

---

# Computing Parametric Ranking Models via Rank-Breaking

---

Hossein Azari Soufiani

David C. Parkes

Harvard University, 33 Oxford Street, Cambridge, MA 02138 USA

Lirong Xia

Rensselaer Polytechnic Institute, Troy, NY 12180, USA

AZARI@FAS.HARVARD.EDU

PARKES@ECS.HARVARD.EDU

XIAL@CS.RPI.EDU

## Abstract

Rank breaking is a methodology introduced by Azari Soufiani et al. (2013a) for applying a Generalized Method of Moments (GMM) algorithm to the estimation of parametric ranking models. Breaking takes full rankings and breaks, or splits them up, into counts for pairs of alternatives that occur in particular positions (e.g., first place and second place, second place and third place). GMMs are of interest because they can achieve significant speed-up relative to maximum likelihood approaches and comparable statistical efficiency. We characterize the breakings for which the estimator is consistent for random utility models (RUMs) including Plackett-Luce and Normal-RUM, develop a general sufficient condition for a full breaking to be the only consistent breaking, and provide a trichotomy theorem in regard to single-edge breakings. Experimental results are presented to show the computational efficiency along with statistical performance of the proposed method.

## 1. Introduction

A standard approach to aggregation and inference with rank order data is to adopt a parametric model and use a maximum likelihood estimator (MLE) to fit model parameters. Based on these parameters we can then perform inference, for example, estimate the modal ranking. This approach has been widely studied in recent years in econometrics (Berry et al., 1995), computational social choice (Conitzer & Sandholm, 2005), and in rank learning settings (Liu, 2011).

However, for many parametric ranking models the MLE is

hard to compute. For example, computing MLE for the Mallows models is  $P_{\parallel}^{\text{NP}}$ -complete (Hemaspaandra et al., 2005). Among the *Random Utility Models (RUMs)*, only the *Plackett-Luce (PL)* model (Plackett, 1975; Luce, 1959) is known to have an analytical solution to the likelihood function. Some previous work has focused on computing specific parametric ranking models. For example, Hunter (2004) propose an *Minorize-Maximization (MM)* algorithm for MLE in the PL model. Others propose an *Monte-Carlo Expectation-Maximization (MC-EM)* algorithm to compute MLE for a general class of RUMs (Azari Soufiani et al., 2012). While this extends the computational reach to more expressive RUMs beyond PL, the running time may still be too large for data sets of practical interest.

An alternative to MLE is to adopt a *Generalized Method of Moments (GMM)* algorithm for estimation.<sup>1</sup> Azari Soufiani et al. (2013a) introduce the idea of *rank-breaking* as a way to apply GMM to full ranking data. In rank-breaking, each ranking in the data is decomposed into a subset of pairwise comparisons, to which GMM is then applied; e.g., one breaking might take as the statistics used for GMM a count of all pairs of alternatives that appear in first position and second position, another *full* breaking considers all possible pairs of positions.

Rank breaking is of interest because it can allow for estimation methods that are considerably faster than MLE. Azari Soufiani et al. (2013a) develop sufficient conditions for the breaking to be *consistent*, such that the GMM is consistent for PL. Consistency is a desired statistical property that says as the size of data generated according to a model within the class assumed by the estimator grows without bound, the output of the estimator converges to the true parameters. In addition, they provide experimental results that demonstrate high computational and statistical efficiency on both synthetic and real world data

---

*Proceedings of the 31<sup>st</sup> International Conference on Machine Learning*, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

---

<sup>1</sup>The method of Negahban et al. (2012) is in this spirit, proposing a graph-based Markov chain algorithm that provides a consistent estimator for pairwise-comparison data.

sets.

But left open was how to extend rank-breaking to other parametric ranking models beyond PL, and whether other consistent breakings beyond a full breaking exist for PL. Finding consistent, partial breakings is interesting because computing the statistics that are used for GMM becomes the bottleneck as the size of datasets grows. We address these questions. For the first question we propose a GMM algorithm (Algorithm 1) for any model in the *location family* of RUMs, which includes PL (Azari Soufiani et al., 2013a) and Normal-RUM and develop a general condition for when the breaking will provide a consistent estimator.

Based on this, and focusing on the location family, we:

(1) Characterize consistent breakings for PL, RUMs with a flipped Gumbel distribution, and RUMs with symmetric utility distributions, providing a negative answer to the second question of (Azari Soufiani et al., 2013a), and showing that for Normal-RUM the full breaking is the only consistent breaking.

(2) Provide a trichotomy theorem that characterizes what is required for *single-edge breakings*, which are simple breakings with only a particular pair of rank positions, to be consistent.

We conduct experimental studies to compare our algorithm to the MC-EM algorithm for RUMs. We consider RUMs with normal distributions and study running time and Kendall correlation. Experimental results show that our algorithm runs much faster than the MC-EM algorithm while achieving comparable, and sometimes even better Kendall correlation.

## 2. Preliminaries

Let  $\mathcal{A} = \{a_1, \dots, a_m\}$  denote the set of alternatives. Let  $D_r = (d_1, \dots, d_n)$  denote the data, where each  $d_j$  is a full ranking over  $\mathcal{A}$ . Let  $\mathcal{L}(\mathcal{A})$  denote the set of all full rankings (that is, all antisymmetry, transitive, and complete binary relationships) over  $\mathcal{A}$ . For any  $d \in \mathcal{L}(\mathcal{A})$  and any pair of alternatives  $a, a'$ , we  $a \succ_d a'$  if and only if  $a$  is preferred to  $a'$  in  $d$ , i.e.,  $(a, a') \in d$ . In a *parametric ranking model*  $\mathcal{M}_r$ , we let  $\Omega \subseteq \mathbb{R}^s$  denote the parameter space and for any  $\vec{\gamma} \in \Omega$ , let  $\Pr_{\mathcal{M}_r}(\cdot | \vec{\gamma})$  denote a distribution over  $\mathcal{L}(\mathcal{A})$ . Sometimes the subscript in  $\Pr_{\mathcal{M}_r}$  is omitted when it does not cause confusion.

### Random Utility Models (RUMs)

In a RUM, each alternative  $a$  is characterized by a utility distribution  $\mu_a$ , parameterized by a vector  $\vec{\gamma}_a$ . Given any ground truth  $\vec{\gamma} = (\vec{\gamma}_1, \dots, \vec{\gamma}_m)$ , an agent generates a full ranking over  $\mathcal{A}$  in the following way: she independently samples a random utility  $U_j$  for each alternative  $a_j$  with

conditional distribution  $\Pr_a(\cdot | \vec{\gamma}_a)$ , then ranks the alternatives according to their respective perceived utilities, such that she prefers  $a$  to  $a'$  if and only if  $U_a > U_{a'}$ .<sup>2</sup> The probability for a ranking  $d$  is the following, where  $d(j)$  is the index of the alternative ranked in the  $j$ th position:

$$\Pr(d | \vec{\gamma}) = \Pr(U_{d(1)} > U_{d(2)} > \dots > U_{d(m)})$$

In this paper, the *location family* refers to the class of RUMs where each distribution is only parameterized by its mean. In other words, the shapes of utility distributions are fixed, though they are not necessarily identical. A *homogeneous location family* is a location family where the shapes of the distributions are identical.<sup>3</sup> In this paper, we will study homogeneous location families with the following distributions:

- Gumbel distribution with  $\lambda = 1$ , whose PDF is  $\Pr_G = e^{-x}e^{-e^{-x}}$ : the corresponding homogeneous location family is PL.
- Flipped Gumbel distribution: the PDF is  $\Pr_G(-x)$ , where  $\Pr_G$  is the PDF of the Gumbel distribution with  $\lambda = 1$ . Flipped Gumbel is not the same as the Gumbel distribution. However it can be seen as a Gumbel distribution case where the smaller the  $x$  the better the alternative in ranking (e.g.  $x$  can be the time each horse takes to finish the race in a horse race competition).
- Normal distribution: no analytic solution to the likelihood function is known.

### Generalized Method-of-Moments

The *Generalized Method-of-Moments (GMM)*<sup>4</sup> provides a wide class of algorithms for parameter estimation. In GMM, we are given a parametric model whose parametric space is  $\Omega \subseteq \mathbb{R}^s$ , an infinite series of  $q \times q$  matrices  $\mathcal{W} = \{W_n : n \geq 1\}$ , and a column-vector-valued function  $g(d, \vec{\gamma}) \in \mathbb{R}^q$ . For any vector  $\vec{h} \in \mathbb{R}^q$  and any  $q \times q$  matrix  $W$ , we let  $\|\vec{h}\|_W = (\vec{h})^T W \vec{h}$ . For any data  $D_r$ , let  $g(D_r, \vec{\gamma}) = \frac{1}{n} \sum_{d \in D_r} g(d, \vec{\gamma})$ , and the GMM method computes parameters  $\vec{\gamma}' \in \Omega$  that minimize  $\|g(D_r, \vec{\gamma}')\|_{W_n}$ , formally defined as follows:

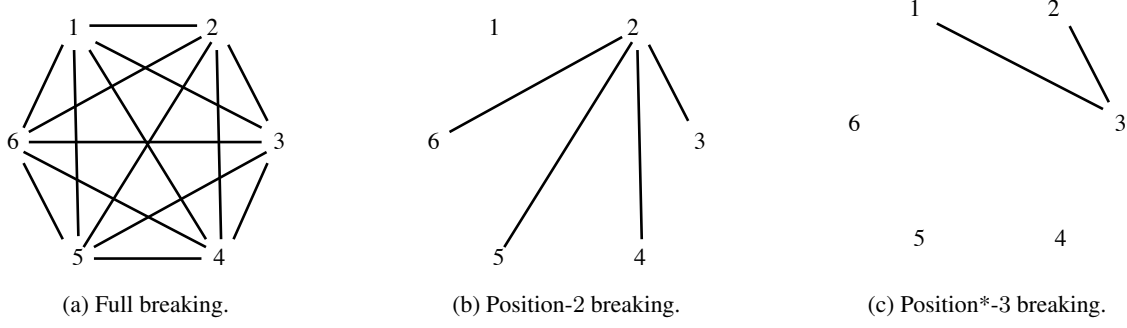
$$\begin{aligned} \text{GMM}_g(D_r, \mathcal{W}) = \\ \{\vec{\gamma}' \in \Omega : \|g(D_r, \vec{\gamma}')\|_{W_n} = \inf_{\vec{\gamma} \in \Omega} \|g(D_r, \vec{\gamma})\|_{W_n}\} \end{aligned} \quad (1)$$

Since  $\Omega$  may not be compact (as in PL), the set of parameters  $\text{GMM}_g(D_r, \mathcal{W})$  can be empty. A GMM is *consistent* if and only if for any  $\vec{\gamma}^* \in \Omega$ ,  $\text{GMM}_g(D_r, \mathcal{W})$  converges in probability to  $\vec{\gamma}^*$  as  $n \rightarrow \infty$  when the data is drawn

<sup>2</sup>We ignore the case of ties where  $U_a = U_{a'}$  since this happens with negligible probability for popular utility distributions.

<sup>3</sup>In this paper we will use  $\Pr(d | \vec{\gamma})$  and  $\Pr(d)$  interchangeably.

<sup>4</sup>Also known as *Z-estimators* (Vaart, 1998).


 Figure 1. Some breaking graphs for  $m = 6$ .

i.i.d. given  $\vec{\gamma}^*$ .

In this paper, we let  $W_n = I$  for all  $n$ . Let  $\|\cdot\|_2$  denote the L-2 norm. Equation (1) becomes

$$\text{GMM}_g(D_r) = \{\vec{\gamma}' \in \Omega : \|g(D_r, \vec{\gamma}')\|_2 = \inf_{\vec{\gamma} \in \Omega} \|g(D_r, \vec{\gamma})\|_2\} \quad (2)$$

### 3. Breakings

In this paper, a *rank-breaking* (*breaking* for short)  $B_G$  is defined as a function  $\mathcal{L}(\mathcal{A}) \rightarrow 2^{\{a \succ a' : a, a' \in \mathcal{A}\}}$  that is represented by an undirected breaking graph  $G$ , whose vertices are  $\{1, \dots, m\}$  that represents the  $m$  positions in a full ranking (rather than the subscripts of the  $m$  alternatives). For any full ranking  $d = [a_{i_1} \succ a_{i_2} \succ \dots \succ a_{i_m}]$ ,  $B_G(d) = \{a_{i_j} \succ a_{i_l} : a_{i_j} \succ_d a_{i_l} \text{ and } \{j, l\} \in G\}$ . That is,  $B_G$  breaks  $d$  into pairwise comparisons for all pairs of alternatives at position  $j$  and  $l$  such that  $\{j, l\}$  is an edge in  $G$ . If  $G$  only contains a single edge, then  $B_G$  is called a *single-edge breaking*.<sup>5</sup>

We extend  $B_G$  definition to apply to data  $D$ , so for any data  $D_r$  composed of full rankings, we let  $B_G(D_r) = \bigcup_{d \in D_r} B_G(d)$  where the union is in multiset sense.

We are interested in the following breakings, illustrated in Figure 1:

- **Full breaking:**  $G_F$  is the complete graph.
- **Position- $k$  breaking:** for any  $k \leq m - 1$ ,  $G_P^k = \{\{k, i\} : i > k\}$ .
- **Position\*- $k$  breaking:** for any  $k \geq 2$ ,  $G_{P^*}^k = \{\{k, i\} : i < k\}$ .

These breakings are of interest because they are easy to be

<sup>5</sup>The direction is implicit in graph  $G$ ; e.g., edge 2-4 will only ever generate a count for the alternative in position 2 being ahead of that in position 4. It doesn't also include a count for the one in position 4 being behind the one in position 2.

characterized analytically and we can generate other breakings using unions of them. We emphasize again that in a breaking  $B_G$ , the edges  $(j, l)$  in  $G$  represents the pairwise comparisons between the alternative ranked in *positions*  $j$  and *position*  $l$  of the input ranking, rather than  $a_j$  and  $a_l$ . Therefore, even though some edges are missing in  $G$ , it does not mean that some pairs of alternatives are never compared, since they can be compared in another ranking in the data where there is an edge in  $G$  between their corresponding positions.

**Example 1.** Let  $D_r = \{[a_1 \succ a_2 \succ a_3], [a_2 \succ a_1 \succ a_3]\}$ . We have:

$$B_{G_F}(D_r) = \{a_1 \succ a_2, a_1 \succ a_3, a_2 \succ a_3, a_2 \succ a_1, a_2 \succ a_3, a_1 \succ a_3\}.$$

$$B_{G_P^1}(D_r) = \{a_1 \succ a_2, a_1 \succ a_3, a_2 \succ a_1, a_2 \succ a_3\}.$$

$$B_{G_{P^*}^3}(D_r) = \{a_1 \succ a_3, a_2 \succ a_3, a_2 \succ a_1, a_1 \succ a_3\}.$$

### 4. A GMM Algorithm for the Location Family of RUM

We recall that in the location family, each utility distribution has only one parameter (its mean). Therefore, we can write  $\vec{\gamma} = (\gamma_1, \dots, \gamma_m)$ , where for any  $i \leq m$ ,  $\gamma_i$  is the mean parameter of the utility distribution for  $a_i$ . W.l.o.g. let  $\gamma_m = 0$ .

To specify the GMM, it suffices to specify the moment conditions. Given a parametric ranking model  $\mathcal{M}_r$  in the location family, for any two alternatives  $a \neq a'$ , any  $\vec{\gamma} \in \Omega$ , and any breaking  $B_G$ , we let  $f_G^{aa'}(\vec{\gamma})$  denote the probability that given  $\vec{\gamma}$ ,  $a \succ a'$  in  $B_G(d)$ . That is,  $f_G^{aa'}(\vec{\gamma}) = \Pr_{\mathcal{M}_r}(a \succ a' \in B_G(d) | \vec{\gamma})$ . When  $G = G_F$ , that is,  $G$  is the complete graph, we use shorthand notation  $f^{aa'} = f_G^{aa'}$ . Since the perceived utilities are generated independently,  $f^{aa'}$  is a function of  $\gamma_a - \gamma_{a'}$ . Therefore, we sometimes write  $f^{aa'}(\gamma_a - \gamma_{a'})$ . We note that in general  $f_G^{aa'}$  may depend on other components of  $\vec{\gamma}$ .

**Definition 1.** Given any breaking  $B_G$ , any  $d \in \mathcal{L}(\mathcal{A})$ , and

any  $a, a' \in \mathcal{A}$ , we let:

- $X_G^{a \succ a'}(d) = \begin{cases} 1 & a \succ a' \in B_G(d) \\ 0 & \text{otherwise} \end{cases}$ ,
- $X_G^{a \succ a'}(D_r) = \frac{1}{n} \sum_{d \in D_r} X_G^{a \succ a'}(d)$ , and

In words,  $X_G^{a \succ a'}(D_r)$  is the normalized frequency of times that alternative  $a$  is preferred to alternative  $a'$  (i.e.,  $a \succ a'$ ). By definition,  $E[X_G^{a \succ a'}(d)] = f_G^{aa'}$ . We now present the moment conditions used in our algorithm, and then comment on why we do not use other seemingly more natural ones. Our moment conditions are: for  $a \neq a'$ ,

$$g_G^{aa'}(d, \vec{\gamma}) = X_G^{a \succ a'}(d) \times f^{a'a}(\vec{\gamma}) - X_G^{a' \succ a}(d) \times f^{aa'}(\vec{\gamma}) \quad (3)$$

We are now ready to present our algorithm as Algorithm 1.

---

**Algorithm 1**  $\text{GMM}_G(D_r)$ 


---

For all  $a, a'$ , compute  $X_G^{a \succ a'}(D_r)$ .  
 Compute  $\text{GMM}_G(D_r)$  according to (2) using the moment conditions in (3) (e.g. using gradient descent).  
**return**  $\text{GMM}_G(D_r)$ .

---

We note that in (3) we use  $f^{aa'}$  and  $f^{a'a}$  instead of  $f_G^{aa'}$  and  $f_G^{a'a}$ . Therefore it is not immediately clear whether the moment conditions equal to 0 in expectation for a graph  $G$  that is not the complete graph. The next definition provides a condition used to guarantee that when a consistent breaking  $G$  is used in Algorithm 1, the moment conditions (3) equal to 0 in expectation.

**Definition 2.** A breaking  $B_G$  is consistent for a location family RUM, if  $G$  has at least one edge and for any  $\vec{\gamma}$  and any  $a \neq a'$ ,<sup>6</sup>

$$\frac{f_G^{aa'}(\vec{\gamma})}{f_G^{a'a}(\vec{\gamma})} = \frac{f^{aa'}(\vec{\gamma})}{f^{a'a}(\vec{\gamma})}$$

Where,

$$\frac{f^{aa'}(\vec{\gamma})}{f^{a'a}(\vec{\gamma})} = \frac{\Pr_{\mathcal{M}_r}(a \succ a' | \vec{\gamma})}{\Pr_{\mathcal{M}_r}(a' \succ a | \vec{\gamma})}$$

We will be interested in understanding when breakings are consistent. By definition, the full breaking is consistent. Let  $\text{CDF}_a$  denote the CDF of  $\Pr_a(\cdot | 0)$ . For the location family we have:

<sup>6</sup>The definition of consistent breakings is more general than the definition in (Azari Soufiani et al., 2013a), which was defined only for PL.

$$f^{aa'}(\vec{\gamma}) = f^{aa'}(\gamma_a - \gamma_{a'}) = \int_{-\infty}^{\infty} \Pr_{a'}(y)(1 - \text{CDF}_a(y - \gamma_a + \gamma_{a'})) dy \quad (4)$$

We have the following proposition for  $f^{aa'}(\gamma_a - \gamma_{a'})$ . All omitted proofs can be found in Appendix A in the supplement material.

**Proposition 1.** For any model in the location family where each utility distribution has support  $(-\infty, \infty)$ ,  $f^{aa'}$  is monotonic increasing (as a function of  $\gamma_a - \gamma_{a'}$ ) on  $(-\infty, \infty)$  with  $\lim_{x \rightarrow -\infty} f^{aa'}(x) = 0$  and  $\lim_{x \rightarrow \infty} f^{aa'}(x) = 1$ . Moreover, if  $\Pr_a$  and  $\Pr_{a'}$  are continuous then  $f^{aa'}$  is continuously differentiable with  $f^{aa'}(x)' = \int_{-\infty}^{\infty} \Pr_{a'}(y) \Pr_a(y - x) dy$ .

**Theorem 1.** For any model in the location family with (possibly) inhomogeneous distributions and any consistent breaking  $B_G$ , if the PDF of every utility distribution is continuous, then Algorithm 1 is consistent.

*Proof.* We prove the theorem by verifying the conditions in Theorem 2.1 in (Hansen, 1982).

Assumption 2.1: The distribution on  $D$  is stationary and ergodic. This holds because in any RUM, data in  $D$  are generated i.i.d.

Assumption 2.2:  $\Omega$  is a separable metric space. Since  $\mathbb{R}^m$  is a metric separable space and  $\Omega$  is an subset of  $\mathbb{R}^m$ ,  $\Omega$  is also separable.

Assumption 2.3:  $g_G^{aa'}(\cdot, \vec{\gamma})$  is Borel measurable for any  $a \neq a'$  and each  $\vec{\gamma} \in \Omega$  and  $g_G^{aa'}(d, \cdot)$  is continuous on  $\Omega$  for each  $d$ . Since the domain of  $g_G^{aa'}(\cdot, \vec{\gamma})$  is discrete,  $g_G^{aa'}(\cdot, \vec{\gamma})$  is continuous, which means that  $g_G^{aa'}(\cdot, \vec{\gamma})$  is Borel measurable. We note that  $g_G^{aa'}(d, \cdot)$  is linear in  $f^{aa'}(\vec{\gamma})$  and by Proposition 1,  $f^{aa'}$  is continuous in  $\vec{\gamma}$ .

Assumption 2.4:  $E_{d|\vec{\gamma}^*}[g_G^{aa'}(d, \vec{\gamma})]$  exists and is finite for all  $\vec{\gamma} \in \Omega$ , and  $E_{d|\vec{\gamma}^*}[g_G^{aa'}(d, \vec{\gamma}^*)] = 0$ . The former is because  $E_{d|\vec{\gamma}^*}[g_G^{aa'}(d, \vec{\gamma})]$  is linear in  $f^{aa'}(\vec{\gamma})$  and by Proposition 1,  $f^{aa'}(\Omega)$  is bounded above by 1. The second part holds because  $E_{d|\vec{\gamma}^*}[X_G^{a \succ a'}(d)] = f_G^{aa'}(\vec{\gamma}^*)$ , which means that  $E_{d|\vec{\gamma}^*}[g_G^{aa'}(d, \vec{\gamma}^*)] = f_G^{aa'}(\vec{\gamma}^*)f^{a'a}(\vec{\gamma}^*) - f_G^{a'a}(\vec{\gamma}^*)f^{aa'}(\vec{\gamma}^*) = 0$ .

Assumption 2.5: The sequence  $\mathcal{W}$  converges almost surely to a positive semi-definite matrix. This holds since  $W_n = I$  for all  $t$ .

Premise (1):  $g_G^{aa'}(d, \vec{\gamma})$  is first moment continuous. Since  $|g_G^{aa'}(d, \vec{\gamma})| \leq 2$ , by Lemma 2.1 of (Hansen, 1982), we have that  $g_G^{aa'}(d, \vec{\gamma})$  is first moment continuous.



Premise (2):  $\Omega$  is compact, which is the assumption of our theorem.

Premise (3):  $E_{d|\vec{\gamma}^*}[g_G^{aa'}(d, \vec{\gamma})]$  has a unique zero at  $\vec{\gamma}^*$ . By Proposition 1 we have that  $f^{aa'}(\gamma_a - \gamma_{a'})$  is monotonic increasing in  $\gamma_a - \gamma_{a'}$  and  $f^{a'a}(\gamma_{a'} - \gamma_a)$  is monotonic increasing in  $\gamma_{a'} - \gamma_a$ . Therefore,  $\frac{f^{aa'}(\gamma_a - \gamma_{a'})}{f^{a'a}(\gamma_{a'} - \gamma_a)}$  is monotonic increasing in  $\gamma_a - \gamma_{a'}$ . Hence if  $\vec{\gamma}'$  is another zero point for  $E_{d|\vec{\gamma}^*}[g_G^{aa'}(d, \vec{\gamma})]$  with  $\gamma'_m = 0$ , then we must have that for all pairs  $(a, a')$ ,  $\gamma'_a - \gamma'_{a'} = \gamma_a^* - \gamma_{a'}^*$ . Given that  $\gamma'_m = \gamma_m^* = 0$ , this means that  $\vec{\gamma}' = \vec{\gamma}^*$ , which is a contradiction. Therefore,  $\vec{\gamma}^*$  is the only zero point of  $E_{d|\vec{\gamma}^*}[g_G^{aa'}(d, \vec{\gamma})]$ .  $\square$

A direct result of the above theorem, for any consistent breaking  $B_G$  for PL, RUM with flipped Gumbel distributions, and RUM with normal distributions (e.g. the full breaking), Algorithm 1 is consistent for PL, RUM with flipped Gumbel distributions, and RUM with normal distributions respectively.

Compared to the MC-EM algorithm (Azari Soufiani et al., 2012), Algorithm 1 runs faster since optimizing Equation (2) is much easier through e.g., gradient descent or Newton-Raphson. This is because  $f^{aa'}(x)'$  is usually easy to compute, and sometimes has a concise analytic solution, as shown in the following example. Breaking is particularly helpful here since it enables analytic expression for gradient.

**Example 2.** Consider RUM with normal distributions whose variances are 1. For any  $a \neq a'$  we have:

$$f^{aa'}(x)' = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} e^{-\frac{(y-x)^2}{2}} dy = \frac{1}{2\sqrt{\pi}} e^{-\frac{x^2}{4}}$$

A similar formula exists for location families with normal distributions whose variances are not identical.

**Why do we use the moment conditions in (3)?** The following moment conditions seem to be more natural.

$$g_G^{aa'}(d, \vec{\gamma}) = X_G^{a \succ a'}(d) \times f_G^{a'a}(\vec{\gamma}) - X_G^{a' \succ a}(d) \times f_G^{aa'}(\vec{\gamma}) \quad (5)$$

The only difference between (5) and (3) is that the former uses  $f_G^{a'a}$  and  $f_G^{aa'}$  while the latter uses  $f^{aa'}$  and  $f^{a'a}$ . However, for models in the location family, usually optimizing (5) is hard due to the lack of analytical solutions to  $f_G^{aa'}$  or  $(f_G^{aa'})'$ . As shown in Example 2,  $(f^{aa'})'$  is easy to compute. This is the main reason we choose (3) over (5).

**Why are we interested in breakings beyond the full breaking?** The optimization problem (2) is

$m$ -dimensional, but requires as input the counts in equation 3 to be computed for every ordered pair of alternatives. Computing these counts scales a  $O(m^2n)$  for full breaking but as  $O(mn)$  for adjacent breaking or position- $k$  breaking. For large  $n$  this can become the bottleneck with the difference between  $O(m^2n)$  and  $O(mn)$  making a meaningful difference and starting to become the bottleneck in computation Azari Soufiani et al. (2013a). In such cases we may would prefer to use a partial breaking and explore the tradeoff between computational efficiency and statistical efficiency. However, it is important to do this while maintaining consistency of the estimator.

## 5. Which Breakings are Consistent?

This section provides theoretical results on the consistency of partial breakings (breakings which take only part of the available ranks) for the location family. We will first present the theorems, then introduce four lemmas in Section 5.1, and finally in Section 5.2 use them as building blocks to provide proofs for the theorems. We start with the following positive results.

**Theorem 2.** For PL, a breaking  $B_G$  is consistent if and only if  $G$  is the union of position- $k$  breakings.

In a similar way the following Theorem holds if we change PL to PL\*.

**Theorem 3.** For the RUM with flipped Gumbel distributions (PL\*),  $B_G$  is consistent if and only if  $G$  is the union of position\*- $k$  breakings.

Theorem 2 gives a complete characterization of consistent breakings for PL (thus answering an open question in (Azari Soufiani et al., 2013a)) and Theorem 3 gives a complete characterization of consistent breakings for the RUM with flipped Gumbel distributions.

**Theorem 4.** Let  $\mathcal{M}_r$  be a model in the (possibly) inhomogeneous location family where each utility distribution has support  $(-\infty, \infty)$ . If the PDF of each utility distribution in  $\mathcal{M}_r$  is symmetric around its mean, then the only consistent breaking is the full breaking.

Since the normal distribution is symmetric, we immediately have the following corollary of Theorem 4.

**Corollary 1.** For the RUM with normal distributions (the variances are not necessary identical), the only consistent breaking is the full breaking.

Theorem 4 and Corollary 1 tell us that for certain natural models in the location family, the only consistent breaking is the full breaking. This will also be demonstrated by experimental results in the next section. The next theorem provides a quick check to see if the full-break is the only consistent breaking by just checking the  $m = 3$  case.

**Theorem 5.** For any model in the homogeneous location family where each utility distribution has support  $(-\infty, \infty)$ , if the full breaking is the only consistent breaking for  $m = 3$ , then the full breaking is the only consistent breaking for any  $m$ .

The last result of this section is a trichotomy theorem for single-edge breakings to be consistent for the homogeneous location family.

**Theorem 6.** For any  $m$  and any model in the homogeneous location family (with support  $(-\infty, \infty)$ ), one and exactly one of the following holds.

1. No single-edge breaking is consistent.
2. Among all single-edge breakings, only  $\{1, 2\}$  is consistent.
3. Among all single-edge breakings, only  $\{m - 1, m\}$  is consistent.

This theorem corresponds to a symmetry notion in the specific location family. Using this theorem and Theorem 4 we know that case (1) corresponds to the symmetric location families and we conjecture that the cases (2) and (3) correspond to negative and positive skewness in the location family distributions respectively.

The next example shows that each of the three cases in Theorem 6 (but not any two of them) holds for some natural location family.

**Example 3.** By Corollary 1, the location family with normal distributions belongs to Case 1 in Theorem 6; by Theorem 2, PL belongs to Case 2 in Theorem 6; by Theorem 3,  $PL^*$  belongs to Case 3 in Theorem 6.

### 5.1. Four Core Lemmas

To prove the theorems we introduce some notation and four core lemmas in this subsection. For any model  $\mathcal{M}_r$  in the location family, let  $\mathcal{M}_r^*$  denote the model in the location family where the PDF of each distribution (conditioned on the mean parameter being 0) is flipped around the y-axis. That is, for any  $i \leq m$  and any  $x$ ,  $\Pr_{\mathcal{M}_r^*, i}(x|0) = \Pr_{\mathcal{M}_r, i}(-x|0)$ . For any breaking  $B_G$ , we let  $B_{G^*}$  denote the breaking such that  $(i, j) \in G^*$  if and only if  $(m + 1 - i, m + 1 - j) \in G$ .

**Example 4.**  $PL^*$  is the RUM with flipped Gumbel distribution. Let  $\mathcal{M}_N$  denote the RUM with normal distributions. We have  $\mathcal{M}_N = \mathcal{M}_N^*$ . For any  $k \geq 2$ , we have  $(G_P^k)^* = G_{P^*}^{m-k}$ .

**Lemma 1.** For any  $\mathcal{M}_r$  in the location family, if  $B_G$  is consistent for  $\mathcal{M}_r$ , then  $B_{G^*}$  is consistent for  $\mathcal{M}_r^*$ .

For any graph  $G$  and any  $1 \leq k_1 < k_2 \leq m$ , we let  $G_{[k_1, k_2]}^*$  denote the subgraph of  $G$  where the vertices  $1, \dots, k_1 - 1$

and  $k_2 + 1, \dots, m$  are removed, and the vertices are renamed to  $1, \dots, k_2 + 1 - k_1$  by subtracting  $k_1 - 1$  from all vertices.

**Example 5.** For  $m = 6$ , a breaking  $B_G$  and its restriction to  $[2, 4]$  are shown in Figure 2.

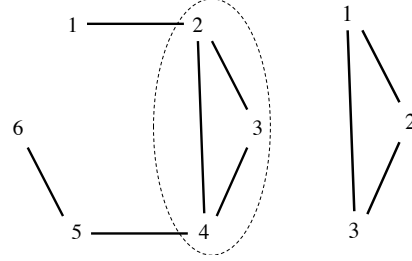


Figure 2. A breaking graph  $G$  and  $G_{[2,4]}$  for  $m = 6$ .

**Lemma 2.** For any model  $\mathcal{M}_r$  in the location family, if  $B_G$  is consistent then for any  $1 \leq k_1 < k_2 \leq m$ , either  $G_{[k_1, k_2]} = \emptyset$ , or  $B_{G_{[k_1, k_2]}}$  is consistent for any location family for  $k_2 - k_1 + 1$  alternatives where the utility distributions can be any combination of  $k_2 - k_1 + 1$  utility distributions in  $\mathcal{M}_r$ .

**Lemma 3.** For any location family where each utility distribution has support  $(-\infty, \infty)$ , the single-edge breaking  $B_{\{1, m\}}$  is not consistent.

The last lemma (specifically, part (3), (4), (5)) is a natural extension of Theorem 4 in (Azari Soufiani et al., 2013a).

**Lemma 4.** Let  $B_{G_1}, B_{G_2}$  be a pair of breakings.

- Suppose both  $B_{G_1}$  and  $B_{G_2}$  are consistent,
  - (1) if  $G_1 \cap G_2 = \emptyset$ , then  $B_{G_1 \cup G_2}$  is also consistent;
  - (2) if  $G_1 \subsetneq G_2$ , then  $B_{G_2 \setminus G_1}$  is also consistent.
- Suppose  $B_{G_1}$  is consistent but  $B_{G_2}$  is not consistent,
  - (3) if  $G_1 \cap G_2 = \emptyset$ , then  $B_{G_1 \cup G_2}$  is not consistent;
  - (4) if  $G_1 \subsetneq G_2$ , then  $B_{G_2 \setminus G_1}$  is not consistent.
  - (5) if  $G_2 \subsetneq G_1$ , then  $B_{G_1 \setminus G_2}$  is not consistent.

*Proof.* The proof is based on the following two observations. 1) If  $G_1 \cap G_2 = \emptyset$ , then  $f_{G_1 \cup G_2}^{aa'}(d) = f_{G_1}^{aa'}(d) + f_{G_2}^{aa'}(d)$  and  $X_{G_1 \cup G_2}^{a \succ a'}(d) = X_{G_1}^{a \succ a'}(d) + X_{G_2}^{a \succ a'}(d)$ . 2) If  $G_1 \subsetneq G_2$ , then  $f_{G_1 \setminus G_2}^{aa'}(d) = f_{G_1}^{aa'}(d) - f_{G_2}^{aa'}(d)$  and  $X_{G_1 \setminus G_2}^{a \succ a'}(d) = X_{G_1}^{a \succ a'}(d) - X_{G_2}^{a \succ a'}(d)$ .  $\square$

### 5.2. Proofs of the Theorems

We are now ready to prove the theorems in this section.

*Proof of Theorem 2.* The “if” direction was proved in (Azari Soufiani et al., 2013a). We now prove the “only if” part by induction on  $m$ . When  $m = 3$ , the theorem

obviously holds. Suppose the theorem holds for  $l$ . When  $m = l + 1$ , we first apply Lemma 2 to  $G_{[2,m]}$ . By the induction hypothesis,  $G_{[2,m]}$  must be the union of position- $k$  breakings for some  $k \geq 2$ . Now apply Lemma 2 to  $G_{[1,m-1]}$ . There are two cases.

**Case 1:** for all  $i \leq m - 1$ ,  $\{1, i\} \in G$ . We claim that  $\{1, m\} \in G$ . This is because  $B_{\{1,m\} \cup G}$  is consistent, and  $B_{\{1,m\}}$  is not consistent due to Lemma 3. Hence  $B_{G \setminus \{1,m\}}$  is not consistent.

**Case 2:** for all  $i \leq m - 1$ ,  $\{1, i\} \notin G$ . In this case  $\{1, m\} \notin G$  following a similar argument as in Case 1.

This means that the theorem holds for  $m = l + 1$ , which proves the theorem.  $\square$

*Proof of Theorem 3.* The proof follows immediately after Theorem 2 and Lemma 1.  $\square$

*Proof of Theorem 4.* Let  $B_G$  denote a consistent breaking. We prove the theorem by induction on  $m$ . When  $m = 3$ , the full breaking is consistent and by Lemma 3, the single edge-breaking  $B_{\{(1,3)\}}$  is not consistent. By Lemma 4 part (5),  $B_{\{(1,2),(2,3)\}}$  is not consistent.

We now prove that the single-edge breaking  $B_{\{(1,2)\}}$  is not consistent. For the sake of contradiction suppose it is. By Lemma 1,  $B_{\{(1,2)\}}^* = B_{\{(2,3)\}}$  is consistent for  $\mathcal{M}_r^*$ . Since all utility distributions in  $\mathcal{M}_r$  are symmetric,  $\mathcal{M}_r^* = \mathcal{M}_r$ . Therefore,  $B_{\{(2,3)\}}$  is consistent for  $\mathcal{M}_r$ . By Lemma 4 part (1),  $B_{\{(1,2),(1,3)\}}$  is consistent, which is a contradiction.

Similarly the single-edge breaking  $B_{\{(2,3)\}}$  is not consistent. It follows from Lemma 4 part (5) that  $B_{\{(1,2),(1,3)\}}$  and  $B_{\{(1,3),(2,3)\}}$  are not consistent. Therefore, the only consistent breaking for  $m = 3$  is the full breaking.

Suppose the theorem holds for  $m = l$ . When  $m = l + 1$ , we first apply Lemma 2 to  $G_{[2,m]}$  and  $G_{[1,m-1]}$ . By the induction hypothesis,  $G_{[2,m]}$  ( $G_{[1,m-1]}$ ) is either empty or the full graph. We have the following two cases.

Since  $m > 3$ , if  $G_{[2,m]}$  is empty, then  $G_{[1,m-1]}$  is empty as well. Since  $G$  is non-empty,  $G = \{(1, m)\}$ , which contradicts Lemma 3.

If  $G_{[2,m]}$  is full, then  $G_{[1,m-1]}$  is full as well. Hence  $G$  can be either the full graph  $G_F$ , or  $G_F \setminus \{(1, m)\}$ . By Lemma 3,  $B_{\{(1,m)\}}$  is inconsistent, which means that  $B_{G_F \setminus \{(1,m)\}}$  is not consistent (Lemma 4 part (5)).

Therefore, the only remaining case is that  $G$  is the full breaking, which means that the theorem holds for  $m = l + 1$ , which proves the theorem.  $\square$

*Proof of Theorem 5.* The proof is similar to the proof of

Theorem 4. We prove the theorem by induction on  $m$ .  $m = 3$  is the assumption. Suppose the theorem holds for  $l$ . When  $m = l + 1$ , we first apply Lemma 2 to  $G_{[2,m]}$ . By the induction hypothesis,  $G_{[2,m]}$  is either empty or full.

If  $G_{[2,m]}$  is empty, then  $G_{[1,m-1]}$  is empty as well. Hence if  $G$  is non-empty, then  $G = \{(1, m)\}$ , which contradicts Lemma 3.

If  $G_{[2,m]}$  is full, then  $G_{[1,m-1]}$  is full as well. Hence  $G$  can be either the full graph  $G_F$ , or  $G_F \setminus \{(1, m)\}$ . By Lemma 3,  $B_{\{(1,m)\}}$  is inconsistent, which means that  $B_{G_F \setminus \{(1,m)\}}$  is inconsistent (since  $G_F$  is always consistent by definition).

Therefore, the theorem holds for  $m = l + 1$ , which completes the proof.  $\square$

*Proof of Theorem 6.* For any  $k_2 > k_1 + 1$ , let us first consider  $G_{[k_1, k_2]}$ . By Lemma 3,  $B_{\{(1, k_2 - k_1 + 1)\}}$  is not consistent. Therefore by Lemma 2, any non-adjacent single-edge breaking is not consistent.

Now for an adjacent single-edge graph  $\{(k_1, k_1 + 1)\}$  that is different from  $\{(1, 2)\}$  and  $\{(m - 1, m)\}$ , by applying Lemma 2 on  $G_{[k_1 - 1, k_1 + 1]}$  and  $G_{[k_1, k_1 + 2]}$ , we have that both  $B_{\{(1,2)\}}$  and  $B_{\{(2,3)\}}$  are consistent for the model in the location family with  $m = 3$  and any combination of 3 utility distributions in  $\mathcal{M}_r$ . By Lemma 4 part (1),  $\{(1, 2), (2, 3)\}$  is consistent, which contradicts Lemma 4 part (5) applied to Lemma 3.

Now, we only need to prove that it is impossible for both  $B_{\{(1,2)\}}$  and  $B_{\{(m-1,m)\}}$  to be consistent. If on the contrary both are consistent, then we apply Lemma 2 on  $G_{[1,3]}$  and  $G_{[m-2,m]}$ . Following a similar argument as in the previous paragraph, we can show a contradiction. This proves the theorem.  $\square$

We conjecture that the converse of Theorem 1 holds for natural models in the location family.

## 6. Experiments

We implemented the MC-EM algorithm (Azari Soufiani et al., 2012), Algorithm 1 with the full breaking, and Algorithm 1 with top-3 breaking for the normal with fixed variance. We evaluate the three algorithms according to run-time and the following two representative criteria. For this, let  $\vec{\gamma}^*$  denote the ground truth parameters, and  $\vec{\gamma}$  denote the output of the algorithm.

- *Kendall Rank Correlation Coefficient:* Let  $K(\vec{\gamma}, \vec{\gamma}^*)$  denote the *Kendall tau distance* between the ranking over components in  $\vec{\gamma}$  and the ranking over components in  $\vec{\gamma}^*$ . The Kendall correlation is  $1 - 2 \frac{K(\vec{\gamma}, \vec{\gamma}^*)}{m(m-1)/2}$ .

The synthetic datasets are generated as follows. Let  $m = 5$ . The ground truth  $\vec{\gamma}^*$  is generated from the Dirich-

$n$	F–T	M–T	M–F	F–T	M–T	M–F
5	$-10^{-4}$ ( $10^{-3}$ )	<b>17 (.05)</b>	<b>17 (.05)</b>	.09 (.55)	.08 (.57)	<b>-.01 (.001)</b>
50	.004 (.005)	<b>198 (1.3)</b>	<b>198 (1.3)</b>	.27 (.4)	.26 (.37)	<b>-.01 (.001)</b>
100	<b>.008 (.0005)</b>	<b>359 (11)</b>	<b>359 (11)</b>	.08 (.08)	.04 (.08)	<b>-.04 (.004)</b>
150	<b>.035 (.004)</b>	<b>970 (10)</b>	<b>970 (10)</b>	<b>.34 (.1)</b>	<b>.33 (.11)</b>	<b>-.01 (.001)</b>
200	<b>.017 (.0015)</b>	<b>1021 (31)</b>	<b>1021 (31)</b>	<b>.29 (.027)</b>	<b>.27 (.022)</b>	<b>-.02 (.0057)</b>

(a) Run time (seconds).

(b) Kendall correlation.

Table 1. Paired t-tests for the three algorithms. F, T, M represents values for full breaking, top-3 breaking, and MC-EM, respectively. Mean (std) are shown. Significance results with 95% confidence are in bold.

let distribution  $\text{Dirichlet}(\vec{1})$  which is a distribution on an  $m$ -dimensional unit simplex. Then, for any given  $\vec{\gamma}^*$  we generate up to  $n = 200$  full rankings from the location family with normal distributions. All experiments are run on a 2.4 Ghz, Intel Core 2 duo 32 bit laptop.

Table 1 (a) shows the paired t-test on running time for the three methods for  $n = 5, 50, 100, 150, 200$ , where F, T, M represents values for full breaking, top-3 breaking, and MC-EM, respectively. We clearly observe that the running time of Algorithm 1 with full breaking and Algorithm 1 with top-3 breaking are significantly lower than the running time of MC-EM.

Table 1 (b) show paired t-tests for the three methods, for Kendall correlation. We note that a higher Kendall correlation means that the estimation is more accurate. Surprisingly, for Kendall correlation, Algorithm 1 with full breaking outperforms MC-EM with 95% confidence for almost all  $n$  in our experiments despite that Algorithm 1 runs much faster. Both algorithms are significantly better than Algorithm 1 with top-3 breaking with 95% confidence when  $n$  is not too small. The latter observation is because Algorithm 1 with top-3 breaking is not consistent for the location family with normal distributions.

## 7. Conclusions

This paper studies consistency of rank breaking for random utility models and provides a fast algorithm to compute parameters of these models. The code is provided in the R package “StatRank” (Chen & Azari Soufiani, 2013). We plan to extend the algorithms and analysis to partial orders, non-location families such as RUMs parameterized by mean and variance, and to GRUMs (Azari Soufiani et al., 2013c) and GRUMs with multiple types (Azari Soufiani et al., 2013b). We also plan to study possible connections between other rank aggregation methods e.g. (Ammar & Shah, 2011) and GMMs and their extension.

## Acknowledgments

Hossein Azari Soufiani is a Siebel scholar. This work is supported in part by NSF Grants No. CCF-0915016 and No. AF-1301976. Lirong Xia acknowledges NSF under

Grant #1136996 to the Computing Research Association for the CIFellows Project and an RPI startup fund for support. Part of this work was done when Lirong Xia was visiting Chinese Academy of Sciences supported by the Overseas Team Program of Academy of Mathematics and Systems Science (AMSS), Chinese Academy of Sciences, NSFC 61232015, and NSFC 61073023. We thank Shivani Agarwal for her comments on different aspects of this work. We thank anonymous ICML-14 reviewers, for helpful comments and suggestions.

## References

- Ammar, Ammar and Shah, Devavrat. Ranking: Compare, don’t score. In *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*, pp. 776–783. IEEE, 2011.
- Azari Soufiani, Hossein, Parkes, David C., and Xia, Lirong. Random utility theory for social choice. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 126–134, Lake Tahoe, NV, USA, 2012.
- Azari Soufiani, Hossein, Chen, William, Parkes, David C., and Xia, Lirong. Generalized method-of-moments for rank aggregation. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, Lake Tahoe, NV, USA, 2013a.
- Azari Soufiani, Hossein, Diao, Hansheng, Lai, Zhenyu, and Parkes, David C. Generalized random utility models with multiple types. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, Lake Tahoe, NV, USA, 2013b.
- Azari Soufiani, Hossein, Parkes, David C., and Xia, Lirong. Preference Elicitation For General Random Utility Models. In *Proceedings of Uncertainty in Artificial Intelligence (UAI)*, Bellevue, Washington, USA, 2013c.
- Berry, Steven, Levinsohn, James, and Pakes, Ariel. Automobile prices in market equilibrium. *Econometrica*, 63 (4):841–890, 1995.



- Chen, William Z. and Azari Soufiani, Hossein. *StatRank*, 2013. URL <http://cran.r-project.org/web/packages/StatRank/index.html>.
- Conitzer, Vincent and Sandholm, Tuomas. Common voting rules as maximum likelihood estimators. In *Proceedings of the 21st Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 145–152, Edinburgh, UK, 2005.
- Hansen, Lars Peter. Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, 50 (4):1029–1054, 1982.
- Hemaspaandra, Edith, Spakowski, Holger, and Vogel, Jörg. The complexity of Kemeny elections. *Theoretical Computer Science*, 349(3):382–391, December 2005.
- Hunter, David R. MM algorithms for generalized Bradley-Terry models. In *The Annals of Statistics*, volume 32, pp. 384–406, 2004.
- Liu, Tie-Yan. *Learning to Rank for Information Retrieval*. Springer, 2011.
- Luce, Robert Duncan. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, 1959.
- Negahban, Sahand, Oh, Sewoong, and Shah, Devavrat. Iterative ranking from pair-wise comparisons. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 2483–2491, Lake Tahoe, NV, USA, 2012.
- Plackett, Robin L. The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):193–202, 1975.
- Vaart, A. W. van der. *Asymptotic Statistics*. Cambridge: Cambridge University Press, 1998.