UNE ALTERNATIVE BAYÉSIENNE EMPIRIQUE À UN ALGORITHME DE TYPE PAGERANK UTILISÉ EN BIBLIOMÉTRIE

Jean-Louis Foulley[1], Julie Josse[2] & Gilles Celeux[3]

[1]*92160 Antony, foulleyjl@gmail.com*
[2]*CMAP, X-Paris-Saclay, 91128 Palaiseau Cedex, julie.josse@polytechnique.edu*
[3]*INRIA-Select, 91405 Orsay Cedex, Gilles.Celeux@math.u-psud.fr*

Résumé. Face aux critiques soulevées par le facteur d'impact (FI) d'évaluation des revues, de nouveaux indices bibliométriques ont vu le jour qui s'appuient sur les citations croisées entre revues et en particulier sur la matrice d'adjacence d'un réseau orienté citant-cité qui est analysée grâce à un algorithme de type Google PageRank (PR): cf notamment « EigenFactor » (EF) de Bergstrom et « SCImago Journal Rating » (SJR) de Gonzalez-Pereira et Moya-Anegon. L'algorithme PR se base sur un lissage de la matrice adjacente combinant une marche aléatoire markovienne dans le réseau des données et une téléportation générale vers l'ensemble des nœuds avec des probabilités respectives ($\alpha = 0.85$ et $1-\alpha = 0.15$) fixes. Nous reprenons cette phase de lissage selon une approche bayésienne empirique qui repose sur un modèle Dirichlet-Multinomiale prenant en compte la condition d'exclusion des autocitations (termes diagonaux de la matrice). L'expression de la matrice de transition lissée s'obtient alors par l'espérance de la distribution a posteriori des paramètres des lois multinomiales sachant ceux des lois Dirichlet. Son expression est proche de celle de l'algorithme PR mais avec un coefficient α qui dépend maintenant de chacune des revues. Nous proposons de remplacer ces derniers paramètres par leur estimation du maximum de vraisemblance marginale. Celle-ci peut s'obtenir par divers procédés notamment grâce à un algorithme de Levenberg-Marquardt ou par simulation MCMC des lois marginales. Cette méthode a le mérite de reprendre la formule de base de PR tout en s'appuyant sur une modélisation probabiliste qui fait bien la distinction entre zéros structuraux (autocitations sur la diagonale) et zéros d'échantillonnage (termes hors diagonale) contrairement aux autres méthodes.

Mots-clés. Bibliométrie, Évaluation des revues, PageRank, Réseaux, Dirichlet-Multinomiale.

Abstract. New journal influence scores have been developed following criticisms against the Journal Impact Factor (JIF). These scores are based on cross-citations among journals using the adjacency matrix of the citing-cited network, the data of which are handled via a PageRank type algorithm: see eg Bergstrom's EigenFactor (EF) and Gonzalez-Pereira & Moya-Anegon's SCimago Journal Rating (SJR). The PR algorithm performs a smoothing of the transition matrix combining a random walk on the data network and a teleportation to all possible nodes with fixed probabilities ($\alpha = 0.85$ et $1-\alpha = 0.15$) respectively. We reconsidered this smoothing step in an Empirical Bayes perspective using a Dirichlet-Multinomial model with self citations excluded to avoid overvalued journal bias. The smoothed matrix can then be expressed by the mean of the posterior distribution of the true multinomial parameters given the Dirichlet parameters. A plug-in approach is proposed by replacing these parameters by their maximum marginal likelihood estimations. These can be obtained in several ways in particular by a Levenberg-Marquardt algorithm or via MCMC simulation of the marginal distributions. This procedure ends up with a formula similar to PR but with an α coefficient varying now according to each journal. It also makes a clear distinction between structural (self citations on the diagonal excluded) and sampling zeroes (off diagonal terms).

Keywords. Bibliometrics, Journal Rating, PageRank, Networks, Dirichlet-Multinomial.

1 Introduction

L'évaluation et le classement des revues à partir d'indices quantitatifs constituent une pratique ancienne (Gross and Gross, 1927), mais qui a pris tout son essor avec l'introduction du facteur d'impact (FI) qui mesure la notoriété d'une revue par le nombre moyen de citations annuelles qu'elle reçoit par article pendant une période antérieure donnée (Garfield, 1972). La publication systématique de FI par Clarivate Analytics (ex Thomson-Reuters ex ISI) dans « Journal of Citation Reports » ponctue et impacte grandement tous les secteurs de la vie et de la politique scientifiques à tous les échelons d'activité et de décision. FI génère une concurrence acerbe entre les revues par la hiérarchie et l'élite de prestige qu'il crée parmi celles-ci et les chercheurs, unités et institutions qui y contribuent. Cependant, l'objectif de FI est contestée: aucune prise en compte de l'appréciation critique (positive, neutre ou négative) des citations ; forte dépendance du champ disciplinaire ; fenêtre de citations d'articles trop restreinte (2 ans) ; distribution très asymétrique du nombre de citations d'un article mal prise en compte par la moyenne ; influence néfaste des autocitations et poids égal attribué à chaque citation quelle que soit sa provenance. Diverses alternatives ont été proposées pour faire face à ces critiques: allongement de la fenêtre de citation ; normalisation par champ disciplinaire, etc…(Zitt et Small, 2008). Nous nous intéresserons ici aux méthodes de prise en compte de l'importance des sources citantes et plus particulièrement à celles dérivées de l'algorithme de Google PageRank (Waltman et van Eck, 2010). Nous nous restreindrons également à celles qui excluent les autocitations telles que celle de l'EigenFactor (EF) pour pallier aux biais de politiques incitatives de certaines revues et aux effets néfastes d'une forme de consanguinité intellectuelle. En un premier temps, nous rappellerons les différentes étapes d'élaboration d'un facteur tel qu'EF. Puis, nous montrerons comment cette construction peut être aménagée dans une perspective bayésienne empirique plus rigoureuse. Enfin nous esquisserons les contours d'une application au cas concret d'une évaluation des revues de statistique.

2 Construction d'un score d'influence de type PageRank

Le matériau de base de cette construction réside dans la matrice carrée -classiquement notée $\mathbf{C}$ - des citations croisées entre les N revues comparées et telle que l'élément ij $[\mathbf{C}]_{ij} = c_{ij}$ représente le nombre de références produites par la revue i lors d'une année donnée, à des articles publiés par la revue j lors d'une période de temps précédente (2,3 ou 5 ans par exemple). À partir de cette matrice, on peut définir un réseau orienté pondéré citant→cité grâce à la matrice d'adjacence $\mathbf{P}$ telle que $[\mathbf{P}]_{ij} = p_{ij} = c_{ij} / c_{i+}$.

Ce type de normalisation est au cœur de l'algorithme PageRank (PR) qui effectue un lissage $\mathbf{G}$ de cette matrice $\mathbf{P}$ par une combinaison linéaire de celle-ci et d'une matrice dite de téléportation $\mathbf{1}\boldsymbol{\pi}^T$ avec des probabilités α et $1-\alpha$.

$$\mathbf{G} = \alpha\mathbf{P} + (1-\alpha)\mathbf{1}\boldsymbol{\pi}^T \qquad (1)$$

En présence de nœuds ballants (« dangling nodes »), c'est-à-dire de revues qui reçoivent des citations mais n'en citent aucune autre des $N-1$ du réseau, on remplacera la ligne i nulle correspondante à chacun des ces nœuds par le vecteur $\boldsymbol{\pi}^T$ (Langville et Meyer, 2006). Cela revient à substituer à $\mathbf{P}$, la matrice $\mathbf{P}^{\#} = \mathbf{P} + \mathbf{b}\boldsymbol{\pi}^T$ où $\mathbf{b} = (b_i)_{1\le i\le N}$ avec $b_i = 1$ si i est ballant et 0 dans le cas contraire. Cela permet de garantir l'existence d'une chaîne de Markov à temps discrets, irréductible et apériodique entre les N nœuds (ici les revues) illustré par l'image du fameux « random sufer » de Google qui, partant de i, avec une probabilité α se meut dans le réseau des données selon une marche aléatoire de probabilité p_{ij} et avec une probabilité $1-\alpha$ se téléporte au hasard vers n'importe quel autre nœud avec une probabilité π_j indépendante du nœud i où il se trouve.

L'algorithme PR tel qu'il est défini dans l'article d'origine (Brin and Page, 1998) est un algorithme récursif qui va hiérarchiser les nœuds en fonction d'une part de leur fréquence de pointage g_{ij} vers

ces nœuds, mais aussi de leur importance propre r_i selon l'expression :

$$r_j^{(n+1)} = \sum_{i=1}^{N} g_{ij} r_i^{(n)} \quad j = 1, 2, ..., N . \qquad (2).$$

Dans notre cas, cela signifie que dans le calcul du score de la revue j, la contribution de la revue i qui la cite est égale à la fréquence g_{ij} avec laquelle i cite j, mais en pondérant cette fréquence par l'influence propre r_i de i. Une citation donnée par une revue de prestige telle que JRRS-B ou JASA n'aura pas le même poids qu'une citation provenant d'une autre revue. Comme les scores d'influence sont inconnus, on procède par itération selon la méthode dite des puissances. En fait la valeur limite de (2) correspond à la distribution stationnaire de la chaîne de Markov définie par la matrice de transition décrite en (1) indépendamment de l'état initial.

C'est le vecteur propre gauche $\mathbf{r}^T = \mathbf{r}^T \mathbf{G}$ normé $(\mathbf{r}^T \mathbf{1} = 1)$ de $\mathbf{G}$ associée à la valeur propre dominante de cette matrice, soit ici l'unité en tant que valeur propre unique, compte tenu de la propriété d'irréductibilité et d'apériodicité de la chaîne (Fouss et al, 2016).

En pratique, cela veut dire que si la matrice $\mathbf{G}$ est connue, c'est-à-dire la matrice lissée de transition relative à $\mathbf{C}$, on peut alors établir une hiérarchie des revues sur la base de leur importance.

Pour revenir à l'image du « random surfer » de Google, l'importance des nœuds correspond à la fréquence limite avec laquelle ces nœuds sont visités lors d'une marche très longue sur le réseau d'un grand nombre de surfers partis de nœuds quelconques.

Dans la méthode EF, les autocitations sont exclues $\left(c_{ii} = 0, p_{ii} = 0, \forall i = 1, 2, ..., N\right)$ et le vecteur $\boldsymbol{\pi}$ n'est pas pris égal à $\mathbf{1}/\mathrm{N}$ comme dans PageRank (équiprobabilité de tous les nœuds), mais égal aux fréquences de références publiées par chacune des revues dans la fenêtre de temps considérée (5 ans pour EF) soit $\boldsymbol{\pi} = \left(\tilde{a}_i\right)_{1 \le i \le N}$ où $\tilde{a}_i = a_i / a_+$, a_i étant le nombre de références publiées par i. De plus, le score EF n'est pas exactement égal au vecteur propre $\mathbf{r}$ associée à $\mathbf{G}$. Pour se préserver des autocitations, Bergstrom (2007) définit le score à partir de la solution de (2) comme suit (cf www.eigenfactor.org/methods.pdf) :

$$r_j^* = \sum_{i=1}^{N} p_{ij} r_i \ , \ EF_j = r_j^* / \sum_{k=1}^{N} r_k^* . \qquad (3)$$

Cet indice est donc une forme de compromis empirique entre l'évaluation PR et celle originale de Pinski et Narin (1976) appliquée à une matrice $\mathbf{P}$ sans autocitations. En fait, on montre aisément, nœuds ballants mis à part, que $EF = \left[\mathbf{r} - (1-\alpha)\boldsymbol{\pi}\right] / \alpha$.

Même si cette formulation paraît opérationnelle, il nous est apparu opportun d'étudier dans quelle mesure on ne pourrait pas la rationaliser davantage tout en conservant les contraintes fortes de départ (autocitations exclues) et la simplicité de son expression. Nous ne reviendrons pas sur le cas où les autocitations sont prises en compte, car il ressort d'un traitement classique dans une approche Dirichlet Multinomiale simple (cf. par exemple Wang et al, 2008).

3 Une alternative bayésienne empirique à la matrice PR excluant la diagonale

Soit $\underline{\mathbf{C}}_i^T = \left(c_{ij}\right)$ for $j \neq i = 1, 2, ..., N$, la ième ligne de la matrice $\mathbf{C}$ privée de ses éléments diagonaux soit $\underline{\mathbf{C}}$ de dimension $\left(N \text{ x } N-1\right)$ s'écrit $\underline{\mathbf{C}}^T = \left[\underline{\mathbf{C}}_1, ... \underline{\mathbf{C}}_i, .., \underline{\mathbf{C}}_N\right]$.

Nous adopterons l'approche probabiliste hiérarchique suivante en deux étapes :

I) Échantillonnage multinomiale des éléments de $\underline{\mathbf{C}}_i^T$ soit

$$\mathbf{C}_i^T \mid \boldsymbol{\theta}_i \sim_{id} \mathcal{M}\left(n_i, \boldsymbol{\theta}_i^T\right), \qquad (4)$$

de paramètres $n_i = \sum_{j \neq i} c_{ij}$ et de vecteur de probabilités $\underline{\boldsymbol{\theta}}_i = \left(\theta_{i1}, ..., \theta_{i,i-1}, \theta_{i,i+1}, .., \theta_{i,N}\right)^T$, ces lois étant supposées indépendantes d'une revue citante (ligne) à l'autre.

II) A priori Dirichlet

En un deuxième temps, les paramètres de la distribution en (6) sont supposés classiquement suivre

des lois de Dirichlet indépendantes entre lignes

$$\boldsymbol{\theta}_i \mid \boldsymbol{\gamma}_{\setminus i}^T \sim_{id} \mathcal{D}\left(\boldsymbol{\gamma}_{\setminus i}^T\right), \tag{5}$$

où $\boldsymbol{\gamma}_{\setminus i}^T = \left(\gamma_1, \gamma_2, ..., \gamma_{i-1}, \gamma_{i+1} ..., \gamma_{iN}\right)$ est le sous vecteur de $\boldsymbol{\gamma}^T = \left(\gamma_k\right)_{1 \le k \le N}$ privé de son ième élément.
Du fait de la propriété de conjugaison de ces lois, l'a posteriori est aussi de Dirichlet

$$\underline{\boldsymbol{\theta}}_i \mid \boldsymbol{\gamma}_{\setminus i}^T, \underline{\mathbf{C}} \sim \mathcal{D}\left(\underline{\mathbf{C}}_i^T + \boldsymbol{\gamma}_{\setminus i}^T\right). \tag{6}$$

On en déduit aisément l'expression de l'espérance a posteriori

$$E_P\left(\underline{\boldsymbol{\theta}}_{ij}\right) = \frac{c_{ij} + \gamma_j}{\sum_{j \ne i} c_{ij} + \sum_{j \ne i} \gamma_j} \text{ pour } j \ne i = 1, 2, ..., N. \tag{7}$$

En posant : $\mathbf{p}_i = \left(p_{ij}\right)_{1 \le j \le N}$, $\alpha_i = n_i / \left(n_i + K_{\setminus i}\right)$ où $K_{\setminus i} = K - \gamma_i$ avec $K = \sum_{i=1}^N \gamma_i$ et $\boldsymbol{\pi}_i^* = \left(\pi_{ij}^*\right)_{j \ne i}$, $\pi_{ij}^* = \gamma_j / K_{\setminus i}$ soit aussi $\pi_{ij}^* = \pi_j / \left(1 - \pi_i\right)$ où $\pi_j = \gamma_j / K$, la ligne i exprimée par l'espérance a posteriori s'écrit :

$$\mathbf{G}_i^{*T} = \alpha_i \mathbf{p}_i^T + \left(1 - \alpha_i\right) \boldsymbol{\pi}_i^T, \tag{8}$$

qui fait le pendant de (1)

$$\mathbf{G}_i^T = \alpha \mathbf{p}_i^T + \left(1 - \alpha\right) \boldsymbol{\pi}^T. \tag{9}$$

On conserve ainsi la même forme d'expression, mais cette fois, le facteur de rétrécissement α n'est plus fixe, mais dépend à la fois du nombre de références n_i produites par chaque revue et d'un paramètre $K_{\setminus i}$ lui-même variable d'une revue à l'autre et fonction des paramètres γ_j. Ce coefficient varie entre 0 et 1. Il est d'autant plus élevé que n_i est élevé et que $K_{\setminus i}$ est faible (variance plus grande de l'a priori Dirichlet) ce qui fait sens et tend à la limite vers 1 quand l'une ou l'autre de ces informations devient prédominante et à l'inverse tend zéro dans le cas contraire. Le cas limite de $n_i = 0$ soit $\alpha_i = 0$ est intéressant puisqu'il correspond à un nœud ballant. Alors la ligne correspondante $\mathbf{G}_i^{*T}$ s'identifie aux probabilités a priori $\boldsymbol{\pi}_i^T$. L'ajustement est alors automatique contrairement à ce qui se passe avec PR.

Diverses options se présentent à ce stade. La première correspond à la situation où la loi Dirichlet en (6) est complètement spécifiée (paramètres connus). C'est le cas notamment des a priori dits « non informatifs » tels que ceux de

-Bayes-Laplace : $K = N$ et $\boldsymbol{\gamma} = \mathbf{1}$,
-Jeffreys : $K = N/2$ et $\boldsymbol{\gamma} = \mathbf{1}/2$,
-Perks: $K = 1$ et $\boldsymbol{\gamma} = \mathbf{1}/N$.

Les avis sont partagés sur les mérites de ces différents a priori (Berger et al, 2015 vs. Tuyl, 2016) notamment en présence d'un nombre important de cellules nulles ce qui était précisément l'objectif principal d'une régularisation de type PR. Dans tous les cas, c'est une solution d'équiprobabilité de la téléportation retenue par PR à l'origine, mais qui n'est pas été retenue en bibliométrie.
Dans un cadre bayésien empirique standard, les paramètres de l'a priori Dirichlet $\boldsymbol{\gamma}$ sont remplacés par l'estimation du maximum de vraisemblance de la loi marginale $m\left(\underline{\mathbf{C}} \mid \boldsymbol{\gamma}\right)$ après intégration des paramètres de probabilité des multinomiales. Cette loi marginale est connue, sa densité est le produit de densités Multinomiales à composantes Dirichlet:

$$m\left(\underline{\mathbf{C}} \mid \boldsymbol{\gamma}\right) = \prod_{i=1}^N m_i\left(\underline{\mathbf{C}}_i^T \mid \boldsymbol{\gamma}_{\setminus i}\right), \tag{10}$$

où

$$m_i\left(\underline{\mathbf{C}}_i^T \mid \boldsymbol{\gamma}_{\setminus i}\right)=\frac{n_i!}{\prod_{j\neq i}c_{ij}!}\frac{\Gamma\left(\sum_{j\neq i}\gamma_j\right)}{\Gamma\left(\sum_{j\neq i}\left(c_{ij}+\gamma_j\right)\right)}\prod_{j\neq i}\frac{\Gamma\left(c_{ij}+\gamma_j\right)}{\Gamma\left(\gamma_j\right)}. \quad (11)$$

On peut maximiser la logvraisemblance par un algorithme adéquat tel que, par exemple, celui de Levenberg-Marquardt :

$$\left[\mathbf{H}\left(\boldsymbol{\gamma}^{(t)}\right)+\lambda^{(t)}Diag\left(\mathbf{H}\left(\boldsymbol{\gamma}^{(t)}\right)\right)\right]\left(\boldsymbol{\gamma}^{(t+1)}-\boldsymbol{\gamma}^{(t)}\right)=-\nabla L\left(\boldsymbol{\gamma}^{(t)}\right), \quad (12)$$

où $\mathbf{H}\left(\boldsymbol{\gamma}\right)=\frac{\partial^2 L\left(\boldsymbol{\gamma}\right)}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^T}$ désigne la $\left(N\text{x}N\right)$ matrice hessienne et $\nabla L\left(\boldsymbol{\gamma}\right)=\frac{\partial L\left(\boldsymbol{\gamma}\right)}{\partial\boldsymbol{\gamma}}$ le $\left(N\text{x}1\right)$ vecteur gradient des dérivées premières. $\lambda^{(t)}$ est un facteur réducteur de valeur décroissante si $L\left(\boldsymbol{\gamma}\right)$ augmente et croissante dans le cas contraire. Si $\lambda^{(t)}=0$, l'algorithme se réduit à celui de Newton-Raphson.

Le gradient s'écrit :

$$\nabla L\left(\boldsymbol{\gamma}\right)=\sum_{i\neq j}\left[\psi\left(K_{\setminus i}\right)\right]-\left(N-1\right)\psi\left(\gamma_j\right)+\sum_{i\neq j}\left[\psi\left(c_{ij}+\gamma_j\right)-\psi\left(n_i+K_{\setminus i}\right)\right], \quad (13)$$

où $\psi\left(x\right)=d\log\Gamma\left(x\right)/dx$ est la fonction digamma et $K=\sum_{j=1}^{N}\gamma_j$.

Cette expression mérite l'attention car on la retrouve de façon quasi-identique si l'on développe une version EM de l'estimation ML des paramètres $\boldsymbol{\gamma}$ qui utilise comme données manquantes les paramètres des lois multinomiales. On a alors à maximiser la fonction suivante :

$$Q\left(\boldsymbol{\gamma};\boldsymbol{\gamma}^{(t)}\right)=\sum_{i=1}^{N}\mathrm{E}_c^{(t)}\left[\log p\left(\underline{\boldsymbol{\theta}}_i \mid \boldsymbol{\gamma}_{\setminus i}\right)\right], \quad (14)$$

où $\mathrm{E}_c^{(t)}\left[.\right]$ désigne une espérance conditionnelle prise par rapport à la distribution de $\underline{\boldsymbol{\theta}}$ sachant les données $\underline{\mathbf{C}}$ et les valeurs courantes des paramètres $\boldsymbol{\gamma}^{(t)}$. La dérivée de (14) présente une grande similarité avec (13),

$$\frac{\partial Q\left(\boldsymbol{\gamma};\boldsymbol{\gamma}^{(t)}\right)}{\partial\gamma_j}=\sum_{i\neq j}\left[\psi\left(K_{\setminus i}\right)\right]-\left(N-1\right)\psi\left(\gamma_j\right)+\sum_{i\neq j}\left[\psi\left(c_{ij}+\gamma_j^{(t)}\right)-\psi\left(n_i+K_{\setminus i}^{(t)}\right)\right]. \quad (15)$$

En conséquence, l'algorithme EM est très proche de l'algorithme de maximisation standard à la matrice hessienne près, plus simple avec EM. Ce distinguo subtil avait été déjà noté par certains auteurs (Minka, 2012). D'autres algorithmes mériteraient d'être envisagés dans le cas Dirichlet-Multinomiale tel que l'algorithme MM de Minoration-Maximisation (Zhou et Zhang, 2012) ou celui itératif d'un point fixe.

L'application concerne la matrice $\mathbf{C}$ des citations croisées entre 47 revues statistiques, établie et étudiée par Varin et al (2016) (citations parues en 2010 relatives à des articles publiés de 2001 à 2010). L'estimation du maximum de vraisemblance du paramètre K de concentration s'établit, quel que soit l'algorithme, à $K=58.10\pm2.82$ avec une variation importante entre revues des valeurs des γ_j allant de 6.61±0.54 pour JASA à 0.06±0.03 pour « The Stata Journal ». Notons que l'ignorance des autocitations par un traitement simple de ces données comme des zéros d'échantillonnage aboutit à un biais non négligeable d'estimation $\left(K=49.00\right)$ et qui n'est pas qu'un effet d'échelle.

4 Discussion-Conclusion

L'estimation ML des paramètres $\boldsymbol{\gamma}$ peut aussi être obtenue comme un sous-produit d'une approche bayésienne hiérarchique comportant une étape supplémentaire de spécification d'a priori non

informatifs adéquats sur les paramètres $\boldsymbol{\gamma}$ et de calcul du mode a posteriori des distributions marginales correspondantes. On pourrait aussi reprendre la formule adoptée par Bergstrom (2007) pour le calcul d'EF soit $\boldsymbol{\pi} = \left(\pi_j = \tilde{a}_j\right)_{1 \le j \le N}$ tel que défini en (9) et estimer ensuite $K = \sum_{j=1}^{N} \gamma_j$ par maximum de vraisemblance à $\boldsymbol{\pi}$ fixé. Cela aurait le mérite de rapprocher les deux méthodes et de faciliter l'application de la nouvelle. Une autre possibilité très simple serait de considérer les fréquences de citations reçues par chacune des revues soit $\boldsymbol{\pi} = \left(c_{+j} / c_{++}\right)_{1 \le j \le N}$.

La méthode développée précédemment constitue une extension de l'algorithme de type PR utilisé en vue de l'établissement de L'EigenFactor selon un modèle probabiliste bien établi (modèle Dirichlet-Multinomiale) qui prend en compte de façon rigoureuse la contrainte d'exclusion des autocitations. Le lissage de la matrice adjacente correspondant au réseau orienté citant→cité s'obtient comme dans PR par une combinaison convexe ligne à ligne du vecteur correspondant $\mathbf{p}_i$ des probabilités de transition observées et d'un vecteur de téléportation $\boldsymbol{\pi}_i$ selon des probabilités respectives $\alpha_i = n_i / \left(n_i + K_{\setminus i}\right)$ et $1 - \alpha_i = K_{\setminus i} / \left(n_i + K_{\setminus i}\right)$ qui varient d'une revue à l'autre en fonction du nombre total n_i de références et d'un coefficient de concentration $K_{\setminus i}$. D'autres méthodes sont envisageables pour établir un facteur d'influence à partir d'une matrice carrée de citations croisées telles que celles basées sur un modèle de Bradley-Terry (Stigler, 1994 ; Varin et al, 2016) ou plus généralement sur des modèles RC symétriques (Goodman, 2002 ; Grah, 2016). Outre sa simplicité conceptuelle et calculatoire, le PR bayésien développé ici a le mérite de bien prendre en compte et de distinguer les zéros de structure de ceux d'échantillonnage.

Références

Berger JO, Bernardo JM, Sun D (2015) Overall objective priors. *Bayesian Analysis*, 10, 189-246.

Bergstrom CT (2007) Eigenfactor: measuring the value and the prestige of scholarly journals. *Coll. Res. Lib. News*

Brin S, Page L. (1998) The anatomy of a large-scale hypertextual web search engine. Comput. Networks ISDN Syst., 30, 107–117.

Fouss F, Saerens M, Shimbo M (2016) Algorithms and Models for Network Data and Link Analysis, Cambridge University Press. NY.

Garfield E. (1972) Citation analysis as a tool in journal evaluation. *Science*, 178, 471–479.

González-Pereira B, Guerrero-Bote, VP, Moya-Anegón F. (2009).The SJR indicator: A new indicator of journals' scientific prestige. *Journal of Informetrics* 4 (2010) 379–391

Goodman LA (2002) Contributions to the statistical analysis of contingency tables: Notes on quasi-symmetry, quasi-independence, log-linear models, log bilinear models, and correspondence analysis models. *Annales de la Faculté des Sciences de Toulouse*, 6ème série, 11, 525-540.

Grah S (2016) Ranking and rating of scientific journals. Rapport de stage Master 1, Paris Sud Orsay

Gross PLK, Gross EM. (1927) College libraries and chemical education. *Science*, **66**, 385–389.

Langville AN, Meyer CD (2006) Google's PageRank and Beyond. The Science of Search Engine Rankings. Princeton University Press, Princeton.

Minka TP (2012) Estimating a Dirichlet distribution. Document web

Pinski G, Narin F (1976) Citation influence for journal aggregates of scientific publications: theory with application to the literature of Physics. *Information Processing & Management*, 12, 297-312.

Stigler SM (1994) Citations Patterns in the Journals of Statistics and Probability. *Statistical Science*, 9, 94-108

Tuyl F (2016) A note on priors for the multinomial model. *The American Statistician*. Accepted

Varin C, Cattelan M, Firth D (2016) Statistical modelling of citation exchange between statistics journals. *Journal of the Royal Statistical Society* A,179,1-63

Zhou H, Zhang Y (2012) EM vs MM: A case study. *Computational Statistics & Data Analysis*, 56, 3909-3920.

Waltman L, van Eck NJ (2010) The relation between Eigenfactor, audience factor and influence weight. *Journal of the American Society for Information Science & Technology*, 61, 1476-1486

Wang X, Tao T, Sun JT, Shakery A, Zhai C (2008) DirichletRank : Solving the Zero-One Problem of Page Rank. *ACM Transactions on Information Systems*, 26,1-29

Zitt M, Small H. (2008) Modifying the journal impact factor by fractional citation weighting: the audience factor. *Journal of the. American. Society for Information. Science &. Technology*., 59, 1856–1860.