10 Championnat de football

L'industrie des paris sportifs, très florissante dans certains pays, a suscité le développement de modèles et de techniques visant à déterminer les cotes auxquelles les bookmakers proposent leurs paris. L'objet de ce projet est l'étude d'un modèle de championnat (où N équipes se rencontrent toutes les unes les autres), l'estimation des probabilités d'événements rares pour ce modèle, et, éventuellement, son amélioration. Ils étudieront en particulier le cas de la *Premier League*[9] *2015/16* ainsi que les régimes asymptotiques mentionnés dans l'article [Esc17].

10.1 Modèle de Bradley-Terry

On considère ici un modèle de compétition entre N équipes (qui pourraient aussi être des agents économiques, des traitements médicaux,...) dans lequel les équipes ont des valeurs intrinsèques $V_1, \ldots, V_N$ (pour tout i, $V_i > 0$ est la valeur de l'équipe i) et se confrontent deux à deux dans des "matchs" ayant un vainqueur et un perdant. L'issue d'un match est aléatoire et distribuée de la façon suivante : pour tous $i \neq j \in \{1, \ldots, N\}$, lors qu'un match entre i et j,

$$\mathbb{P}(i \text{ gagne}) = \frac{V_i}{V_i + V_j}. \tag{7}$$

On suppose par ailleurs que les matchs successifs, même si ils impliquent les mêmes équipes, sont indépendants. Chaque match rapporte un point au vainqueur (et zéro au perdant), et à la fin du championnat, le score S_i de l'équipe i est la somme des points qu'elle a gagnés lors de ses matchs. Si les équipes se rencontrent deux à deux une et une seule fois, en notant

$$X_{i,j} = 1 - X_{j,i} = \begin{cases} 1 & \text{si } i \text{ est le vainqueur du match entre } i \text{ et } j, \\ 0 & \text{si } j \text{ est le vainqueur du match entre } i \text{ et } j, \end{cases}$$

on a donc

$$S_i = \sum_{j \neq i} X_{i,j}. \tag{8}$$

Si, comme dans les championnats de football, les équipes se rencontrent toutes 2 fois, on a

$$S_i = \sum_{j \neq i} X_{i,j}^{(1)} + X_{i,j}^{(2)}, \tag{9}$$

9. La *Premier League* est la première division du championnat masculin anglais de football.

où $X^{(1)}_{i,j}$ (resp. $X^{(2)}_{i,j}$) désigne l'issue du premier (resp. deuxième) match entre i et j. Les variables aléatoires

$$(X_{i,j})_{1\leq i<j\leq N} \qquad (\text{ou } (X^{(k)}_{i,j})_{1\leq i<j\leq N, k=1,2})$$

sont des variables aléatoires de Bernoulli indépendantes telles que

$$\mathbb{P}(X_{i,j}=1)=\frac{V_i}{V_i+V_j} \qquad (\text{ou } \mathbb{P}(X^{(k)}_{i,j}=1)=\tfrac{V_i}{V_i+V_j}).$$

La loi de l'issue de la compétition est entièrement déterminée par la valeur des paramètres $V_1,\dots,V_N$ (et invariante par multiplication de ces nombres par un même nombre strictement positif).

10.2 Premier League 2015/16

La *Premier League 2015/16* a vu un événement particulièrement surprenant se produire : la petite équipe de Leicester a gagné le championnat, reléguant les grosses cylindrées de Premier League au second plan. Si les exploits de "petits poucets" lors des coupes (compétitions avec élimination directe, dont l'issue repose sur un petit nombre de matchs) sont courants, ce genre de surprises est rare dans les championnats.

a) Les élèves implémenteront le modèle décrit à l'équation (9). Pour cela, il est nécessaire de choisir des valeurs pour les V_i. La littérature fournit des méthodes d'estimation de ces paramètres (basées sur les matchs passés, voir par exemple [Hun04, YYX12]), mais nous nous contenterons ici des V_i donnés par une des deux possibilités suivantes (κ, b, θ sont des paramètres qui ne dépendant pas de i) :

- pour tout i, $V_i=(P_i)^\kappa$ où P_i est le nombre de points obtenus par l'équipe i en Premiere League l'année précédente, 2014/15 (pour les équipes promues de la seconde division, on prend le nombre de points obtenus en seconde division l'année précédente et on divise par 1.8),

ou bien :

- pour tout i, $V_i=(b-(R_i)^\theta)$ où R_i est le classement de l'équipe i en Premiere League l'année précédente, 2014/15 (pour les équipes promues de la seconde division, on prend $R_i=15$).

Les P_i, R_i, ainsi que des suggestions de valeurs pour les paramètres κ, b, θ, sont disponibles dans le fichier [Data].

Les élèves compareront les probabilités de victoires des équipes de Chelsea, Manchester City, Arsenal, Manchester United et Tottenham données par ce modèle avec celles choisies, avant le début du championnat, par les bookmakers, telles que présentées par exemple en [Sch16] (qui sont donc, pour ces équipes, respectivement, $8/(8+13)$, $2/(5+2)$, $2/(7+2)$, $1/(5+1)$ et $1/(100+1)$).

b) Les élèves estimeront alors, avec ce modèle et ces V_i, la probabilité que Leicester gagne le championnat. Ils pourront aussi estimer les probabilités d'autres événements plus rares qui se sont réalisés, comme la probabilité que Leicester gagne avec les deux équipes de Manchester hors du podium et Liverpool et Chelsea hors des *7 places européennes*. Ils pourront aussi estimer la probabilité de l'événement (non réalisé, celui-ci) de la victoire de l'équipe la plus faible selon ces V_i. Ils commenteront les probas obtenues. Ils devront pour cela utiliser l'échantillonnage préférentiel.

10.3 Régimes limites

Dans l'article [CDL16] (voir aussi la note de vulgarisation [Esc17]), les auteurs considèrent, pour des valeurs très élevées de N, l'effet de la répartition des V_i sur la question de savoir si l'équipe favorite (celle ayant le plus grand V_i) gagne toujours, dans le modèle de l'équation (8). De façon très surprenante, ils montrent que l'équipe favorite ne gagne pas toujours : ils modélisent la répartition des V_i en les supposant choisis au hasard selon certaines lois et mettent en évidence plusieurs régimes différents, certains d'entre eux donnant en fait assez rarement la victoire à l'équipe favorite.

a) Les élèves pourront implémenter cette sophistication du modèle et mettre en évidence les tendances caractéristiques des différents régimes de l'article.

b) Ils pourront ensuite estimer, au moyen de méthodes vues en cours, la probabilité d'événements rares comme celui où une équipe ayant un petit V_i gagne finalement ou bien celui où l'équipe ayant le plus grand V_i ne gagne pas, alors que celui-ci est très au dessus des autres V_i.

10.4 Perfectionnement du modèle

Le modèle de championnat décrit par les équations (7) et (9) est assez rudimentaire : il ne tient compte ni de la possibilité de matchs nuls, ni de l'évolution des équipes dans le temps, ni de la différence entre les matchs à l'extérieur et les matchs à domicile, etc... Les élèves pourront essayer d'affiner le modèle, en gardant en tête que l'augmentation du nombre de paramètres augmente le nombre de données à estimer avant toute implémentation pratique.

Références

[CDL16] R. Chetrite, R. Diel, M. Lerasle *The number of potential winners in Bradley-Terry model in random environment*, Ann. Appl. Probab., 2016. https://arxiv.org/abs/1509.07265

[Esc17] S. Escalón. *Le ballon rond à l'épreuve des probabilités,* Le Journal du CNRS, 2017. https://lejournal.cnrs.fr/articles/le-ballon-rond-a-lepreuve-des-probabilites.

[Sch16] A. Schooler *Premier League 2015/16 : How odds changed as Leicester claimed title* http://www.skysports.com/football/news/11712/10261535/premier-league-201516-how-the-odds-changed-as-leicester-claimed-the-title

[Data] `DataPL201516.py` Données disponibles sur le Moodle du cours.

[Hun04] D. Hunter. *MM algorithms for generalized Bradley-Terry models.* Ann. Statist. 32, 384–406, 2004.

[YYX12] T. Yan, Y. Yang, J. Xu *Sparse paired comparisons in the Bradley-Terry model*. Statist. Sinica 22 1305–1318, 2012.