Estimation des paramètres du modèle de Mallows-Bradley-Terry par le maximum de vraisemblance

Simplice Dossou-Gbété, Dominique Lafon, Amadou Sawadogo

MONTE CARLO MAXIMUM LIKELIHOOD METHOD FOR MALLOWS-BRADLEY-TERRY MODELS

[(1)]SIMPLICE DOSSOU-GBÉTÉ, [(2)]DOMINIQUE LAFON & [(3)]AMADOU SAWADOGO

RÉSUMÉ. Le modèle de Mallow-Bradley-Terry est très souvent utilisé pour l'analyse des données de préférence lorsqu'il s'agit d'étudier le classement simultané de q objets selon un critère de préférence. En nous fondant sur la remarque que le modèle de Mallows-Bradley-Terry est une famille exponentielle courbe, nous montrons dans ce travail que les paramètres du modèle de Mallows-Bradley-Terry peuvent être estimés par le maximum de vraisemblance et nous proposons un algorithme pour la mise en oeuvre de cette méthode lorsque le nombre d'objets n'est pas élevé (inférieur à 10 pour les classements sans-ex-aequo). Lorsque le nombre d'objets à classer est supérieur à 10 nous proposons un algorithme qui fait appel à l'évaluation du gradient du log de la vraisemblance par une méthode de Monte Carlo par Chaines de Markov. Des résultats de simulations seront présentés pour l'évaluation des performances des algorithmes proposés

ABSTRACT. This paper is devoted to the maximum likelihood estimation of the parameters of the Bradley-Terry parameters when dealing with the Mallows-Bradley-Terry model for complete ranking data analysis. The maximum likelihood method is avoid until now because of the normalising constant that involves an untractable sum when the number of terms to be ranked is large, i.e. 10 or over. After we recalled that the Mallows-Bradley-Terry model is a curved exponential family, we pointed out that the Hessian matrix of the likelihood function is lower bounded. We show how to implement a MM-algorithm that incorporates a Markov Chain Monte Carlo scheme in the gradient step as viable option for the calculation of the maximum likelihood estimates. This algorithm uses the lower bound of the Hessian matrix in order to avoid repetitive calculations of the Hessian matrix required by Newton-Raphson type of algorithm. The method is also valid for models where covariates are associated with the Bradley-Terry parameters through a link function.

1. INTRODUCTION

Let's consider a set of q objects or items labelled by distinct numbers from 1 to q and submitted to a judge for ranking. Let's denote r the resulting ranks vector, that is a finite sequence of integers of length q where each value appears at least once and belongs to the set of consecutive integers from 1 to q. Only ranking without ties will be considered in the work, thus each integer value of the sequence r appears exactly once. With this constraint, a ranking model is defined to be a family of probability distributions defined on rankings considered as a set of permutations of the finite sequence $1:q$, the symmetric group $\mathscr{S}(q)$. Distributions belonging to such a family are indexed by a finite set of parameters, where certain parameters may be interpreted as item (object) parameters which reflect solely properties of the objects being ranked.

Ranking of q items are related to paired comparisons of that items. On one hand it is clear a ranking can be broken down into a set of $\frac{q(q-1)}{2}$ paired comparisons. On the other hand, a complete paired comparisons design may lead to a ranking provided the multiple pairwise judgements are consistent. Consistency means that there are no circular triads of comparisons such as x foregoes y, y is prior to z and z foregoes x. The idea underlying the statistical models for ranking data based on paired comparisons as a possible way of generating rankings data can

Date: January 2009.

be formulated as follows: a judge constructs a ranking by starting with paired comparisons, but reports his/her ranking only after having a consistent set of comparisons that yields an unambiguous ranking [13, 2]. This paradigm dates back at least to Babington Smith who introduced a general statistical model based on paired comparisons probabilities. Statistical models based on paired comparisons constitute an important class of probability distributions models available for the statistical analysis of ranking data along with order statistics models, multistage ranking models, distance-based ranking models.

2. Statistical models for complete rank data using paired comparison approach

2.1. **Babington Smith general model for complete rankings.** The idea behind the Babington Smith probability model for ranking is that all of the ranking structure can be recovered from the pairwise comparisons. Thus It is assumed there is a set of $\frac{q(q-1)}{2}$ parameters $\theta_{ij} \in]0,1[$, $1 \leq i < j \leq q$ such that the probability the ranking r occurs by chance is $p(r,\theta) = c(\theta) \prod_{(i,j)} \theta_{ij}^{I[r(i)<r(j)]}$ where $\theta_{ji} = 1 - \theta_{ij}$ for $i < j$ and $I[A]$ denotes the indicator function of the event A , that is $\mathrm{I}[A] = 1$ when A occurs and $I[A] = 0$ otherwise. $c(\theta)$ stands for the normalising constant. θ_{ij} is interpreted as the probability that item i is ranked lower than item j if the random experiment reduces to the comparison of this pair of items only. One should notice that θ_{ij} is different from the probability that item i is ranked lower then item j if the ranking experiment concern $q > 2$ items including these specific items. The Babington Smith model involves $\frac{q(q-1)}{2}$ parameters and this may be too large to deal with and the data analysis may be not easy and reliable. Moreover, burden of computation may occurs, specifically when dealing with Newton-Raphson like algorithm for the likelihood maximisation. Attempt to reduce the number of the model's parameters and to obtain a more tractable and interpretable models yielded the Mallows-Bradley-Terry model. It is this latter that will be under consideration in this work.

2.2. **Mallows-Bradley-Terry submodel for complete rank data.** A sub-model of the Babington Smith model for ranks without tie, that reduces its forementioned drawback, has been proposed by Mallows by assuming a Bradley-Terry model for the preference probabilities θ_{ij} as $\theta_{ij} = \frac{\pi_i}{\pi_i + \pi_j}$ with $\pi_i > 0$. Since this model is scale invariant, identifiability of Babington Smith rankings probability model is achieved by putting constraints on the parameters space. To this end it is common to consider the sum constraint $\sum_{i=1}^{q} \pi_i = 1$ but other constraints may be used. Thus $\frac{\theta_{ij}}{1-\theta_{ij}} = \frac{\pi_i}{\pi_j}$ and the Babington Smith ranking distribution model has the following form with this new parametrisation $p(r,\pi) \propto \exp\left\{ \sum_{1 \leq i < j \leq q} \left(\log \pi_i - \log \pi_j\right) I[r(i) < r(j)] \right\}$. It is well known that the probability $p(r,\pi)$ is proportional to the product $\prod_{j=1}^{q} \pi_j^{q-r(j)}$ and this allows to rewrite the Mallows-Bradley-Terry ranking probability as $p(r,\pi) = C(\pi) \prod_{j=1}^{q} \pi_j^{q-r(j)}$. This result implies $q - r = (q - r(j))_{j=1:q}$ is a sufficient statistic for the Mallows-Bradley-Terry model. If the constraint $\sum_{j=1}^{q} \pi_i = 1$ is considered for the model identifiability, one can

reparametrized the model as follows: $\theta_j = \log\left\{\frac{\pi_j}{\pi_q}\right\}$ for $j = 1, \cdots, q$ leading to the following relationship between the π_js and the θ_js: $\pi_j = \frac{\exp(\theta_j)}{\sum_{l=1}^{q}\exp(\theta_l)}$, $j = 1:q$ with $\theta_q = 0$. Whence one can state that the Mallows-Bradley-Terry model is a curved exponential family model with parameters θ_l, $l = 1, \cdots, q-1$.

3. Maximum likelihood estimation of Mallows-Bradley-Terry model by using a Monte Carlo MM-algorithm

In what follows we shall consider the Mallows-Bradley-Terry model as $p(r,\pi) = c(\pi)\prod_{j=1}^{q}\pi_j^{q-r(j)}$ where $\log(c(\pi)) = -\log\left\{\sum_r \exp\left[\sum_{j=1}^{q}(q-r(j))\log\pi_j\right]\right\}$. It results that the log-likelihood of the canonical parameters $\theta_j = \log\left(\frac{\pi_j}{\pi_q}\right)$, $j = 1:q$ given a random sample of n rankings r_i, $i = 1:n$ is $l(\theta) = \sum_{j=1}^{q}\log\{\pi_j(\theta)\}\sum_{i=1}^{n}(q-r_i(j)) - n\psi(\theta)$ with $\theta = \left(\theta_j\right)_{j=1:q}$, $\theta_q = 0$ and $\psi(\theta) = -\log(c(\pi(\theta)))$. The main difficulty in dealing with the maximum likelihood estimation (mle) method is the computation of the normalising constant $c(\pi(\theta)) = \exp\{-\psi(\theta)\}$. When the number q of ranked items is equal or greater than 10, the computation of $c(\pi(\theta))$ involves an explicit sum over more than 3×10^6 terms. To get round this problem, Critchlow D.J. and Fligner M. [2] has proposed to handle the Mallows-Bradley-Terry model as a GLM with a log link function and a multinomial family; $\psi(\theta)$ is considered as a nuisance parameter to be estimated as the intercept of the linear predictor. This approach is until now widely used in the literature on preference data analysis, e.g. [1, 3]. An alternative approach to overcome the computation of the normalising constant $c(\pi(\theta))$ has been mentionned by Marden [13] as an EM-algorithm, but he didn't give indications on how to implement this solution. This idea is pursued in what follows.

3.1. Introducing to MM-algorithms and ingredients for the likelihood maximisation by using a MM-algorithm.

The MM-algorithms consist in a class of algorithms that includes EM-algorithm. These algorithms aim to tranform untractable optimisation problems into tractable ones [5]. A MM-algorithm scheme is made of the following two steps, given an estimate θ^{old} of the parameters vector θ:

(1) Substitute a surrogate function $S\left(\theta,\theta^{old}\right)$ for the objective function $l(\theta)$ such that $l(\theta) \geq S\left(\theta,\theta^{old}\right)$ and $l\left(\theta^{old}\right) = S\left(\theta^{old},\theta^{old}\right)$;
(2) Updating the estimate of θ by maximising the surrogate function $S\left(\theta,\theta^{old}\right)$ with respect to θ.

It is readily seen if $\theta^{new} = \arg\max\left\{S\left(\theta,\theta^{old}\right)\right\}$ then

$$l\left(\theta^{old}\right) = S\left(\theta^{old},\theta^{old}\right) \leq S\left(\theta^{new},\theta^{old}\right) \leq l(\theta^{new}) = S(\theta^{new},\theta^{new}).$$

and therefore the MM-algorithms are monotone. Moreover if $\nabla S\left(\theta^{old},\theta^{old}\right) = 0 \Rightarrow \nabla l\left(\theta^{old}\right) = 0$ any cluster point of a sequence defined by the recurrence $\theta^{(t+1)} = \arg\max S\left(\theta,\theta^{(t)}\right)$ is a local maximum of $l(\theta)$.

3.2. **An MM-algorithm for the maximum likelihood method.** Given a ranking data r_i, $i=1:n$ the log-likelihood of the parameters $\theta=\left(\theta_j\right)_{j=1,\cdots,q-1}$ is $l(\theta)=\sum_{i=1}^{n} l(\theta,r_i)$ where $l(\theta,r_i)=\sum_{j=1}^{q}\left[q-r_i(j)\log\left\{\pi_j(\theta)\right\}\right]+\log\{C(\pi(\theta))\}$. Let's denote $\mathbb{E}(\theta)=\left[\mathbb{E}(\theta,j)\right]_{j=1:q}$ the mathematical expectation of the Mallow-Bradley-Terry distribution with parameter θ. The score vector of the log-likelihood function is $\nabla_\theta l(\theta)=\left\{\sum_{i=1}^{n}\left\{\mathbb{E}(\theta,l)-r_i(l)\right\}\right\}_{l=1,\cdots,q-1}$. The Hessian matrix is $\nabla^2_\theta l(\theta)=-n^t U\left[\operatorname{diag}\{p(\theta)\}-p(\theta)^t p(\theta)\right]U$, where $p(\theta)=\{p(s,\pi(\theta))\}_s$ is the vector of the probability masses of the rankings $s\in\mathscr{S}(q)$, $u_s=(s(l))_{l=1,\cdots,q-1}$ and $U=\{s(l)\}_s^{l=1:(q-1)}$ is a matrix with $q!$ rows and $q-1$ columns. The matrix $-\nabla^2_\theta l(\theta)$ is symmetric and nonnegative and then the log-likelihood is concave. Furthermore one can show that $-\nabla^2_\theta l(\theta)$ is uppper bounded by $nq(q-1)^2 I$ since the ranks $s(l)$ obey the constraint$1\leq s(l)\leq q$ for any ranking s. It results that the likelihood $l(\theta)$ is lower bounded as $S(\theta,\theta')\leq l(\theta)$ where

$$S\left(\theta,\theta'\right)=-nq(q-1)^2\,{}^t\left(\theta-\theta'\right)\left(\theta-\theta'\right)+{}^t\left(\theta-\theta'\right){}^t\nabla_\theta l\left(\theta'\right)+l\left(\theta'\right).$$

One observes $\nabla S(\theta',\theta')=\nabla_\theta l(\theta')$ and thus $\nabla S(\theta',\theta')=0\Rightarrow\nabla l(\theta')=0$. Using $S(\theta,\theta')$ as surrogate function for the optimisation of $l(\theta)$ leads to the following update rule:

$$\theta^{new}=\theta^{old}+\frac{1}{2q(q-1)^2}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{E}\left(\theta^{old}\right)-r_i\right]\right\}.$$

The update rule requires the computation of the mathematical expectations $\mathbb{E}\left(\theta^{old},j\right)$, $j=1,\cdots,q-1$ which involve sums with a huge number of terms when the number q of items if large (e.g. 10 or over). This problematical computation may be overcome by using Monte Carlo maximum likelihood approach.

4. MM-ALGORITHM VIA MONTE CARLO MARKOV CHAIN PROCEDURE

Monte Carlo maximum likelihood methods are known as procedures for computing maximum likelihood estimation that avoids untractable computations for a large class of probability models that includes exponential families. The main feature of these methods for exponential families of distributions as Mallows-Bradley-Terry models is that the maximum likelihood estimation can be carried out without the analytic computation of the normalising constant $c(\pi(\theta))$ and its derivatives by using Markov Chain Monte Carlo approximations [6]. One approach in dealing with this class of methods, named Simulated maximum likelihood, consists in a direct approximation of $\psi(\theta)$ by using importance sampling [6, 8]. An alternative approach is derived from Monte Carlo Newton-Raphson procedure [11, 12, **?**] and consists in incorporating a Markov Chain Monte Carlo scheme in the gradient step of a likelihood maximisation algorithm like a MM-algorithm. In this section, we attempt to implement a Monte Carlo Newton-Raphson type procedure by incorporating a Monte Carlo procedure in the MM-algorithm scheme to obtain reliable mle estimations of the Bradley-Terry parameters.

4.1. **Generalized Gibbs sampling.** Let $\tau(a,b)$ denote the transposition of the integers a and b that belong to the set of integers $1,\cdots,q$. If r is a complete ranking, $s=\tau(a,b)r$ means $s(j)=r(j)$ if $r(j)\notin\{a,b\}$, $s(j)=b$ if $r(j)=a$ and $s(j)=a$ if $r(j)=b$. One has $\frac{\Pr(\tau(a,b)r)}{\Pr(r)}=$

$\left\{\frac{\pi_{[t==a]}}{\pi_{[r==b]}}\right\}^{a-b}$. It is well known that for two rankings r and s there is a finite sequence (a_i,b_i), $i = 1:n$ such that $s = \prod_{i=1}^{n} \tau(a_i,b_i)\, r$. Thus a generalized Gibbs scheme can be stated as follows:

Algorithm 1 (Generalized Gibbs sampling)

Repeat until convergence

(1) sample two distincts integers a and b according to the uniform distribution on the set of consecutive integers from 1 to q.

(2) $r = \tau(a,b)\, r$ if $U(0,1) < \left\{\frac{\pi_{[t==a]}}{\pi_{[r==b]}}\right\}^{a-b}$

4.2. **MM-algorithm using MCMC scheme.** The algorithm above provides us with a scheme to generate a path of a Markov chain with stationary distribution $p(\theta)$. Thus we can compute a Monte Carlo approximation of the score vector of the parameter likelihood. We can state a MCMC MM-algorithm as follows:

Algorithm 2 (MCMC MM-algortithm)

Repeat until convergence

(1) $\theta_{old} = \theta_{new}$

(2) Generate a MCMC sample r_k, $k = 1:N$ with stationary distribution $p(\theta_{old},r)$

(3) Compute a Monte Carlo approximation of the mathematical expectation $\tilde{\mathbb{E}}_{N,\theta_{old}} = \frac{1}{N}\sum_{k=1}^{N} r_k$

(4) Update the parameter estimation $\theta_{new} = \theta_{old} + \frac{2}{q(q-1)^2}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\tilde{\mathbb{E}}_{N,\theta_{old}}(r) - r_i\right]\right\}$

When dealing with the analytic MM-algorithms, the update rules are repeated until some numerical convergence criterion is fullfilled. Usually this stopping criterion is stated as the convergence of the sequence of the updapted parameters vectors or the convergence of the sequence of gradient vectors $\nabla S\left(\theta^{old},\theta^{old}\right)$ to 0 with respect to some numerical tolerance. Such a stopping criterion is no more admissible in the case of the Monte Carlo MM-algorithm because of noise induced by Monte Carlo sampling. Thus one should seek a less tight stopping criterion that prevents from the error (type 1 error) of stopping to late, that is after $\nabla_\theta l(\theta^{new}) = 0$ as suggested in [11].

4.3. **A stopping criterion for the Monte Carlo MM-algorithm and assessment of the accuracy of the parameters estimates.** As mentioned earlier, one has $\nabla_\theta S(\theta^{new},\theta^{new}) = \nabla_\theta l(\theta^{new})$ after an update of the parameters vector and $\nabla_\theta l(\theta^{new}) = \left\{\frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{E}_{\theta^{new}}\{s(j)\} - r_i(j)\right]\right\}_{j=1:q}$. Hence, by following the idea in [11, 12], a stopping criterion might be proposed, based on an overall statistic $W(\theta^{new}) = {}^t\nabla_\theta \tilde{l}_N(\theta^{new})\,\tilde{\Sigma}^{-1}\nabla_\theta \tilde{l}_N(\theta^{new})$ where $\tilde{\Sigma}$ is some consistent estimate of the covariance matrix of $\left(\tilde{\mathbb{E}}_{N,\theta^{new}}\{s(j)\}\right)_{j=1:q}$. Given a sample of Markov Chain Monte Carlo replicates of rankings r_k, $k = 1:N$ with stationary probability $p(r,\pi(\theta))$, it is well known by the Ergodic Theorem that $\tilde{\mathbb{E}}_{N,\theta}\{r(j)\} = \frac{1}{N}\sum_{k=1}^{N} r_k(j)$ converges to $\mathbb{E}_\theta\{r(j)\}$ with probability 1 as N increases to infinity (cf. , [14]). Furthermore, since the states space of

the Markov chain r_k^{mc} is finite, the Central Limit Theorem holds for $\left(\tilde{\mathbb{E}}_{N,\theta}\{r(j)\}\right)_{j=1,\cdots,q-1}$. Thus $\sqrt{N}\nabla_\theta \tilde{l}_N(\theta) = \sqrt{N}\left(\tilde{\mathbb{E}}_{mN,\theta}\{s(j)\} - \frac{1}{n}\sum_{i=1}^{n} r_i(j)\right)_{j=1,\cdots,q-1}$ is asymptotically gaussian.

A consistent estimate of the covariance matrix of the limiting gaussian distribution of $\sqrt{N}\nabla_\theta \tilde{l}_N(\theta)$ is yield by using the batch means method as it is described in [7, 10, 4]. Such an estimate of the covariance matrix of the random vector $\left(\tilde{\mathbb{E}}_{N,\theta^{new}}\{s(j)\}\right)_{j=1:q}$ can be yielded by using the batch means method. This method aims to estimate the steady-state mean of a weakly stationary sequence of random variables and the variance of the sample mean (for fixed size). To achieve this goal a sequence of sub-samples (batch-means) is built by aggregating b successive observations in order to overcome the problem of auto-correlation between consecutive sample elements. It is frequently used and owes its popularity to its simplicity and effectiveness.

Since $\sqrt{N}\nabla_\theta \tilde{l}_N(\theta)$ is asymptotically gaussian, it results that the distribution of $W(\theta^{new})$ is a chi-square with degrees of freedom equal to the dimension of parameters space, under the null hypothesis that the analytic gradient $\nabla_\theta l(\theta^{new})$ is 0, as it is claimed in [12]. Hence a stopping criterion is proposed as $W(\theta^{new}) < \chi_{q-1}(1-\alpha)$ where $\chi^2_{q-1}(1-\alpha)$ is the lower $(1-\alpha)$-quantile of a chi-square distribution with $q-1$ degrees of freedom.

REFERENCES

[1] Courcoux Ph. & Séménou M., (1997): Preference data analysis using a paired comparison model. Food Quality and Preference, 8(5/6), pp. 353 – 358.

[2] Critchlow D.E. & Fligner M.A. (1991): Paired comparison, triple comparison, and ranking experiments as generalized linear models, and their implementation on glim. Psychometrika, 51(3), pp.517 – 533.

[3] Dittrich R., Katzenbeisser W. and Reisinger H., (2000): The analysis of rank ordered preference data based on Bradley-Terry Type Models. OR Spektrum, 22(1), pp.117 – 134.

[4] Flegal J.M., Haran M. & Jones G.L. (2008): Markov Chain Monte Carlo: Can We Trust the Third Significant Figure? arXiv:math/0703746v4

[5] Hunter D. R. & Lange K. (2004): A Tutorial on MM Algorithms. The American Statistician, 58(1), pp.30-37.

[6] Geyer C.J. (1991): Markov Chain Monte Carlo Maximum Likelihood. Computing Science and Statistics: Proceedings of 23rd Symp. Interface, pp.156-163.

[7] Geyer C.J. (1992): Practical Markov Chain Monte Carlo. Statistical Science, 7(4), pp.473 – 483

[8] Geyer C.J. (1994): On the Convergence of Monte Carlo Maximum Likelihood Calculations. Journal of the Royal Statistical Society, B-56(1), pp.261-274.

[9] Geyer C.J. & Thompson E. (1992): Constrained Monte Carlo maximum likelihood for dependent data. Journal of the Royal Statistical Society, B-54, pp.657-699.

[10] Jones G.L., Haran M., Caffo B. S. & Neath R. (2006): Fixed-width output analysis for Markov Chain Monte Carlo. http://arxiv.org/abs/math/0601446. Journal of the American Statistical Association, 101:1537–1547 .

[11] Kuk A.Y.C. & Cheng Y.W. (1997): The Monte Carlo Newton-Raphson algorithm. Journal of Statistical Computing and Simulation, 59(2), pp.233-250

[12] Kuk A.Y.C. & Cheng Y.W. (1999): Pointwise and functional approximation in Monte Carlo maximum likelihood estimation, Statistics and Computing 9, pp.91-99.

[13] Marden J.I. (1995): Analysing and modelling rank data. Chapman & Hall ed., London.

[14] Robert C.P. & Casella G. (2004): Monte Carlo statistical methods. 2nd edition, Springer.

(1) LABORATOIRE DE MATHÉMATIQUES ET DE LEURS APPLICATIONS DE PAU (LMAP) UMR CNRS 5142 UNIVERSITÉ DE PAU ET DES PAYS DE L'ADOUR AVENUE DE L'UNIVERSITÉ B.P. 576 64012 PAU CEDEX (FRANCE)

(2) CENTRE DES MATÉRIAUX DE GRANDE DIFFUSION ECOLE DE MINES D'ALÈS, HELIOPARC AVENUE DU PROFESSEUR ANGOT 64000 PAU CEDEX (FRANCE)