Modèlisation statistique d'un championnat de football. Extension au tournoi de fighting

Eric Parent, Edouard Kolf

le 20 décembre 2012, AppliBUGS, Paris

Contents

Structure du championnat de France 2012

Exemple du championnat de ligue

- 20 équipes (1,...,20) soit 20*20-20=380 rencontres
- 38 journées de championnats
- 10 matchs par journée
- 3 résultats possibles : victoire, nul ou défaite

Objectif

- Estimer les forces relatives des équipes $(\lambda_1, ..., \lambda_{20})$
- Estimer les probabilités des différentes rencontres
- Parier sur les 3 résultats possibles (cf. LotoFoot)

Un exemple de Grille de LotoFoot

Résultats du Loto Foot Matches N°1034

1.	GUINGAMP	NANTES	X	N	2
2.	BASTIA	MARSEILLE	X	N	2
3.	SEDAN	METZ	X	N	2
4.	LENS	LILLE	1	X	2
5.	LORIENT	MONACO	X	N	2
6.	MONTPELLIER	SOCHAUX	1	X	2
7.	TROYES	LYON	1	N	X
8.	P.S.G.	RENNES	X	N	2
9.	GUEUGNON	CAEN	1	X	2
10.	LE HAVRE	GRENOBLE	X	N	2
11.	MARTIGUES	NANCY	1	X	2
12.	WASQUEHAL	ST-ETIENNE	1	X	2
13.	LAVAL	NIMES	X	N	2

1 Victoire de la première équipe (à domicile).
N Match nul
2 Victoire de la deuxième équipe (à l'extérieur).

Modèle de comparaison par paires

Notations

- Chaque équipe possède une force *latente* propre : λ_i
- 3 issues possibles :
  - i bat j avec probabilité Π_{ij}^{+}
  - i et j font match nul avec probabilité $\Pi_{ij}^{=}$
  - i perd contre j avec probabilité Π_{ij}^{-}

Version Duel Bradley-Terry sans match nul

- $\Pi_{ij} = \frac{\lambda_i}{\lambda_i + \lambda_j}$
- Interprétation Diaconis : v. latentes
- $Z_i \sim Exp(\lambda_i)$; $Z_j \sim Exp(\lambda_j)$
- $\Pi_{ij} = [Z_i < Z_j]$

Modele Bradley-Terry-Caron-Doucet

Avec match nul θ

- $\Pi_{ij}^{+} = \frac{\lambda_i}{\lambda_i+\theta\lambda_j}$
- $\Pi_{ij}^{=} = \frac{(\theta^2-1)\lambda_i\lambda_j}{(\lambda_i+\theta\lambda_j)(\theta\lambda_i+\lambda_j)}$
- $\Pi_{ij}^{-} = \frac{\lambda_j}{\lambda_j+\theta\lambda_i} = \Pi_{ji}^{+}$
- Et oui ! $\frac{(\theta^2-1)\lambda_i\lambda_j+\lambda_i(\theta\lambda_i+\lambda_j)+\lambda_j(\lambda_i+\theta\lambda_j)}{(\lambda_i+\theta\lambda_j)(\theta\lambda_i+\lambda_j)} = 1$

Avantage terrain δ

- $\lambda_i \to \delta\lambda_i$ si i joue à domicile
- $\Pi_{ij}^{+} = \frac{\delta\lambda_i}{\delta\lambda_i+\theta\lambda_j}$ si i joue à domicile
- $\Pi_{ij}^{=} = \frac{(\theta^2-1)\delta\lambda_i\lambda_j}{(\delta\lambda_i+\theta\lambda_j)(\theta\delta\lambda_i+\lambda_j)}$
- $\Pi_{ij}^{-} = \frac{\lambda_j}{\lambda_j+\theta\delta\lambda_i} \neq \Pi_{ji}^{+}$

Vraisemblance Bradley Terry

Vraisemblance

- rencontre (ij) : l'équipe i rencontre j et joue sur son terrain
- $\log[N|\lambda,\theta,\delta] = \sum_{(ij)} \left\{ n_{ij}^{+} log\Pi_{ij}^{+} + n_{ij}^{=} log\Pi_{ij}^{=} + n_{ij}^{-} log\Pi_{ij}^{-} \right\}$
- $\log[N_{(ij)}|\lambda,\theta,\delta] =$
  $\left(n_{(ij)}^{+} + n_{(ij)}^{=}\right) \log \frac{\delta\lambda_i}{\delta\lambda_i+\theta\lambda_j} + n_{(ij)}^{=} log(\theta^2-1) + \left(n_{(ij)}^{-} + n_{(ij)}^{=}\right) \log \frac{\lambda_j}{\lambda_j+\theta\delta\lambda_i}$
- $n_{(ij)}^{+} + n_{(ij)} + n_{(ij)}^{-} = 1$: match entre i et j
- Pb d'inférence : les dénominateurs !
  - Force brute : SIR séquentiel au cours des rencontres (cf Parent/Bernier)
  - Force aveugle : go to WinBUGS (cf Parent/Bernier)
  - Approximation Normale bayésienne de la logistique(cf Glickman)
  - Asymptotique Normale fréquentiste (cf David)
  - Elégance bayésienne : Augmentation de données (cf Caron-Doucet)

Astuce Caron-Doucet : Augmentation !

Introduction de variables latentes $Z_{ij}|\lambda_i, \lambda_j$ pour chaque rencontre (ij)

- $Z^+_{(ij)} \sim \Gamma(n^+_{(ij)} + n^=_{(ij)}, \delta\lambda_i + \theta\lambda_j)$; $Z^-_{(ij)} \sim \Gamma(n^-_{(ij)} + n^=_{(ij)}, \delta\theta\lambda_i + \lambda_j)$
-

$$\log[N_{(ij)}, Z_{(ij)}|\lambda, \theta, \delta] \equiv \left(n^+_{(ij)} + n^=_{(ij)}\right) \log \delta\lambda_i + \left(n^-_{(ij)} + n^=_{(ij)}\right) \log\lambda_j$$
$$-(\delta\lambda_i + \theta\lambda_j)z^+_{(ij)} - (\delta\theta\lambda_i + \lambda_j)z^-_{(ij)} + n^=_{(ij)} log(\theta^2 - 1)$$

- On prend des priors appropriés :
  - $\delta \sim \Gamma(a_\delta, b_\delta)$,
  - $\lambda_i \sim \Gamma(a_i, b_i)$, éventuellement fixé en fonction du classement de la saison précédente
  - NB $\lambda_i \sim \Gamma(a_i, b)$, $\Pi_{ij} = \frac{\lambda_i}{\lambda_i+\lambda_j}$ Dirichlet
- Jolie vraisemblance complète : Gibbs ou MAP par EM (ou Max Vrais) !

Méthode Caron-Doucet : Algorithme de Gibbs

Des conditionnelles complètes très Gamma

- $Z^{+}_{(ij)} \sim \Gamma(n^{+}_{(ij)} + n^{=}_{(ij)}, \delta\lambda_i + \theta\lambda_j)$; $Z^{-}_{(ij)} \sim \Gamma(n^{-}_{(ij)} + n^{=}_{(ij)}, \delta\theta\lambda_i + \lambda_j)$
- $\lambda_i \sim \Gamma(a_i + \sum_j \left(n^{+}_{(ij)} + n^{=}_{(ij)}\right) + \sum_k \left(n^{+}_{(ki)} + n^{=}_{(ki)}\right), b_i + \sum_j \left(\delta z^{+}_{(ij)} + \delta\theta z^{-}_{(ij)}\right) + \sum_k \left(\theta z^{+}_{(ki)} + z^{-}_{(ki)}\right))$
- $\delta \sim \Gamma(a_\delta + \sum_{(ij)} \left(n^{+}_{(ij)} + n^{=}_{(ij)}\right), b_\delta + \sum_i \lambda_i \sum_j \left(z^{+}_{(ij)} + \delta z^{-}_{(ij)}\right))$
- $[\theta|Z, N, \lambda, \delta] \propto [\theta] \times (\theta^2 - 1)^{\sum_{(ij)} n^{=}_{(ij)}} \exp(-K\theta) \times 1_{\theta>1}$

Comment générer selon la conditionnelle du paramètre de match nul ?

Comment générer selon $\left(\theta^2-1\right)^S \exp -K\theta \times 1_{\theta>1}$?

S est entier.

$$\theta = 1 + x$$

$$\theta^2 - 1 = x \times (x+2)$$

$$x^S (x+2)^S = S! \sum_{k=0}^{S} x^{S+k} \frac{2^{S-k}}{k!\,(S-k)!}$$

$$x^S (x+2)^S \exp -Kx \propto \sum_{k=0}^{S} \left(\frac{\Gamma(S+k+1)2^{-k}}{K^{k+1} \times k!\,(S-k)!} \right) \left(\frac{K^{S+k+1} \times x^{(S+k+1)-1}}{\Gamma(S+k+1)} \exp -Kx \right)$$

Un mélange de $S+1$ lois gamma avec $P_k \propto \frac{\Gamma(S+k+1)2^{-k}}{K^{k+1} \times k!(S-k)!}$

Remarques

Sans problème pour un bayésien...

- chaque force λ_i à une constante près, on n'apprend rien sur leur somme
- possibilité de "parameter extension", améliore drastiquement, pas fait par WinBUGS
- on peut aussi faire de l'EM, bien sûr...

Pour rappel, modèle Parent-Bernier

Conditions de comportement vis à vis de l'écart des forces Δ

- $\frac{\Pi_{ij}^{+}}{\Pi_{ij}^{=}} = g_1(\Delta)$
- $\frac{\Pi_{ij}^{=}}{\Pi_{ij}^{-}} = g_2(\Delta)$
- $g_1(\Delta)$ et $g_2(\Delta)$ fonctions croissantes de Δ
- On choisit $g_1(\Delta) = K_1 g(\Delta)$ *et* $g_2(\Delta) = K_2 g(\Delta)$ *avec* $K_1 > 0,\ K_2 > 0$

Par conséquent, le modèle doit s'écrire

- $\Pi^{-}(\Delta) = \frac{\frac{1}{K_2 g(\Delta)}}{\frac{1}{K_2 g(\Delta)} + 1 + K_1 g(\Delta)})$
- $\Pi^{=}(\Delta) = \frac{1}{\frac{1}{K_2 g(\Delta)} + 1 + K_1 g(\Delta)}$
- $\Pi^{+}(\Delta) = \frac{K_1 g(\Delta)}{\frac{1}{K_2 g(\Delta)} + 1 + K_1 g(\Delta)}$

Pour rappel, modèle Parent-Bernier

Conditions de symétrie (sans avantage jeu à domicile)

- $\Pi^{+}(\Delta) = \Pi^{-}(-\Delta)$
- $\Pi^{=}(\Delta) = \Pi^{=}(-\Delta)$
- Si on choisit $g(\Delta) = exp(\Delta)$, $K_1 \times K_2 = 1$

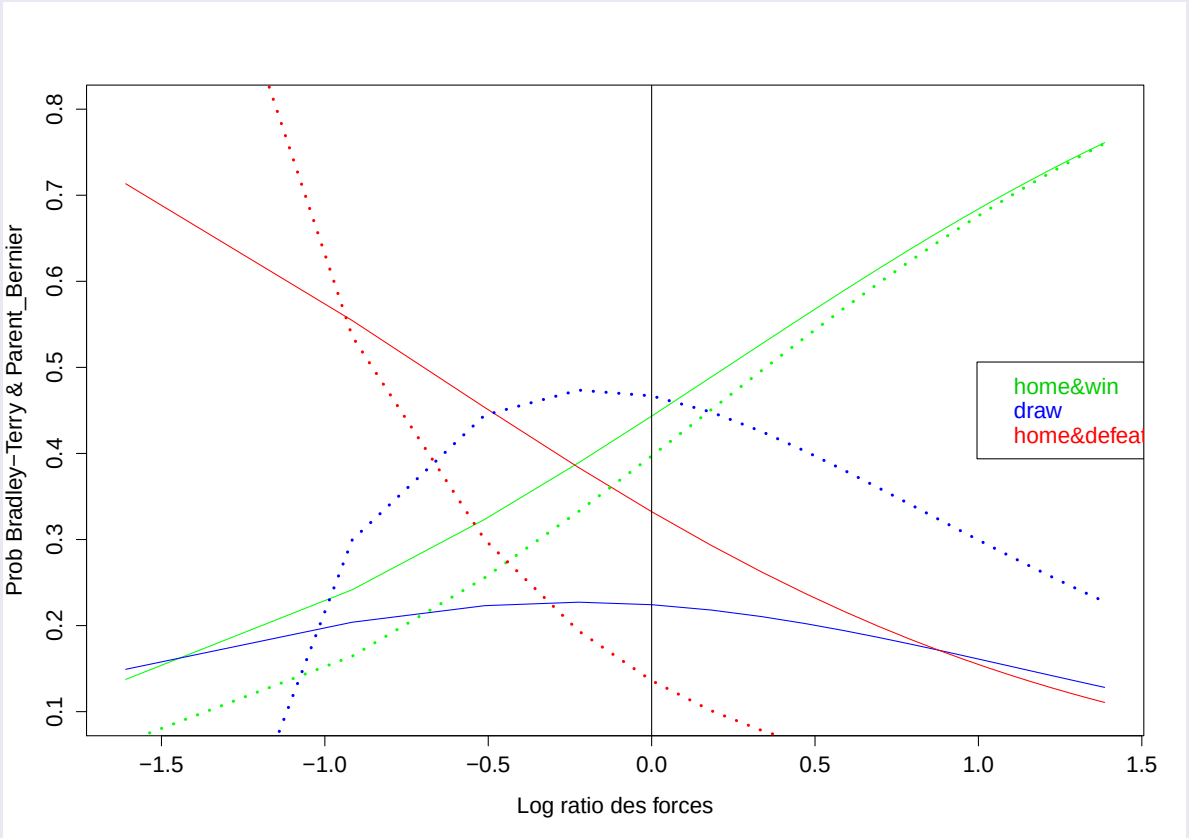
Modèle Parent-Bernier

- $\Pi^{+}(\Delta_{ij}) = \frac{Kexp(\Delta_{ij})}{Kexp(-\Delta_{ij})+1+Kexp(\Delta_{ij})}$
- $\Pi^{=}(\Delta_{ij}) = \frac{1}{Kexp(-\Delta_{ij})+1+Kexp(\Delta_{ij})}$
- $\Pi^{-}(\Delta_{ij}) = \frac{Kexp(-\Delta_{ij})}{Kexp(-\Delta_{ij})+1+Kexp(\Delta_{ij})}$

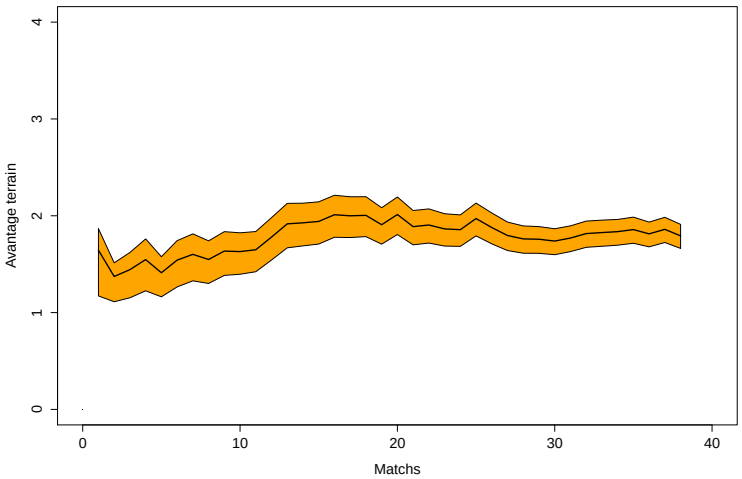
Le dénominateur de ce modèle ne donne pas une écriture du type Bradley-Terry-Caron-Doucet...mais se factorise et permettrait la même astuce de data augmentation

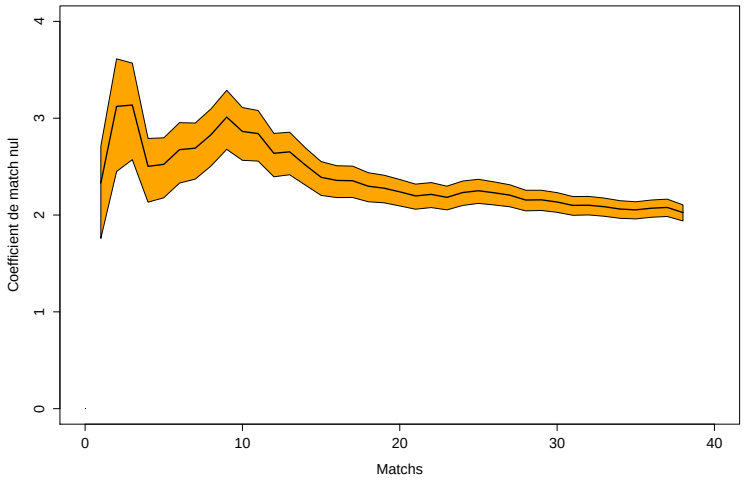
Quel modèle ?

Comportement Bradley-Terry versus Parent-Bernier

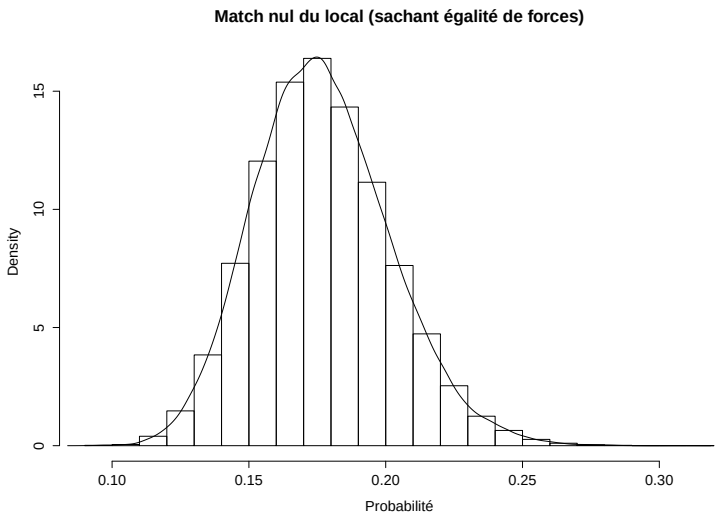


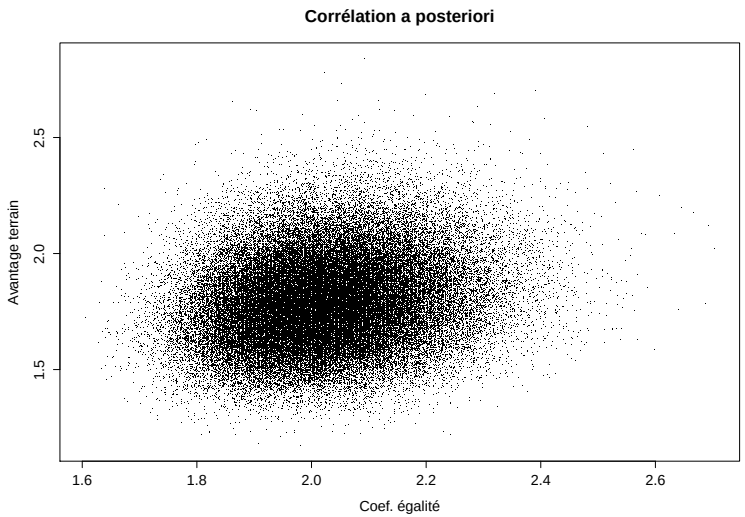
Résultats d'inférence sur les paramètres δ et θ de réglage





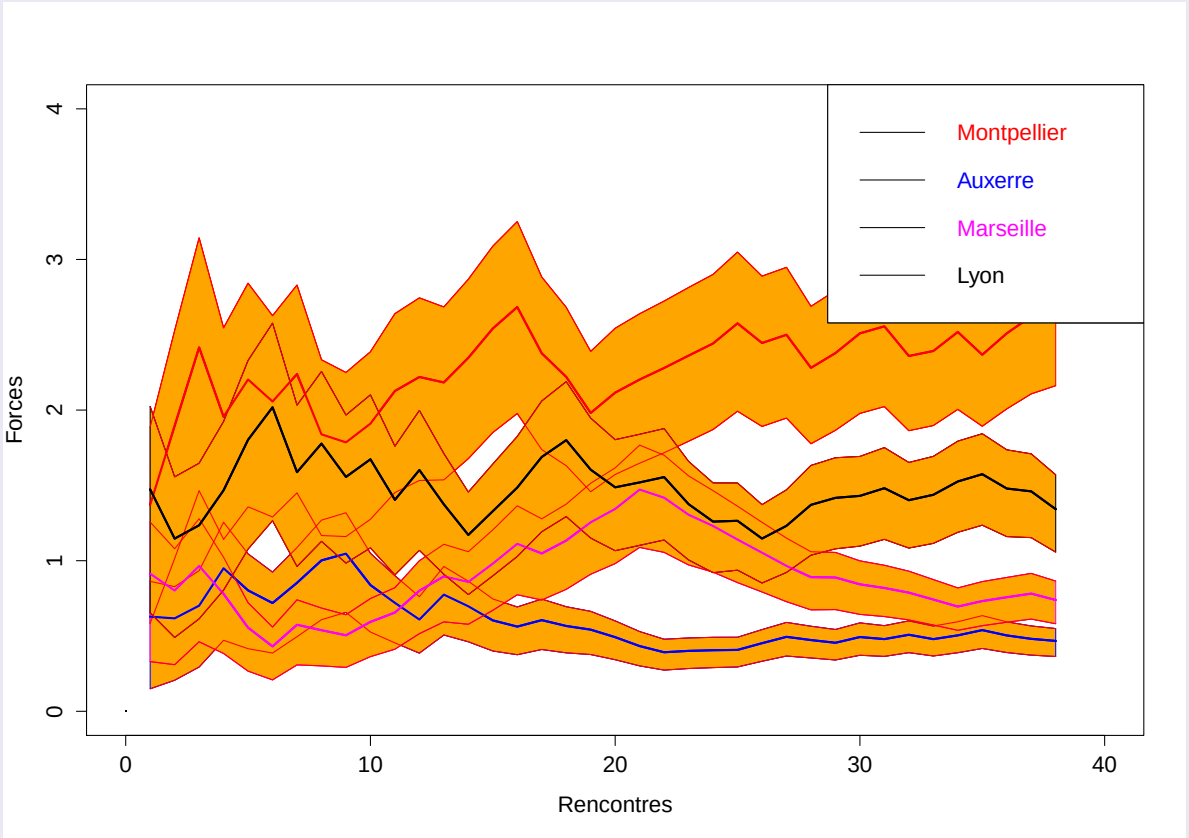
Résultats d'inférence sur les paramètres δ et θ de réglage (dernière journée)



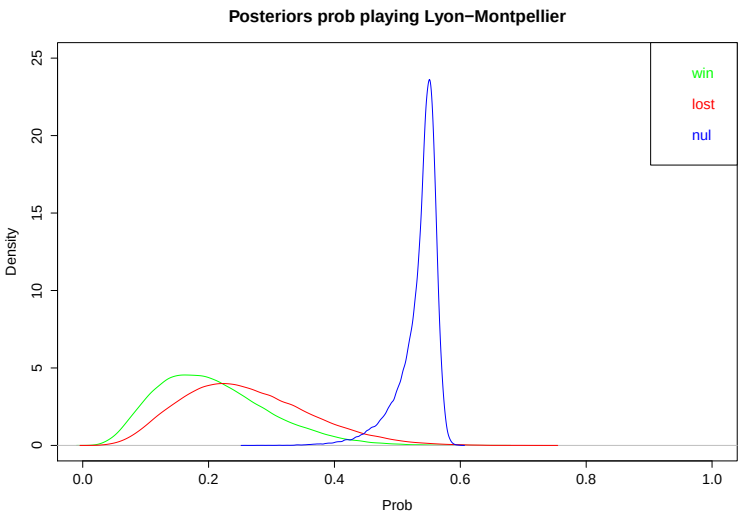


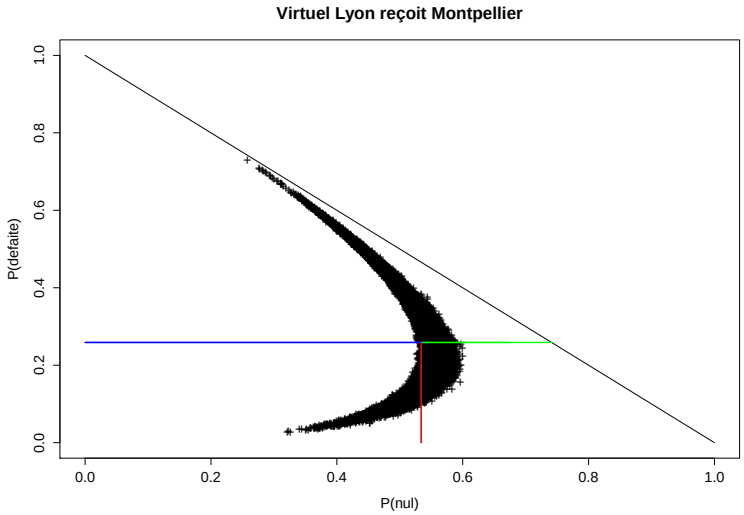
Résultats d'inférence sur la force des équipes

Estimation des λ_i au cours du temps



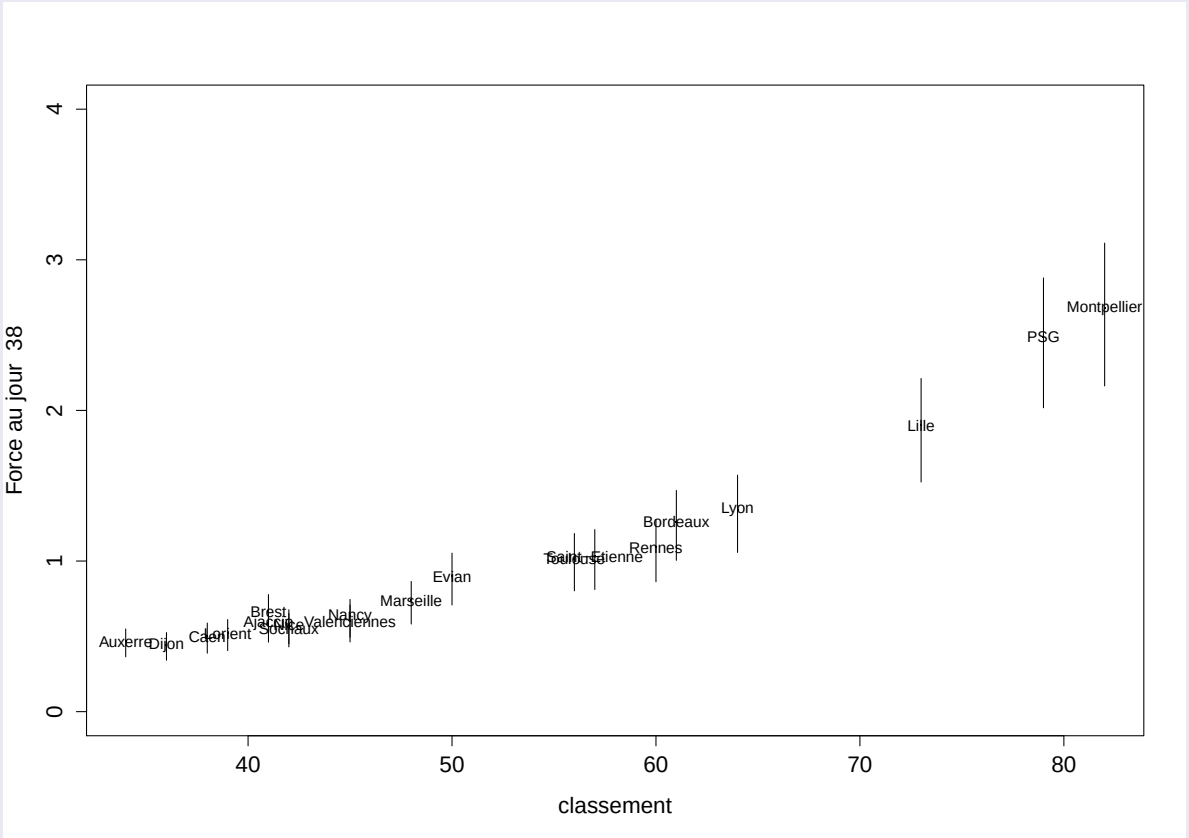
Résultats prédictifs Lyon recoit Montpellier (après la dernière journée)





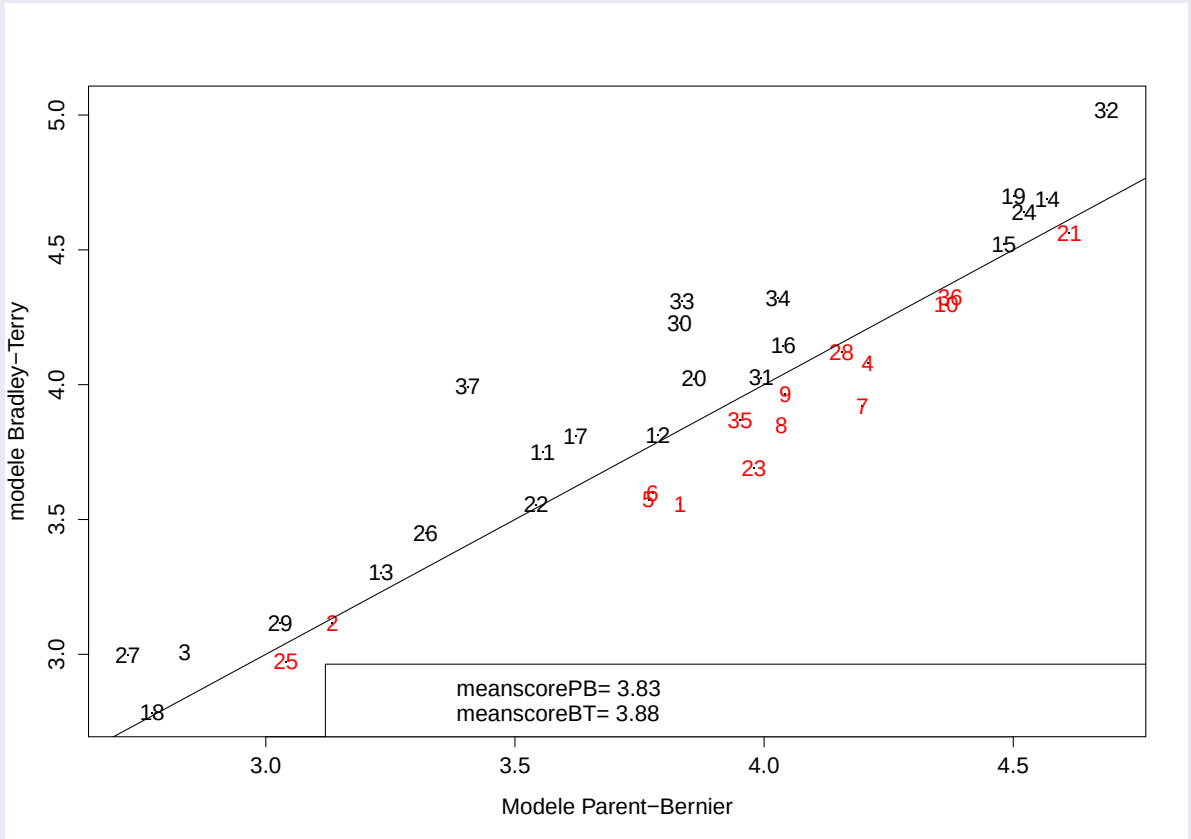
Le classement versus l'estimation statistique de la force des équipes

Dernière journée



Un bayésien peut-il gagner au lotofoot ?

Nombre de matchs (sur 10) prévus correctement



Conclusions : il reste du travail !

Temps de calcul

- Facteur $\times 3$ entre augmentation de données+parameter extension et inférence brutale

Part de bons résultats

- Moyenne de 40% de matchs correctement prévus sur tout le championnat
- Grimpe parfois à 50% sur la seconde partie du championnat

Améliorations possibles

- Rechercher l'information : expertise a priori, annees précédentes
- Changer le modèle : nb buts marqués, etc
- Covariables de composition des équipes

Extension au Fighting System du JuJitsu

Données

- 3 phases de combats (Atemi (pieds-poings), Judo debout, Judo sol)
- 3 forces en présence pour chacun des combattants $\lambda_i = (\lambda_i^1, \lambda_i^2, \lambda_i^3)$
- 3 manières de gagner un combat
- ...à faire

Références Bibliographiques


R.A. Bradley and M.E. Terry.
Rank analysis of incomplete block designs, 1. the method of paired comparisons.
Biometrika, 39 :324Ű-345, 1952.


F. Caron and A. Doucet.
Efficient Bayesian inference for generalized bradley-terry models.
Research Report 7445, INRIA, 2010.


H.A. David.
The Method of Paired Comparisons.
Oxford University Press, New York, 1988.


M.E. Glickman.
Parameter estimation in large dynamic paired comparison experiments.
JRSB Appl. Statist., 48(3) :377–394, 1999.


E. Parent and J Bernier.
Le Raisonnement Bayésien : Modélisation et Inférence.
Springer France, Paris, 2007.


M. H. Tanner.
Tools for Statistical Inference : Observed Data and Data Augmentation Methods.
Springer-Verlag, New York, 1992

C’est fini

Merci