Journal of the Korean Data & Information Science Society
2013, **24**(5), 1089–1099

http://dx.doi.org/10.7465/jkdi.2013.24.5.1089
한국데이터정보과학회지

Revisiting the Bradley-Terry model and its application to information retrieval

Jong-June Jeon[1] · Yongdai Kim[2]

[12]Department of Statistics, Seoul National University

Received 1 July 2013, revised 28 August 2013, accepted 16 September 2013

Abstract

The Bradley-Terry model is widely used for analysis of pairwise preference data. We explain that the popularity of Bradley-Terry model is gained due to not only easy computation but also some nice asymptotic properties when the model is misspecified. For information retrieval required to analyze big ranking data, we propose to use a pseudo likelihood based on the Bradley-Terry model even when the true model is different from the Bradley-Terry model. We justify using the Bradley-Terry model by proving that the estimated ranking based on the proposed pseudo likelihood is consistent when the true model belongs to the class of Thurstone models, which is much bigger than the Bradley-Terry model.

Keywords: Bradley-Terry model, consistency, model misspecification, probability model for ranking, Thurstone model.

1. Introduction

Information retrieval (IR) is a process of activities to seek information relevant to a topic of interest. In a narrow sense, IR is a task to produce a relevant list of items (documents) for a given query. For example, a Internet user searches documents with the query of 'big data' in web portal. A search engine in the web portal provides a list of documents expected to be relevant to the query.

As the keyword search service becomes a part of major industries, the research of IR is one of the most important fields of the big data analysis. To satisfy Internet users, the list of numerous documents should be ordered according to the relevances to the query periodically, and the relevances should be measured continuously. If a relevant document is located at the end of the provided list, the user may not find the document and he (she) may become unsatisfied with the keyword search service. The efficient measurement of relevances is a core problem in the IR.

The relevance is usually measured by the probability of the rankings of documents. Even though the relevance is conceptually numeric, it is enough to know the ranking of relevances of the documents for IR. The lower the ranks are, the more relevant the documents are to

[1] Corresponding author: Postdoctoral researcher, Department of Statistics, Seoul National University, Seoul 151-742, Korea. E-mail: jj.jeon@gmail.com
[2] Professor, Department of Statistics, Seoul National University, Seoul 151-742, Korea.

a given query. In statistics a lot of probability models for ranking are developed such as the Thurstone model (Thurstone, 1927), the Bradley-Terry model (Bradley and Terry, 1952), the Luce model (Luce, 1959), Mallows' $\phi-$model (Mallows, 1957), gamma ranking model (Stern, 1990), distance based ranking model (Flinger and Verducci, 1986). A ranking model for IR based on the Bradley-Terry model is proposed by many researchers (Joachims, 2003; Freund *et al.*, 2003; Burges *et al.*, 2005; Xia *et al.*, 2008). The popularity of the Bradley-Terry model is due to its resemblance to the logistic regression model and computational advantages (Yan *et al.*, 2010; Hong *et al.*, 2010).

However, it is doubtful to use the Bradley-Terry model for IR in a sense of estimating the ranking of relevances, since some assumptions required for the Bradley-Terry model are not met for ranking data of documents. First, only two documents are compared in each trial in the Bradley-Terry model while several documents (i.e. documents appeared on the computer screen at the same time) are compared simultaneously. Second, the Bradley-Terry model assumes a certain structure of the probability model for ranking and it is not clear how restrictive the model structure of the Bradley-Terry model is. If the probability model corresponding to the Bradley-Terry model is too small, the bias from model misspecification may be serious.

The purpose of this paper is to add positive justification of using the Bradley-Terry model for IR. We propose a pseudolikelihood for ranking data of documents based on the Bradley-Terry model, and prove that the relevances of documents for a given query can be estimated consistently when the true model belongs to the class of Thurstone models, which is the most natural probability model for ranking and is much bigger than the Bradley-Terry model. Roughly speaking, we can ignore the model misspecification problem of the Thurstone model in estimating the rank of relevances by using the Bradley-Terry model.

The paper is organized as follows. Section 2 introduces the Thurstone model without covariates and the pseudolikelihood method based on the Bradley-Terry model. Section 3 considers the Thurstone model with covariates and studies asymptotic properties of the estimator based on the proposed pseudo likelihood. Some algorithms of the IR system are investigated comparing the pseudolikelihood in the Thurstone model, and concluding remarks follow in Section 4.

2. Thurstone model for ranking without covariates

The IR system manages a massive data set of documents, and gives a relevant document list for a given query. If the query of 'big data' is submitted into an IR system (e.g. Google or Naver), it returns an ordered set of documents according to the computed relevances. In the IR system the relevances is estimated using the click-through data (Joachims, 2003), which is an implicit feedback to the list of documents. Figure 2.1 shows the process of cumulating the click-through data in the IR system. The user selects or clicks some documents in the given list. Then, the IR system records the document list and the sequence of clicks. With recorded click-through data, the IR system updates the relevances of the documents to the queries. The order of clicks obtained in the click-through data is used for estimating relevances to the documents. For convenience we assume that the order of relevances is the order of clicks. Then, we can assign the rank of the documents. In the Figure 2.1, the rank 3,1,2,4 are assigned to the documents A,B,C,D respectively.

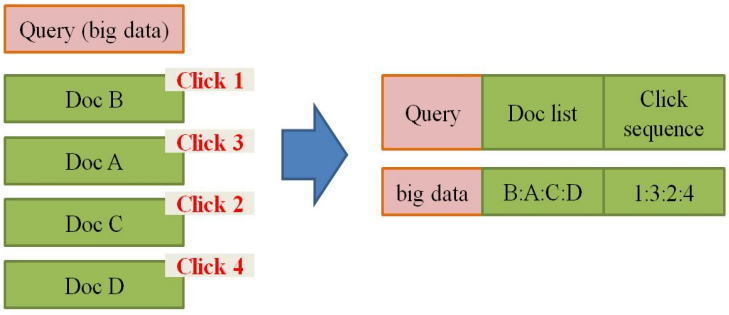

Figure 2.1 Process of cumulating the click-through data in the IR system

If relevant documents do not locate at the top of the list, users have hard time to find them and hence are unsatisfactory to the IR system. That is, to satisfy users, the IR system should give a list of documents in which the relevant documents are highly ranked. In addition, algorithms in IR systems should be sufficiently fast since the number of documents is huge and users are usually impatient. The two objectives of IR systems - accurate estimation of relevancies and fast computations, seem to be conflict to each other. A refined probability model is necessary for accurate estimation while a simple model is computationally beneficial. In this section, we consider the Thurstone model (Thurstone, 1927) as a probability model for rankings which is general and flexible, and show that the Bradley-Terry model, which has computational advantages, can be used to estimate the parameters in the Thurstone model.

2.1. Probability model

Suppose that there are p documents to be ranked with no ties. Let $S = \{1, \cdots, p\}$ be the index set of the documents and $\mathcal{S}$ be a class of all permutations of S. Denote the ranking vector of the document set by $R = (R_1, \cdots, R_p) \in \mathcal{S}$ where R_j is the ranking of the document j. Let $Z_j = \lambda_j + \epsilon_j$ be the latent variable associated with the document j, where $\lambda_j \in \mathbb{R}$ is a location parameter to represent the average relevance of document j, and ϵ_j for $j = 1, \cdots p$ is a sequence of independent random variables having a continuous distribution function F. The random variable Z_j can be regarded as unobserved relevance of a document j. Naturally the ranking is derived from the order of the latent variables Z_j for $j = 1, \cdots, p$.

Definition 2.1 (Thurstone model) Let $Z_j = \lambda_j + \epsilon_j$, where $\lambda_j \in \mathbb{R}$ for $j = 1, \cdots, p$ and ϵ_j for $j = 1, \cdots, p$ are independent and identically random variables with F. If $R = (R_1, \cdots, R_p) \in \mathcal{S}$ be ranking vector is defined by $R_j = \sum_{k=1}^{p} I(Z_j \leq Z_k)$ for $j = 1, \cdots, p$, we call this ranking model the Thurstone model with F.

Let (k) be the index of the document whose rank is k, then the probability of ranking R is given by

$$P(R) = P(\epsilon_1 + \lambda_{(1)} > \cdots > \epsilon_p + \lambda_{(p)}), \tag{2.1}$$

where $\epsilon_j \sim_{iid} F$ for $j = 1, \cdots, p$. For a special case, if F is the standard Gumbel distribution,

(2.1) becomes

$$P(R) = \sum_{k=1}^{p} \frac{\exp(\lambda_{(k)})}{\sum_{k \le j \le p} \exp(\lambda_{(j)})}, \tag{2.2}$$

which is called the Plackett-Luce model (Luce, 1959; Plackett, 1975).

Let r_i for $i = 1, \cdots, n$ be independent observations of ranking vectors, which are obtained from the ith searching in the IR system. The likelihood is given by

$$\mathcal{L}(\boldsymbol{\lambda}) = \prod_{i=1}^{n} P(R = r_i; \boldsymbol{\lambda}), \tag{2.3}$$

where $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_p)$ is a vector of parameters. For identifiability of the model, we let $\lambda_p = 0$. Here, the parameters $\boldsymbol{\lambda}$ can be used as the measurement of relevancy of documents to a given query.

The parameter can be estimated by maximizing (2.3). In spite of good asymptotic properties, the maximum likelihood estimator is generally hard to compute, since the likelihood does not have a closed form except very few cases. To avoid the problem, we suggest an estimation method based on the Bradley-Terry model.

2.2. Pairwise comparison pseudo likelihood

In the probability model for ranking, the marginal probability for two documents is given by $P(R_j < R_k) = \sum_{R:R_j<R_k} P(R)$. In the Thurstone model with F the marginal probability is $P(R_j < R_k) = G(\lambda_j - \lambda_k)$ where G is the distribution function of $\epsilon_1 - \epsilon_2$ where ϵ_1 and ϵ_2 are independent random variables following F. In the Plackett-Luce model, G is given by the standard logistic distribution function such that its marginal probability is $G(\lambda_j - \lambda_k) = \exp(\lambda_j)/(\exp(\lambda_j) + \exp(\lambda_k))$, which is known as the preference probability in Bradley-Terry model (Bradley and Terry, 1952). To estimate the parameter $\boldsymbol{\lambda}$ in the Thurstone model, we propose to use the following pairwise comparison likelihood, which is a kind of pseudo likelihood, based on the marginal probability.

Definition 2.2 (Pairwise comparison likelihood) Assume that the probability model for ranking is the Thurstone model with F. Let $r_i = (r_{i1}, \cdots, r_{ip})$ for $i = 1, \cdots, n$ be independent observations of ranking vector, and let G be a distribution function corresponding to the marginal probability in the Thurstone model. Then, the pairwise comparison likelihood is defined as

$$\mathcal{L}^{pw}(\boldsymbol{\lambda}) = \prod_{i=1}^{n} \prod_{j<k} G(\lambda_j - \lambda_k)^{I(r_{ij}<r_{ik})} G(\lambda_k - \lambda_j)^{I(r_{ij}>r_{ik})}. \tag{2.4}$$

Denote the logarithm of the pairwise comparison likelihood, the pairwise comparison log-likelihood, by $l^{pw}(\boldsymbol{\lambda})$. The maximum pairwise comparison likelihood estimator (MPCLE) is given by

$$\widehat{\boldsymbol{\lambda}} = \text{argmax}_{\lambda} l^{pw}(\boldsymbol{\lambda})$$

with $\hat{\lambda}_p = 0$. The following theorem proves that the MPCLE is consistent.

Theorem 2.1 Let $\boldsymbol{\lambda}^* = (\lambda_1^*, \cdots, \lambda_p^*) \in C$ be the true parameter of the Thurstone model with F, where C is a compact subset of $\mathbb{R}^{p-1} \times \{0\}$ and $\lambda_j^* \neq \lambda_k^*$ for all $j \neq k$. If G is differentiable and strictly logconcave, then the estimator $\hat{\boldsymbol{\lambda}}$ is consistent as n goes to infinity.

Proof: By law of large number, $\sum_{i=1}^n I(r_{ij} < r_{ik})/n \to_p G(\lambda_j^* - \lambda_k^*)$ as $n \to \infty$ for all j, k. It follows that

$$l^{pw}(\boldsymbol{\lambda})/n \to_p El^{pw}(\boldsymbol{\lambda}) = \sum_{j \neq k} G(\lambda_j^* - \lambda_k^*) \log G(\lambda_j - \lambda_k)$$

for each $\boldsymbol{\lambda} \in C$ as $n \to \infty$. It is easily shown that the first derivative of $El^{pw}(\boldsymbol{\lambda})$ is zero at $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$. By convexity lemma (Pollard, 1991), $l^{pw}(\boldsymbol{\lambda})/n$ uniformly converges to $El^{pw}(\boldsymbol{\lambda})$ on the compact set C in probability. It suffices to prove that $El^{pw}(\boldsymbol{\lambda})$ is strictly convex function. Without loss of generality, we can assume C is a convex set in $\mathbb{R}^p$. Set $\boldsymbol{\lambda}^1 \neq \boldsymbol{\lambda}^2$ where $\boldsymbol{\lambda}^1 = (\lambda_1^1, \cdots, \lambda_{p-1}^1, 0)$, $\boldsymbol{\lambda}^2 = (\lambda_1^2, \cdots, \lambda_{p-1}^2, 0) \in C$. Since $\log G$ is strictly concave, $\log G(tv_1 + (1-t)v_2) > t \log G(v_1) + (1-t) \log G(v_2)$ for all $v_1 \neq v_2$ and $t \in (0, 1)$. Let $v_1 = \lambda_j^1 - \lambda_k^1$ and $v_2 = \lambda_j^2 - \lambda_k^2$, then there exists $j, k \in \{1, \cdots, p\}$ such that

$$\log G(t(\lambda_j^1 - \lambda_k^1) + (1-t)(\lambda_j^2 - \lambda_k^2)) > t \log G(\lambda_j^1 - \lambda_k^1) + (1-t) \log G(\lambda_j^2 - \lambda_k^2)$$

for all $t \in (0, 1)$. Since $G(\lambda_j^* - \lambda_k^*) > 0$ for all j, k, $El^{pw}(t\boldsymbol{\lambda}^1 + (1-t)\boldsymbol{\lambda}^2) > tEl^{pw}(\boldsymbol{\lambda}^1) + (1-t)El^{pw}(\boldsymbol{\lambda}^2)$. Therefore, $El^{pw}(\boldsymbol{\lambda})$ is strictly concave, and it concludes that

$$\|\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^*\|_2 \to_p 0$$

as $n \to \infty$. □

Theorem 2.1 implies that we can obtain the consistent estimator of the model with easier computation, provided that we choose the correct distribution of G derived from the Thurstone model.

Remark 2.1 Another famous model for ranking is the Bradley-Terry-Mallows model (Critchlow *et al.*, 1991) , which assumes that the probability of ranking $R = (R_1, \cdots, R_p)$ is given by $P(R) = C(\mathbf{u}) u_j^{p-R_j}$ where u_j for $j = 1, \cdots, p$ is a nonnegative parameter corresponding to the item j and $C(\mathbf{u}) = 1/\sum_{R \in \mathcal{S}} \sum_{j=1}^p u_j^{p-R_j}$ for $\mathbf{u} = (u_1, \cdots, u_p)$. The log odds ratio of the marginal probability for two items depends on $\mathbf{u}$, hence no computational advantage can be obtained to use the corresponding pairwise comparison likelihood.

Remark 2.2 When $p = o(n^\alpha)$, $\alpha < 1/2$, the asymptotic properties of the MPCLE can be shown by the use of the central limit theorem for m-dependent random variables (Berk, 1973).

3. Thurstone model for ranking with covariates

The consistency of the estimator $\hat{\boldsymbol{\lambda}}$ holds as the number of each document being ranked goes to infinity. In case when the number of documentations is large and the number of comparison is relatively small, the MPCLE may not have desired large sample properties.

Moreover, the Thurstone model without covariates is not adequate to assign the rank to new documentation.

To overcome these deficiency, in many studies of information retrieval, λ_j is parameterized with low dimensional covariates. Each document has information about author, length and number of keywords related to a given query in the document j, which are called the attributes of document. If the average relevance is a model of its attributes, such as $\lambda_j = x_j^T\boldsymbol{\beta}$ for $x_j \in \mathbb{R}^q$ and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_q)^T$, the probability of ranking conditioning on $x_1, \cdots, x_p \in \mathbb{R}^q$ is given by

$$P(R|x_1, \cdots, x_p) = P(\epsilon_1 + x_{(1)}^T\boldsymbol{\beta} > \cdots > \epsilon_p + x_{(p)}^T\boldsymbol{\beta})) \tag{3.1}$$

in the Thurstone model. (3.1) is called the constrained model with covariates (Critchlow and Flinger, 1991). This model allows a change of documents in the list for each searching.

Let $r_i = (r_{i1}, \cdots, r_{ip})$ be the independent observed ranking and $x_{ij} \in \mathbb{R}^q$ be the associated covariates of a document j in the ith search. Let $y_{ijk} = I(r_{ij} < r_{ik})$ and $l_{ijk}(\boldsymbol{\beta}) = y_{ijk}\log G((x_{ij} - x_{ik})^T\boldsymbol{\beta}) + (1 - y_{ijk})\log(1 - G((x_{ij} - x_{ik})^T\boldsymbol{\beta}))$, then the pairwise comparison loglikelihood is given by

$$l^{pw}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j<k} l_{ijk}(\boldsymbol{\beta}) \tag{3.2}$$

Let $\widehat{\boldsymbol{\beta}} = \text{argmax}_\beta l^{pw}(\boldsymbol{\beta})$ be the MPCLE. In the following subsection, we investigate the properties of $\widehat{\boldsymbol{\beta}}$.

3.1. Specification of G

Assume that x_{ij} for $i = 1, \cdots, n$, $j = 1, \cdots, p$ are iid random variables with distribution Q. Let $\mathbf{x} \sim \tilde{Q}$ and $y|\mathbf{x} \sim Bernoulli(G(\mathbf{x}^T\boldsymbol{\beta}))$, where $\tilde{Q}$ is a distribution of $x_{ij} - x_{ik}$. Note that $l_{ijk}(\boldsymbol{\beta})$ is independent of $l_{i'j'k'}(\boldsymbol{\beta})$ for all j', k' and $i \neq i'$. By Chebyshev's inequality,

$$\frac{1}{np(p-1)/2} l^{pw}(\boldsymbol{\beta}) \to_p E_{\mathbf{x},y}\, y\log(G(\mathbf{x}^T\boldsymbol{\beta})) + (1-y)\log(G(-\mathbf{x}^T\boldsymbol{\beta}))$$

as $n \to \infty$. We can prove that the consistency of $\widehat{\boldsymbol{\beta}}$ similarly as that of Theorem 2.1, which is stated in the following lemma without proof.

Lemma 3.1 Assume that $\sum_{i=1}^n \sum_{j<k} x_{ijk}x_{ijk}^T/n$ converges to a strictly positive definite matrix where $x_{ijk} = x_{ij} - x_{ik}$. If G is logconcave then $\widehat{\boldsymbol{\beta}}$ is consistent.

The Bradley-Terry model is a special version of the Thurstone model and its loglikelihood of the model is equally written as the pairwise comparison loglikelihood of the Thurstone model, (3.3), of which G is logistic distribution.

$$l^{pw}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j<k} y_{ijk}x_{ijk}^T\boldsymbol{\beta} + \log(1 + \exp(x_{ijk}^T\boldsymbol{\beta})). \tag{3.3}$$

Since G of the Thurston model is unknown, the assumed model for G should be checked, and it is difficult in practice (Pregibon, 1980). If misspecifying G can lead to a serious

bias of the estimated model, it is undesirable to use of the Bradley-Terry model in spite of its computational advantage without goodness of fit test for G. To clarify the influence of misspecifying G caused by use of the Bradley-Terry model, we investigate the consistency of $\widehat{\boldsymbol{\beta}}$, MPCLE of (3.3), and we obtain an optimistic result in misspecification problem of the Thurstone model. In the next theorem, the consistency of $\widehat{\boldsymbol{\beta}}$ is proved regardless of the true distribution of G.

Theorem 3.1 Assume that the probability model for ranking follows (3.1), and let $\boldsymbol{\beta}^*$ be the true parameter of the model. Assume two following conditions:

(i) r_i for $i = 1, \cdots, n$ are independent random samples.

(ii) $\tilde{Q}$ is a multivariate normal distribution with mean 0.

Then, $\widehat{\boldsymbol{\beta}}$ is consistent up to scale. That is, there exists a positive constant c such that

$$\|\hat{\boldsymbol{\beta}} - c\boldsymbol{\beta}^*\|_2 \to_p 0$$

as $n \to \infty$.

Proof: By (ii), there exists a constant c for each $\boldsymbol{\beta}$ such that $E(\mathbf{x}^T\boldsymbol{\beta}|\mathbf{x}^T\boldsymbol{\beta}^*) = \mathbf{x}^T(c\boldsymbol{\beta}^*)$. Let $l(y, \nu) = y\nu - \log(1 + \exp(\nu))$ for $y \in (0, 1)$, which is concave in $\nu \in \mathbb{R}$. Then, by Jensen's inequality, we obtain that

$$\begin{aligned} & E_{\mathbf{x},y} y\mathbf{x}^T\boldsymbol{\beta} + \log(1 + \exp(\mathbf{x}^T\boldsymbol{\beta})) \\ = \; & E_{\mathbf{x},y} y l(y, \mathbf{x}^T\boldsymbol{\beta}) \\ = \; & E_{\mathbf{x},y} E[l(y, \mathbf{x}^T\boldsymbol{\beta})|\mathbf{x}^T\boldsymbol{\beta}^*, y] \\ \geq \; & E_{\mathbf{x},y} l(y, E\mathbf{x}^T\boldsymbol{\beta}|\mathbf{x}^T\boldsymbol{\beta}^*) \\ = \; & E_{\mathbf{x},y} l(y, \mathbf{x}^T(c\boldsymbol{\beta}^*)). \end{aligned}$$

Combining the convexity lemma (Pollard, 1991), this inequality implies that $\widehat{\boldsymbol{\beta}}$ falls on the line with direction of $\boldsymbol{\beta}^*$ asymptotically.

Next, it suffices to prove that c is positive. Consider $h(a, b) = b\log(1/2 + a) + (1 - b)\log(1/2 - a)$, which is a bivariate function defined on $(-1/2, 1/2) \times (0, 1)$. It is easily shown that $\{0 < a < 1/2\} = \{a : h(a, b) > h(-a, b)\}$ for $b > 1/2$ and $\{-1/2 < a < 0\} = \{a : h(a, b) < h(-a, b)\}$ for $b < 1/2$. Through the function $h(a, b)$, the conditional expectation of $l(y, \mathbf{x}^T\boldsymbol{\beta})$ can be written by

$$\begin{aligned} & E[l(y, \mathbf{x}^T\boldsymbol{\beta})|\mathbf{x}] \\ = \; & G(\mathbf{x}^T\boldsymbol{\beta}^*)\log\left(\frac{\exp(\mathbf{x}^T\boldsymbol{\beta})}{1 + \exp(\mathbf{x}^T\boldsymbol{\beta})}\right) + (1 - G(\mathbf{x}^T\boldsymbol{\beta}^*))\log\left(\frac{1}{1 + \exp(\mathbf{x}^T\boldsymbol{\beta})}\right) \\ = \; & b(\mathbf{x}^T\boldsymbol{\beta}^*)\log(0.5 + a(\mathbf{x}^T\boldsymbol{\beta})) + (1 - b(\mathbf{x}^T\boldsymbol{\beta}^*)\log(1 - a(\mathbf{x}^T\boldsymbol{\beta})). \end{aligned}$$

Let $\tilde{c}$ be an arbitrary positive constant, then it follows that $E[l(y, \mathbf{x}^T(\tilde{c}\boldsymbol{\beta}^*)|\mathbf{x}] > E[l(y, \mathbf{x}^T(-\tilde{c}\boldsymbol{\beta}^*)|\mathbf{x}]$ for all $\mathbf{x}$. Therefore, it concludes that c is positive. □

The proof is almost same as that of theorem 3.1 of Li and Duan (1989) who consider the problem of the misspecification of the link function in generalized linear model. According to

the proof of Theorem 3.1, the consistency of $\widehat{\boldsymbol{\beta}}$ depends on the concavity of the misspecified model and the distribution of the design matrix. In the first condition, we know that Theorem 3.1 can be extended to wider concave models. However, in this paper, we focus on the Bradley-Terry model which is a specific concave model. We discuss some examples of other concave models in Section 3.2. The second condition is relatively restrictive, but it seems reasonable in the circumstance where the number of documents compared is huge.

We conduct a simple simulation to confirm the consistency of $\widehat{\boldsymbol{\beta}}$ under the Thurstone model. Let $x_{ij} \sim_{iid} N_q(0, \Sigma)$ and $\epsilon_{ij} \sim_{iid} N(0,1)$ for $i = 1, \cdots, n$, $j = 1, \cdots, p$, respectively. The ranking vector $r_i = (r_{i1}, \cdots, r_{ip})$ is given by $r_{ij} = \sum_{k=1}^{p} I(x_{ij}^T \boldsymbol{\beta}^* \leq x_{ik}^T \boldsymbol{\beta}^*)$. Let $p = 4$, $q = 3$, $\boldsymbol{\beta}^* = (\beta_1^*, \beta_2^*, \beta_3^*) = (1, 0.5, -1)^T$, and $\Sigma = I_3$. The estimates, $\widehat{\boldsymbol{\beta}}$, are obtained from 200 iterated simulations. Table 3.1 shows that the ratio of $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$. It confirms that the ratio converges to a constant as the number of observations is increasing. Note that the constant is not required to obtain the order of the relevances of interested documents to a given query.

Table 3.1 Consistency of the Bradley-Terry model

		number of observations		
		$n = 10^2$	$n = 10^3$	$n = 10^4$
mean (sd)	$\hat{\beta}_1/\beta_1^*$	.566 (.132)	.568 (.042)	.576 (.013)
	$\hat{\beta}_2/\beta_2^*$	.608 (.036)	.568 (.066)	.576 (.020)
	$\hat{\beta}_3/\beta_3^*$	.557 (.134)	.573 (.038)	.575 (.012)

3.2. Application to information retrieval system

The method of maximizing the pairwise comparison likelihood can be compared with the method of minimizing the empirical risk function with a surrogated loss function in classification. Let f^* be the ranking function satisfying $P(R_j < R_k | x_j, x_k) > 1/2$ if and only if $f^*(x_j) > f^*(x_k)$. For a given class $\mathcal{F}$ of functions including f^*, it can be shown that

$$f^* = \text{argmin}_{f \in \mathcal{F}} E\left\{ I(R_j < R_k)(f(x_j) - f(x_k)) < 0 \right\}.$$

A natural way of estimating f^* is to find a function minimizing the empirical risk

$$l_{0-1}(f) = \sum_{i=1}^{n} \sum_{j<k} I\{y_{ijk}(f(x_{ij}) - f(x_{ik})) < 0\}/n,$$

where $y_{ijk} = I(r_{ij} < r_{ik}) - I(r_{ij} > r_{ik})$. Since this empirical risk function is almost impossible to be minimized due to computational complexity, an alternative way is to use a convex surrogate loss ϕ of the 0-1 loss (Bartlett *et al.*, 2006; Zhang, 2004). That is, we estimate f by minimizing the surrogated empirical risk

$$l_\phi(f) = \sum_{i=1}^{n} \sum_{j<k} \phi\{y_{ijk}(f(x_{ij}) - f(x_{ik}))\}/n,$$

This idea is widely used in the researches of information retrieval system. The RankSVM (Joachims, 2003) uses the hinge loss function as the surrogated loss function. RankNet

(Burges *et al.*, 2005) and Rankboost (Freund *et al.*, 2003) adopt the logistic loss function and exponential loss function, respectively. Especially, RankNet is directly connected with the Thurstone model with the standard Gumbel distribution. While the three methods are based on the pairwise comparison surrogated loss functions, ListRank (Xia *et al.*, 2008) uses the full likelihood of the Thurstone model with Gumbel distribution.

Good performances of various ranking models in information retrieval are explained by Fisher consistency (Xia *et al.*, 2008) which means that the rankings are estimated consistently with surrogated loss functions. A limitation of Fisher consistency is that the class $\mathcal{F}$ of functions is sufficiently large so that all measurable functions are included. When only linear models are considered, Fisher consistency does not hold in general. However, through the proof of Theorem 3.1, we know that the estimators of the regression coefficients obtained from the methods based on pairwise comparison, RankSVM (Joachims, 2003), RankNet (Burges *et al.*, 2005), and Rankboost (Freund *et al.*, 2003), are consistent up to scale when $\mathcal{F}$ is linear function class. Hence, we can estimate the relevances of documents consistently. In this paper we do not show the proof, but we show the simulation results of some algorithms for information retrieval. Table 3.2 and Table 3.3 show the results of the RankSVM and RankNet by simulation, which is conducted in Section 3.2. The estimates $\widehat{\boldsymbol{\beta}}$ in two tables are obtained by RankSVM and RankBoost, respectively. As the number of observations are increasing, we also find the convergences of the ratios to some constants.

Table 3.2 Consistency of RankSVM

		number of observations		
		$n = 10^2$	$n = 10^3$	$n = 10^4$
mean (sd)	$\hat{\beta}_1/\beta_1^*$	0.743 (.175)	0.744 (0.05)	0.744 (.001)
	$\hat{\beta}_2/\beta_2^*$	0.865 (1.321)	0.752 (.102)	0.744 (.003)
	$\hat{\beta}_3/\beta_3^*$	0.761 (.192)	0.747 (.006)	0.744 (.016)

Table 3.3 Consistency of RankBoost

		number of observations		
		$n = 10^2$	$n = 10^3$	$n = 10^4$
mean (sd)	$\hat{\beta}_1/\beta_1^*$	1.064 (.275)	1.127 (0.09)	1.134 (.002)
	$\hat{\beta}_2/\beta_2^*$	1.278 (.954)	1.133 (.015)	1.131 (.004)
	$\hat{\beta}_3/\beta_3^*$	1.079 (.280)	1.125 (.009)	1.132 (.002)

4. Conclusion

We proposed the pairwise comparison likelihood to estimate the parameters in the Thurstone model, and proved the consistency of the MPCLE. Also, when there are covariates distributed according to the multivariate normal distribution, we can estimate the relevances of documents consistently based on the pairwise comparison likelihood, even when the distribution of the Thurstone model is misspecified. Conclusively, the main contribution of the paper is to give theoretical guidelines about ranking algorithms for IR. When there is no covariates, the distribution (i.e. G or equivalently the surrogated loss function) is selected with great care. When there are covariates which look like normally distributed, any reasonable surrogated loss can be used.

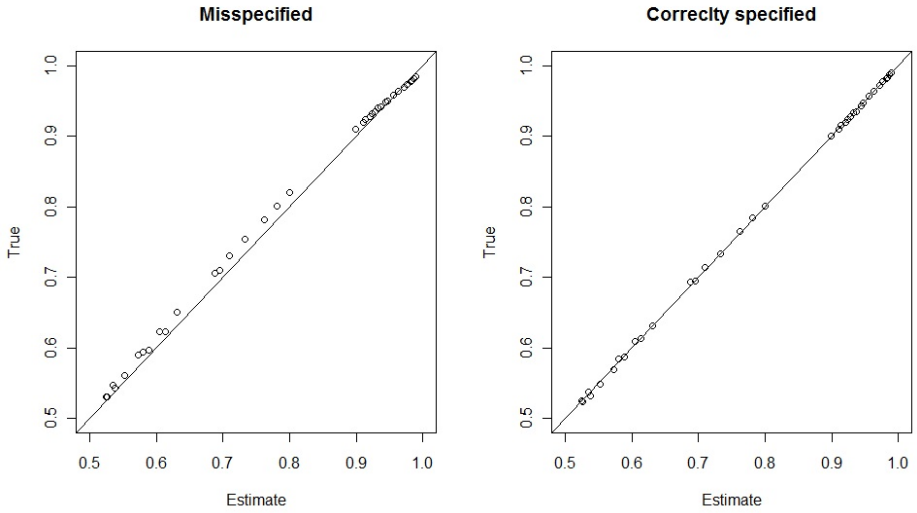

Figure 4.1 Plot of the estimated marginal probabilities

However, Theorem 3.1 does not depreciate the importance of specifying F correctly in the Thurstone model. Especially, under the Thurstone model with no covariate, the estimated marginal probability of the ranking can be biased under the misspecified model. Assume the Thurstone model with $F \sim N(0, 1/2)$. Let $\exp(\boldsymbol{\lambda}^*) = (10, 8, 6, 4.9, 4.6, 4.3, 1.2, 1.1, 1)$, and let the misspecified model be Bradley-Terry model. For $n = 10^4$, we obtain the estimated marginal probability for rankings. In Figure 4.1, the vertical and horizontal values of the dots are the true and estimated marginal probabilities in the ranking respectively. In the right panel, we found that the marginal probabilities are underestimated due to the model misspecification, which is known as bias of the model .

For incomplete design in the ranking model (Bradley and Terry, 1952), the correct model specification is important to estimate the marginal probability of ranking. In general, the pairwise comparison likelihood is hard to compute with an arbitrary symmetric function G. We leave developing a computationally efficient method in the Thurstone model with unknown F as a future work.

References

Bartlett, P. L., Jordan, M. I. and Mcauliffe, J. D. (2006). Convextity, classification, and risk bounds. *Journal of the American Statistical Association*, **101**, 138-156.

Berk, N. K. (1973). A central limit theorem for m-dependent random variables with unbouned m. *Annals of Probability*, **1**, 352-354.

Bradley, R. A. (1972). A biometrics invited paper. science, statistics, and paired comparisons. *Biometrcs*, **39**, 213-239.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, **39**, 324-345.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N. and Hullender, G. (2005). Learning to rank using gradient descent. *Proceedings of the 22nd International Conference on Machine Learning*, 89-96.

Critchlow E. E. and Flinger M. A. (1991). Paired comparison, triple comparison, and ranking experiments as generalized linear models, and their implementation on GLIM. *Psychometirka*, **56**, 517-533.

Critchlow E. E., Flinger M. A. and Verducci J. S. (1991). Probability models on rankings. *Journal of Mathematical Psychology*, **35**, 294-318.

Fligner, M. A. and Verducci, J. S. (1986). Distance based ranking models. *Journal of the Royal Statistical Society B*, **48**, 359-369.

Freund, Y., Iyer, R., Schapire, R. E. and Singer, Y. (2003). An efficient boosting algorithm for combining preferences. *The Journal of Machine Learning Research*, **4**, 933-963.

Hong C., Jung M. and Lee J. (2010). Prediction model analysis of 2010 South Africa World Cup. *Journal of the Korean Data & Information Science Society*, **21**, 1137-1146.

Joachims, T. (2003). Optimizing search engines using clickthrough data. *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 133-142.

Li K-C. and Duan N. (1989). Regression analysis under link violation. *Annals of Statistics*, **17**, 1009-1052.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*, John Wiley, New York, 213-239.

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

Plackett, R. L. (1975). The analysis of permutations. *Applied Statistics*, **24**, 193–202.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, **7**, 186-199.

Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Applied Statistics*, **29**, 15-24.

Stern, H. (1990). Models for distributions on permutations. *Journal of the American Statistical Association*, **85**, 558-564.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, **44**, 273-286.

Xia, F., Liu, T-Y., Wang, J., Zhang, W. and Li, H. (2008). Listwise approach to learning to rank: Theory and algorithm, *Proceedings of the 25th International Conference on Machine Learning*, 1192-1199.

Yan, T., Yang, Y. and Xu, J. (2010). Sparse paired comparisons in the Bradley-Terry model. *Statistica Sinica*, **22**, 1305-1318.

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, **32**, 56-134.