si ces deux équipes ont le même nombre de victoires et de défaites, l'équipe 1 est favorisée parce qu'elle a battu l'équipe 5 qui a battu l'équipe 6, tandis que l'équipe 2 a battu l'équipe 6. De même, parmi les trois équipes qui ont une victoire et deux défaites, l'équipe 3 est considérée supérieure à 4 et 5 parce qu'elle a obtenu des victoires contre des adversaires plus forts.

Existence et unicité de la solution

On peut se demander si le système linéaire engendré par la méthode de Colley possède toujours une solution unique. La réponse est oui, en vertu d'un théorème d'algèbre linéaire appelé théorème de Levy-Desplanques. Ce théorème dit que si une matrice carrée est telle que sur chaque ligne, la valeur absolue de l'entrée située sur la diagonale principale est strictement supérieure à la somme des valeurs absolues des autres entrées sur la ligne, alors cette matrice est inversible.

Dans un système linéaire de Colley, l'entrée de la diagonale sur la ligne i est $2+p_i$ tandis qu'il y a p_i autres entrées non nulles sur la même ligne, chacune valant ± 1 . La somme des valeurs absolues des entrées d'une ligne, en excluant la diagonale, est donc égale à p_i qui est inférieur à $2+p_i$. L'hypothèse du théorème est donc vérifiée, ce qui implique que la matrice A est inversible. L'équation matricielle $A\vec{x} + \vec{b}$ possède donc toujours une solution unique.

La méthode de Massey

La méthode de Colley présentée à la section précédente ne tient pas compte de la marge de victoire des parties. Une victoire par la marque de 3 à 2 a le même poids qu'une victoire écrasante de 10 à 1. On peut raisonnablement penser que la marge de victoire (c'est-à-dire l'écart entre le pointage de l'équipe gagnante et le pointage de l'équipe perdante) nous donne de l'information qui peut être utile pour évaluer la force d'une équipe, et que la méthode de Colley n'est pas parfaite puisqu'elle n'utilise pas cette information.

Une méthode différente, due à Kenneth Massey, établit un classement des équipes en utilisant la marge de victoire pour chaque partie. Un avantage de cette méthode est qu'elle est applicable même si une partie peut être nulle (la marge de victoire sera alors de zéro). La méthode de Massey, comme la méthode de Colley, utilise l'algèbre linéaire pour résoudre un système d'équations mais là s'arrête la ressemblance.

L'idée est la suivante : on considère que lorsque deux équipes jouent une partie, la marge de victoire représente la différence de force entre les deux équipes. Par exemple, si les équipes i et j jouent une partie et que l'équipe i gagne par k points, alors il est naturel de considérer que $f_i - f_j = k$, où f_i représente la force de l'équipe i et f_j celle de l'équipe j .

Chaque partie nous donne une équation qui relie les forces des deux équipes impliquées. On obtient donc un système d'équations linéaires, chaque équation impliquant seulement deux variables. Ce système peut être écrit sous la forme matricielle $A\vec{x} = \vec{b}$.

Cependant, le système contient autant d'équations qu'il y a de parties, et autant d'inconnues qu'il y a d'équipes. Comme le nombre de parties sera le plus souvent supérieur au nombre d'équipes, le système sera surdéterminé : il contiendra plus d'équations que d'inconnues. Il sera donc possible, et en fait, très probable, qu'il n'ait aucune solution. La méthode de Massey consiste à utiliser la méthode des moindres carrés pour normaliser le système sous la forme

$$A^T A \vec{x} = A^T \vec{b}$$

et trouver la valeur des forces d'équipes qui est aussi près que possible d'une solution. Ce sont ces forces qui seront utilisées pour établir le classement des équipes.

Un dernier détail est que le système normal ainsi construit n'a pas une solution unique mais une infinité de solutions. Massey impose donc une condition supplémentaire (la somme des forces des équipes doit sommer à zéro) pour obtenir une solution unique. Cela se fait simplement en ajoutant une équation linéaire au système normal et ne complique pas le calcul.



Kenneth Massey

Un exemple

Par exemple, considérons quatre équipes, numérotées de 1 à 4, qui jouent chacune une fois contre chaque autre équipe. Les résultats des six parties sont donnés ci-dessous.

Partie	Equipes	Résultat		Marge de la victoire
		Équipe gagnante	Pointage	
1	1 vs 2	1	$\frac{5}{2}$	3
2	3 vs 1	3	$\frac{4}{3}$	1
3	1 vs 4	1	$\frac{3}{2}$	1
4	2 vs 3	2	$\frac{4}{3}$	1
5	4 vs 2	4	$\frac{3}{1}$	2
6	3 vs 4	3	$\frac{3}{2}$	1

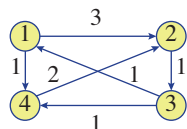
On peut aussi représenter cette situation au moyen d'un graphe orienté pondéré. Les sommets représentent les équipes et il y a un arc du sommet X au sommet Y si l'équipe X a gagné sa partie contre l'équipe Y . Chaque arc est pondéré par la marge de victoire.

Si on note f_i la force de l'équipe i , on obtient les six équations linéaires

$$\begin{aligned} f_1 - f_2 &= 3, & f_2 - f_3 &= 1, \\ f_3 - f_1 &= 1, & f_4 - f_2 &= 2, \\ f_1 - f_4 &= 1, & f_3 - f_4 &= 1. \end{aligned}$$

qui donnent lieu à l'équation matricielle $A\vec{x} + \vec{b}$

$$\begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \\ 1 \\ 2 \\ 1 \\ 1 \end{pmatrix}$$



Il est facile de voir que ce système surdéterminé ne possède pas de solution. En effet, si on additionne la deuxième et la troisième équation, on obtient $f_3 - f_4 = 2$, ce qui contredit la dernière équation $f_3 - f_4 = 1$.

Utilisons la méthode des moindres carrés : au lieu de chercher un vecteur $\vec{x}$ tel que $A\vec{x}$ est égal au vecteur $\vec{b}$, cherchons un vecteur $\vec{x}$ ayant la propriété que $A\vec{x}$ est le plus près possible de $\vec{b}$. Autrement dit, cherchons le vecteur $\vec{x}$ qui minimise la norme euclidienne $\|A\vec{x} - \vec{b}\|$. Comme une norme n'est jamais négative, minimiser la norme revient à minimiser son carré. La fonction à minimiser est donc

$$\|A\vec{x} - \vec{b}\|^2 = (A\vec{x} - \vec{b})^T (A\vec{x} - \vec{b}),$$

dont le gradient est $2A^T(A\vec{x} - \vec{b})$. Ce qui donne l'équation normale

$$A^T A \vec{x} = A^T \vec{b}$$

qu'il suffira ensuite de résoudre. Pour cela, nous allons multiplier l'équation à gauche par A^T .

$$\begin{aligned} A^T A &= \begin{pmatrix} 1 & -1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & 1 & -1 \end{pmatrix} \\ &= \begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{pmatrix} \end{aligned}$$

et

$$A^T \vec{b} = \begin{pmatrix} 1 & -1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} 3 \\ 1 \\ 1 \\ 2 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ -4 \\ 1 \\ 0 \end{pmatrix}$$

L'équation normale est donc

$$\begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_3 \end{pmatrix} = \begin{pmatrix} 3 \\ -4 \\ 1 \\ 0 \end{pmatrix}.$$

Ce système possède une infinité de solutions car la matrice $A^T A$ n'est pas inversible : la somme de toutes les lignes donne la ligne nulle. Cela est dû au fait que toutes les équations originales impliquant les forces des équipes sont de la forme $f_i - f_j = k$ et qu'ajouter une même constante à toutes les forces ne change aucune équation.

Pour s'assurer qu'il y a une solution unique, on peut simplement enlever la première équation (ce qui ne change pas les solutions du système puisque cette équation est linéairement dépendante des autres) et ajouter une nouvelle équation disant que la somme des forces est nulle : $f_1 + f_2 + f_3 + f_4 = 0$. Autrement dit, on remplace la première ligne de la matrice par une ligne de 1 et on remplace le premier coefficient du vecteur de droite par un 0.

Le système normal modifié est alors

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_3 \end{pmatrix} = \begin{pmatrix} 0 \\ -4 \\ 1 \\ 0 \end{pmatrix}$$

et la solution est :

$$\begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_3 \end{pmatrix} = \begin{pmatrix} 3/4 \\ -1 \\ 1/4 \\ 0 \end{pmatrix}.$$

Ainsi, selon la méthode de Massey, l'équipe 1 est la plus forte, suivie des équipes 3, 4 et 2 dans cet ordre.

Existence et unicité de la solution

La théorie spectrale des graphes nous apprend que la matrice $A^T A$ obtenue par la méthode de Massey (connue sous le nom de matrice laplacienne du graphe des parties) est de rang $n - c$, où n est le nombre de sommets du graphe des parties (donc le nombre d'équipes) et c est le nombre de composantes connexes de ce graphe (vu comme un graphe non orienté). Puisqu'on remplace une de ses lignes par une ligne de 1, linéairement indépendante des autres lignes, la matrice résultante sera de plein rang, donc inversible, à condition que le graphe des parties soit connexe, c'est-à-dire qu'il existe une chaîne de parties reliant n'importe quelle paire d'équipes. Il y aura donc une solution unique à l'équation normale. La condition d'avoir un graphe de parties connexe sera satisfaite dans tous les cas raisonnables, puisqu'on ne peut pas imaginer classer deux équipes qui n'ont pas de chaîne de parties qui les relient.

Finalement, il reste à montrer que c'est bien un minimum. Pour cela, il suffit de montrer que la matrice hessienne H des dérivées secondes est semi-définie positive, c'est-à-dire que $\vec{v}^T H \vec{v} \geq 0$ pour tout vecteur $\vec{v}$ non nul. On peut montrer que $H = 2A^T A$, donc pour tout $\vec{v} \neq \vec{0}$,

$$\begin{aligned} \vec{v}^T H \vec{v} &= \vec{v}^T (2A^T A) \vec{v} \\ &= 2(\vec{v}^T A^T)(A \vec{v}) \\ &= 2(A \vec{v})^T (A \vec{v}) \\ &= 2\|A \vec{v}\|^2 \geq 0, \end{aligned}$$

ce qu'on voulait montrer. Donc le point où le gradient s'annule, c'est-à-dire la solution de $A^T A \vec{x} = A^T \vec{b}$, est bien un minimum.