Modèles génératifs de rangs relatifs à un algorithme de tri par insertion

Christophe Biernacki, Julien Jacques

Modèles génératifs de rangs relatifs à un algorithme de tri par insertion

Christophe Biernacki & Julien Jacques

Laboratoire P. Painlevé, UMR 8524 CNRS, Université Lille I,
Bât. M2, Cité Scientifique, F-59655 Villeneuve d'Ascq Cedex, France.

Résumé. Les données de rang proviennent d'un processus de tri dont la nature est généralement inaccessible au statisticien. Faisant l'hypothèse que ce tri repose sur la comparaison entre paires d'objets et que par ailleurs le processus retenu vise à en minimiser le nombre, l'algorithme de tri par insertion s'impose comme l'un des meilleurs candidats. Par l'introduction d'une erreur de Bernoulli sur les comparaisons de paires, on obtient une modélisation générative probabiliste des données de rang dont une des originalités est de dépendre de l'ordre de présentation initiale des objets à classer. En fonction des hypothèses d'échantillonnage relatives à cet ordre de présentation (inconnu), plusieurs modèles réalistes sont obtenus. Des expériences numériques sur des données réelles permettent de comparer ces modèles avec le modèle standard Φ de Mallows.
Mots clés : algorithme de tri, ordre de présentation, vraisemblance intégrée.

Abstract. Rank data arise from a sorting mechanism which is generally unobservable for the statistician. Assuming both that this mechanism relies on paired-comparisons and that it aims to minimize their number, the insertion sorting algorithm is one of the best candidates. A Bernoulli event can be naturally introduce in the paired-comparison step, leading to an original probabilistic generative model for rank data which depends on the initial presentation order. According to this (unknown) initial rank, several reaslistic models are proposed. Then, experiments on real data sets compare these models with the usual Mallows Φ model.
Keywords : sorting algorithm, presentation order, integrated likelihood.

1 Introduction

Les données de rang sont très présentes dans les activités humaines impliquant des préférences ou des choix, typiquement le classement de pages web, les résultats sportifs ou économiques, des tests en tout genre, le domaine du marketing ou de la biologie, *etc.* Les rangs sont à ce point utiles pour l'Homme qu'il n'est pas rare qu'ils résultent d'une transformation d'autres types de données. Ce sont des données par nature multivariées mais hautement structurées, conduisant ainsi à l'élaboration de méthodes spécifiques pour leur visualisation ou leur modélisation probabiliste (voir par exemple Marden (1995) et les nombreuses références associées). Nous nous intéressons ici à ce dernier point.

L'unité statistique correspondant à un rang est un classement de m objets $\mathcal{O}_1, \ldots, \mathcal{O}_m$ par un juge (non nécessairement humain) selon un ordre de préférence que nous supposerons ici, sans perte de généralité, décroissant. Deux codages sont alors communément utilisés : le *classement* ou le *rangement*, désignés en anglais respectivement par *ranking* et *ordering*. Un *classement* $x^{-1} = (x_1^{-1}, x_2^{-1}, \ldots, x_m^{-1})$ contient les rangs de classement de chaque objet et signifie que l'objet $\mathcal{O}_1$ se trouve en position x_1^{-1}, l'objet $\mathcal{O}_2$ est en position x_2^{-1}, et ainsi de suite. Un classement est donc un élément de $\mathcal{P}_m$, l'ensemble des permutations des m premiers entiers. Le *rangement* $x = (x_1, x_2, \ldots, x_m)$ est aussi un élément de $\mathcal{P}_m$ et signifie que l'objet $\mathcal{O}_{x_1}$ est en 1ère position, l'objet $\mathcal{O}_{x_2}$ en seconde, *etc.*

Le modèle Φ de Mallows (Mallows, 1957) est certainement l'un des modèles génératifs les plus simples et les plus utilisés pour les données de rang. Notant μ un rangement de référence, il s'appuie sur la distance de Kendall τ entre ce dernier et n'importe quel rang x de $\mathcal{P}_m$:

$$pr(x; \mu, p) = \left\{1 - \frac{1-p}{p}\right\} \left\{\sum_{j=1}^{m-1} \left[1 - \left(\frac{1-p}{p}\right)^{m-j+1}\right]\right\}^{-1} \left(\frac{1-p}{p}\right)^{\tau(x,\mu)} \quad (1)$$

avec $0 < p < 1$ un paramètre indiquant la dispersion de la distribution autour de μ. La valeur $p = 1/2$ correspond à la loi uniforme sur $\mathcal{P}_m$, tandis que $p > 1/2$ donne une distribution de mode μ, de plus en plus concentrée autour de μ quand p augmente. Les valeurs $p < 1/2$ procèdent de façon symétrique avec le rangement $\bar{\mu}$ inverse de μ.

Le modèle Φ de Mallows peut être interprété à l'origine comme un processus de comparaison indépendante de toutes les paires d'objets deux à deux (Mallows, 1957), dans la limite de garantir la cohérence du classement obtenu. Alternativement, il peut être vu également comme le résultat d'un algorithme de *tri par sélection* particulier, où la notion de comparaison par paires est beaucoup moins explicite (Fligner and Verducci, 1986).

Nous proposons dans ce travail de revenir à la comparaison de paires d'objets comme étape élémentaire et fondamentale du processus de fabrication d'un rang. Dans ce contexte, il est naturel d'introduire de la variabilité dans le classement en modélisant le risque de réaliser une mauvaise comparaison entre deux paires d'objets. Le minimisation du nombre d'erreurs en moyenne repose donc sur une stratégie visant à minimiser le nombre de comparaisons de paires à réaliser, ce qui nous conduit à retenir un algorithme de *tri par insertion* si le nombre d'objets ne dépasse pas une dizaine (Knuth, 1973).

2 Modèle de tri par insertion

Comme de nombreux modèles de rangs, nous supposons tout d'abord qu'il existe un rangement de référence $\mu = (\mu_1, \ldots, \mu_m)$ sur les m objets, ce qui signifie qu'un juge parfait retournerait systématiquement ce rang de référence. Supposant que le juge cherche à minimiser le nombre de comparaisons par paires d'objets ($m \leq 10$), l'algorithme de tri par insertion est optimal. Partant d'un rangement initial $\sigma = (\sigma_1, \ldots, \sigma_m) \in \mathcal{P}_m$, cet

algorithme procède successivement de la façon suivante. L'objet $\mathcal{O}_{\sigma_1}$ est tout d'abord sélectionné. Puis le second objet $\mathcal{O}_{\sigma_2}$ est placé sa gauche et lui est comparé. Si l'objet $\mathcal{O}_{\sigma_2}$ n'est pas préféré, selon μ, à $\mathcal{O}_{\sigma_1}$ alors l'objet $\mathcal{O}_{\sigma_2}$ est décalé à la suite de $\mathcal{O}_{\sigma_1}$, donc sur sa droite. Sinon la position de $\mathcal{O}_{\sigma_2}$ est inchangée et on peut passer à l'introduction du nouvel objet $\mathcal{O}_{\sigma_3}$ à l'extême gauche des objets déjà classés. Le nouvel objet introduit est décalé vers la droite tant qu'il n'est pas préféré (selon μ) à l'objet déjà classé se trouvant à sa droite (s'il existe). Et ainsi de suite jusqu'à épuisement des objets à classer.

L'action fondamentale du juge repose sur la comparaison d'une paire d'objets et il est naturel d'y introduire une possibilité d'erreur en interprétant cet événement comme le résultat d'un expérience de Bernoulli dont le résultat est une comparaison correcte avec probabilité p ($0 < p < 1$) et erronée avec probabilité $1 - p$. De plus, il est raisonnable de supposer l'indépendance de chaque opération de comparaison par paires, ce qui conduit à la loi de probabilité suivante sur tout rang x :

$$pr(x|\sigma;\mu,p) = \prod_{j=1}^{m} p^{\eta_j(x,\sigma,\mu)}(1-p)^{\#j(x,\sigma)-\eta_j(x,\sigma,\mu)} \tag{2}$$

où j indique l'étape de l'algorithme de tri consistant à ranger l'objet $\mathcal{O}_{\sigma_j}$ ($1 \leq j \leq m$) et

- $\#j(x,\sigma) = \#\left\{\{i : x^{-1}_{\sigma_i} < x^{-1}_{\sigma_j}, 1 \leq i < j\} \bigcup \{i : i = \arg\min_{1\leq i'<j}\{i' : x^{-1}_{\sigma_{i'}} > x^{-1}_{\sigma_j}\}\}\right\}$ est le nombre total de comparaisons de paires à l'étape j ;
- $\eta_j(x,\sigma,\mu) = \sum_{i\in j(x,\sigma)} \delta_{\sigma_i\sigma_j}(\mu) + \delta_{\sigma_j\sigma_i}(\mu)$ est le nombre total de *bonnes* comparaisons à l'étape j, où $\delta_{ii'}(\mu) = \mathbf{1}\{\mu_i^{-1} < \mu_{i'}^{-1}\}$ est égal à 1 si $\mathcal{O}_i$ est correctement rangé avant $\mathcal{O}_{i'}$ et 0 sinon ($i, i' = 1, \ldots, m, i \neq i'$).

Lorsque le rang initial σ est inconnu et en fixant par défaut l'uniformité pour la distribution de cette donnée manquante, c'est-à-dire $pr(\sigma) = 1/m!$, on obtient la loi marginale des x par $pr(x;\mu,p) = \frac{1}{m!}\sum_{\sigma\in\mathcal{P}_m} pr(x|\sigma;\mu,p)$. Les propriétés de cette distribution ont été étudiées par Biernacki and Jacques (2010), parmi lesquelles l'uniformité si $p = 1/2$, l'unimodalité en μ si $p > 1/2$, la concentration de la masse autour de μ quand p augmente, une symétrie entre les couples (p,μ) et $(1-p,\bar{\mu})$ et l'identifiabilité.

3 Ordre initial et modèles d'échantillonnage

Une des principales originalités du modèle (2) est de prendre en compte l'ordre initial σ de présentation des objets. Lorsque l'on dispose d'un échantillon de rangements $(x^1, \ldots, x^n)$, l'estimation des paramètres μ et p du modèle repose nécessairement sur les hypothèses d'échantillonnage qui seront faites sur les couples de données (x^1,σ^1), $\ldots$, (x^n,σ^n), même si les σ^i sont des données manquantes. Dans ce cas on supposera comme précédemment que la loi est uniforme sur les σ^i.

Une première hypothèse est de supposer l'indépendance des n couples (x^i, σ^i), ce qui conduit aussi à l'indépendance des x^i et à maximiser la vraisemblance suivante :

$$L(\mu, p) \;=\; \prod_{i=1}^{n} \frac{1}{m!} \sum_{\sigma \in \mathcal{P}_m} pr(x^i|\sigma; \mu, p). \tag{3}$$

Un algorithme EM peut être utilisé avec profit grâce aux données manquantes σ^i (voir Biernacki and Jacques, 2010). Alternativement, les rangs σ^i peuvent être inconnus mais tous égaux ($\sigma = \sigma^1 = \ldots = \sigma^n$), situation réaliste dans un questionnaire par exemple où chaque juge aurait le même ordre initial des questions/objets. Dans ce cas, les x^i sont indépendants mais uniquement conditionnellement à σ et la vraisemblance associée s'écrit:

$$L(\mu, p) = \frac{1}{m!} \sum_{\sigma \in \mathcal{P}_m} \prod_{i=1}^{n} pr(x^i|\sigma; \mu, p). \tag{4}$$

De façon similaire, les rangs intiaux σ peuvent être égaux à une inversion près, ce qui signifie que le juge peut tout aussi bien faire un algorithme d'insertion à gauche ou à droite. Dans ce cas, on a $\sigma^i \in \{\sigma, \bar{\sigma}\}$ $(i = 1, \ldots, n)$ et, en supposant l'uniformité sur le choix entre σ et son inversion $\bar{\sigma}$, la vraisemblance devient

$$L(\mu, p) \;=\; \frac{1}{m!} \sum_{\sigma \in \mathcal{P}_m} \prod_{i=1}^{n} \frac{1}{2} \left(pr(x^i|\sigma; \mu, p) + pr(x^i|\bar{\sigma}; \mu, p) \right). \tag{5}$$

Les estimations peuvent de nouveau se faire dans les deux cas par un algorithme EM dont l'étape M est spécifiquement adaptée au modèle d'échantillonnage considéré.

Les trois modèles précédents seront respectivement notés ISR, ISR_σ et $\text{ISR}_{\sigma,\bar{\sigma}}$, la notation générique ISR indiquant *Insertion Sorting Rank*. Ils peuvent être comparés en calculant la vraisemblance intégrée suivante :

$$pr(x) = \frac{1}{m!} \sum_{\mu \in \mathcal{P}_m} \int L(\mu, p) pr(p) dp \tag{6}$$

où la loi *a priori* uniforme a été retenue pour μ et la loi *a priori* non informative de Jeffreys, qui est une loi bêta de paramètres $(1/2, 1/2)$, a été retenue pour p. La somme sur les valeurs de μ est réalisable pour des petites valeurs de m et l'intégrale sur p se fait par une simple méthode de Monte-Carlo.

4 Illustration numérique

Notre objectif est ici d'évaluer expérimentalement, sur des jeux de données réelles, la pertinence des trois modèles d'insertion ISR, ISR_σ et $\text{ISR}_{\sigma,\bar{\sigma}}$ ainsi que du modèle Φ de Mallows pour lequel on supposera un échantillonnage avec indépendance des $x^1, \ldots, x^n$.

Les trois 1^ers jeux de données proviennent de questionnaires Q_1, Q_2 and Q_3 donnés à 40 de nos étudiants et décrits comme suit :

- Q_1 : classer mentalement ces nombres par ordre croissant
  $\mathcal{O}_1 = \pi/3$, $\mathcal{O}_2 = \log 1$, $\mathcal{O}_3 = \exp 2$, $\mathcal{O}_4 = \frac{1+\sqrt{5}}{2}$.
- Q_2 : classer ces écrivains français par ordre croissant de date de naissance
  $\mathcal{O}_1$ = V. Hugo, $\mathcal{O}_2$ = Molière, $\mathcal{O}_3$ = A. Camus, $\mathcal{O}_4$ = J.-J. Rousseau.
- Q_3 : classer ces pays par ordre croissant de victoires à la coupe du monde de football
  $\mathcal{O}_1$ = France, $\mathcal{O}_2$ = Allemagne, $\mathcal{O}_3$ = Brésil, $\mathcal{O}_4$ = Italie.

Le second jeu de données correspond aux classements du tournoi des quatre nations de rugby dans la période 1884 à 1909 (19 observations) qui opposa l'Angleterre ($\mathcal{O}_1$), l'Écosse ($\mathcal{O}_2$), l'Irlande ($\mathcal{O}_3$) et le Pays de Galles ($\mathcal{O}_4$). Les quatre jeux de données sont représentés par des polytopes dans l'espace des classements sur la figure 1.

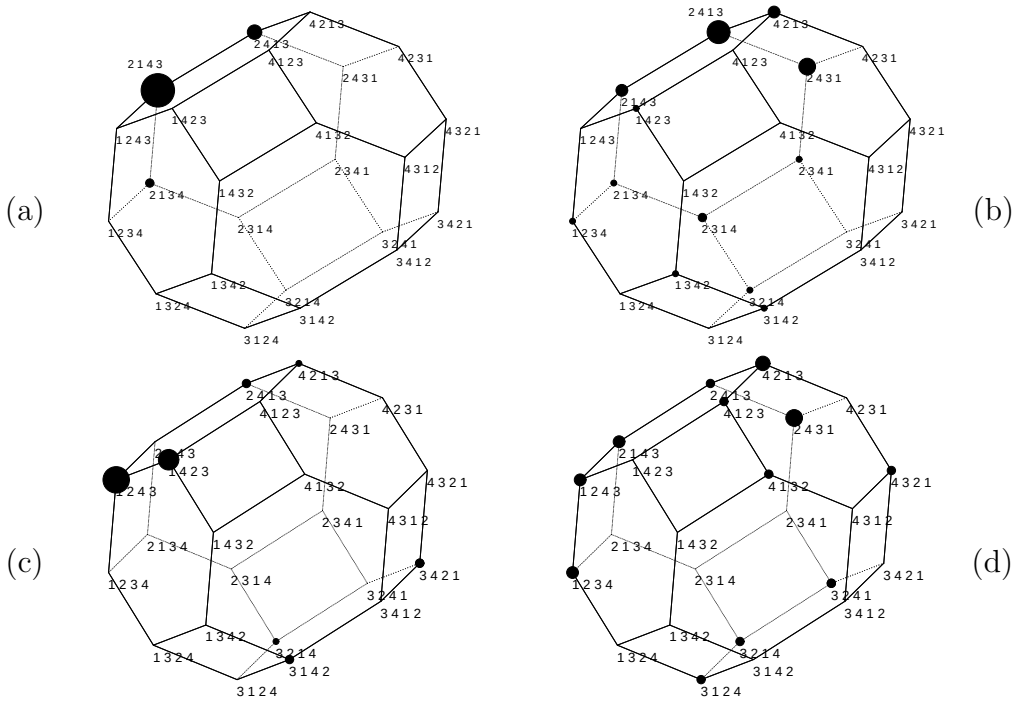

Figure 1: Distributions empiriques pour les questionnaires Q_1 (a), Q_2 (b) et Q_3 (c) et les résultats de la ligue des quatre nations de rugby (d).

La log-vraisemblance intégrée est évaluée pour les quatre modèles ISR, ISR_σ, $\text{ISR}_{\sigma,\bar{\sigma}}$ et le Φ de Mallows sur chacun des jeux de données. Dans le cas du modèle Φ de Mallows, la

vraisemblance intégrée associée est évaluée sous les mêmes conditions que la formule (6) déjà utilisée pour les autres modèles. Les résultats sont donnés dans la table 1.

Modèle	Q_1 (nombres)	Q_2 (écrivains)	Q_3 (football)	Rugby
ISR	76.1801	201.3385	186.2254	**125.2258**
ISR_σ	72.6682	**196.5624**	189.8696	127.3351
$\text{ISR}_{\sigma,\bar{\sigma}}$	75.7878	198.4372	**182.6972**	126.5051
Mallows Φ	**71.1467**	197.0821	188.6102	125.2408

Table 1: Log-vraisemblance intégrée pour les modèles ISR, ISR_σ, $\text{ISR}_{\sigma,\bar{\sigma}}$ et Φ de Mallows.

On remarque tout d'abord qu'aucun des modèles n'est hégémonique, chacun pouvant donc être un candidat d'intérêt en pratique pour modéliser des données de rang en fonction des différentes situations. Notons cependant la bonne tenue des modèles de type ISR, mis en tête dans 75% des cas par la vraisemblance intégrée. Par ailleurs, il est remarquable que les modèles de type ISR_σ et $\text{ISR}_{\sigma,\bar{\sigma}}$ soient systématiquement préférés au modèle ISR pour les questionnaires, situation où il est en effet bien réaliste d'avoir des contraintes présentes sur le vrai ordre de présentation. Au contraire, dans le cas du rugby, qui n'est pas un questionnaire, le modèle libre ISR est préféré, indiquant ainsi l'absence d'ordre initial en commun entre les individus.

Bibliographie

Biernacki, C. and Jacques, J. (2010). A generative model for rank data based on sorting algorithm. *Preprint IRMA Lille*, 70-III.

Fligner, M. A. and Verducci, J. S. (1986). Distance based ranking models. *J. Roy. Statist. Soc. Ser. B*, 48(3):359–369.

Knuth, D. (1973). *Sorting and Searching: Volume 3. The art of Computer Programming.* Addison-Wesley, Massachusetts.

Mallows, C. L. (1957). Non-null ranking models. I. *Biometrika*, 44:114–130.

Marden, J. I. (1995). *Analyzing and modeling rank data*, volume 64 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.