MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **66** (2011) 219-229

ISSN 0340 - 6253

Approximation of Average Ranks in Posets

K. De Loof[a], B. De Baets[a] and H. De Meyer[b]

[a] *Department of Applied Mathematics, Biometrics and Process Control, Ghent University Coupure links 653, B-9000 Gent, Belgium*, karel.deloof@ugent.be
[b] *Department of Applied Mathematics and Computer Science, Ghent University Krijgslaan 281 S9, B-9000 Gent, Belgium*

(Received October 21, 2010)

Abstract Objects that are described by attribute vectors often need to be ranked. A popular approach not requiring subjective assumptions ranks the objects on the basis of their average rank in the linear extensions of the induced partially ordered set, or poset for short. Since the exact computation of average ranks in posets with many incomparable objects is infeasible with current technology, approximations are required. In this paper we introduce a new formula that approximates the average ranks more accurately than presently known formulae.

1 Introduction

In many contexts objects such as chemicals need to be ranked on the basis of objective criteria. The emergence of initiatives aimed at protecting the environment, such as the REACH (Registration, Authorisation and Restriction of Chemicals) project of the European Union [1], have recently increased the need for such ranking methods. In computer models, objects such as chemicals are typically represented as attribute vectors. Two objects are comparable when the attribute vector of the first object is componentwise smaller than or equal to, or larger than or equal to the attribute vector of the second object, but remain incomparable when there is no such relationship. The fact that different objects can have the same attribute vector causes equivalence classes to arise. Since there is no way to discern objects residing in the same equivalence class, one will often opt to retain only one representative object in each class. The set that consists of these representative objects is a partially ordered set, or poset for short.

Objects that are incomparable are considered as obstacles in obtaining a ranking. A popular approach that is not based on subjective assumptions and that overcomes this problem, computes the average rank of each object in the linear extensions of the poset. A linear extension of a poset consists of the set of objects equipped with a linear order that is compatible with the partial order of the poset. Although algorithms to compute the average ranks are known [11, 12], due to their exponential nature they are not suitable for posets with many incomparable objects. As a consequence, one often has to resort to approximative approaches. Markov chain Monte Carlo methods allow to sample uniformly at random from the set of linear extensions of the poset [9]. However, considerable time is required to generate a random linear extension and a large sample of linear extensions is needed to allow for the estimation of the average ranks [15]. Albeit at the expense of accuracy, in many applications fast approximations are needed. Approximative formulae using simple features of the poset to approximate the average ranks have already been developed [6, 7] and are used in practice. In this contribution, we introduce a new formula based on approximations of the so-called mutual rank probabilities. For larger posets, it turns out to perform considerably better than the formulae presently used.

2 Preliminaries

A binary relation $\leq_P$ on a set P is called an *order relation* if it is reflexive ($x \leq_P x$), antisymmetric ($x \leq_P y$ and $y \leq_P x$ imply $x =_P y$) and transitive ($x \leq_P y$ and $y \leq_P z$ imply $x \leq_P z$). If for an order relation it furthermore holds that every two elements are comparable ($x \leq_P y$ or $y \leq_P x$), it is called a *linear order relation*. If neither $x \leq_P y$ nor $x \geq_P y$, we say that x and y are *incomparable* and write $x \,||_P\, y$. A couple $(P, \leq_P)$, where P is a set of objects and $\leq_P$ is an order relation on P, is called a partially ordered set or *poset* for short. The size of a poset $(P, \leq_P)$, denoted as $|P|$, is defined as the number of elements in P.

We will assume, as described in the introduction, that each object $x \in P$ can be described by an attribute vector $q(x) = (q_1(x), q_2(x), \ldots, q_k(x))$, where $q_i(x) \in Q_i$ for each $i \in \{1, \ldots, k\}$. Each set Q_i is equipped with a linear order relation $\leq_i$. This reflects the fact that q_i can be considered as a true criterion: if $q_i(x) \leq_i q_i(y)$, then x is at most as good as y with respect to criterion q_i. We say that x is smaller than or equal to y,

denoted as $x \leq_P y$ if $q_i(x) \leq_i q_i(y)$ for all $i \in \{1, \ldots, k\}$. Without loss of generality, we assume that all attribute vectors are unique. If this is not the case, we choose an arbitrary representative element from each equivalence class of objects having identical attribute vectors. The relation $\leq_P$ is an order relation; it is the restriction to P of the product ordering on $Q_1 \times \ldots \times Q_k$.

Let Q be an ordinary set and R and S be two binary relations on Q. If $R \subset S$, then (Q, S) is called an extension of (Q, R). A *linear extension* of a poset $(P, \leq_P)$ is an extension $(P, \leq_L)$ for which $\leq_L$ is a linear order. Let us denote the set of linear extensions of $(P, \leq_P)$ as $\mathcal{E}(P)$. The *rank probability* $\mathcal{P}(\text{rank}(x) = i)$ of an element $x \in P$ is defined as the fraction of linear extensions in which element $x \in P$ has rank i, or in other words have exactly $i - 1$ elements that are smaller than x. The *average rank* $\rho(x)$ of an element $x \in P$ is then defined as the expected value of the rank of x, *i.e.*

$$\rho(x) = \sum_{i=1}^{|P|} i \cdot \mathcal{P}(\text{rank}(x) = i)\,.$$

Finally, the *mutual rank probability* $\mathcal{P}(x > y)$ of two elements $x, y \in P$ is defined as the fraction of linear extensions in which element x is ranked higher than element y.

3 Approximating the average ranks

3.1 Known formulae

Consider a poset $(P, \leq_P)$. The original local partial order model (LPOM) developed by Brüggemann *et al.* [6, 7] obtains a simple approximation of the average rank of an element $x \in P$ by considerably simplifying the structure of the poset. Let us denote the set of elements incomparable to x as $I(x)$ and its cardinality as $i(x)$, the set of elements smaller than x as $S(x)$ and its cardinality as $s(x)$, and the set of elements larger than x as $L(x)$ and its cardinality as $l(x)$. The approximation considers all elements from $I(x)$ as isolated elements, *i.e.* as elements that are incomparable to all other elements of P, and the sets $S(x)$ and $L(x)$ as linearly ordered. Furthermore, the elements in $I(x)$ are considered to be either all ranked before x or all ranked after x in linear extensions of $(P, \leq_P)$. When denoting the size of $(P, \leq_P)$ as n, the average rank is of $x \in P$ is approximated as

$$\hat{\rho}_L(x) = \frac{[s(x) + 1]\,[n + 1]}{n + 1 - i(x)}\,. \tag{1}$$

Very recently, Brüggemann *et al.* [3] have introduced an extended local partial order model approximating the average rank of $x \in P$ as

$$\hat{\rho}_E(x) = s(x) + 1 + \sum_{y \in I(x)} \frac{|S(x) \cap I(y)| + 1}{|I(y) \setminus I(x)| + 1}. \tag{2}$$

Note that in the specific case where the elements in $I(x)$ are isolated elements, expression (2) can be rewritten as

$$s(x) + 1 + \frac{i(x)\,[s(x)+1]}{n - i(x) + 1},$$

which simplifies to the approximation in (1). The extended local partial order model can thus be seen as a generalization of the local partial order model.

Note that, given a poset $(P, \leq_P)$, the approximations in (1) and (2) both have a time complexity of $\mathcal{O}(n^2)$.

3.2 New formula

In this section we introduce a new formula that approximates the average rank of $x \in P$. Let us first put forward an interesting relationship between the average ranks and mutual rank probabilities in a poset.

Theorem 1 *For a poset $(P, \leq_P)$ where $P = \{p_1, p_2, \ldots, p_n\}$ and $p_l \in P$, the following relationship holds between the average ranks and the mutual rank probabilities:*

$$\rho(p_l) = \sum_{i=1}^{n} i \cdot \mathcal{P}(\mathrm{rank}(p_l) = i) = 1 + \sum_{j=1}^{n} \mathcal{P}(p_l > p_j).$$

Proof We will prove a slightly more general identity. Let A be any list of m permutations of n different symbols $p_1, p_2, \ldots, p_n$. Denote by $n_i^{p_j}$ the number of times symbol p_j occurs at position i in A, and by $n_{p_j > p_i}$ the number of times symbol p_j occurs after symbol p_i in A.

We want to prove that for any symbol $p_l \in \{p_1, p_2, \ldots, p_n\}$ it holds that

$$\sum_{i=1}^{n} i \cdot n_i^{p_l} = m + \sum_{j=1}^{n} n_{p_l > p_j}. \tag{3}$$

Dividing both sides by m, the left-hand side describes the average position of symbol p_l in A, whereas the right-hand side represents the sum of the elements in column l of the

matrix B with elements $b_{ij} = \mathcal{P}(p_j > p_i|A)$ for all $i, j \in \{1, 2, \ldots, k\}$ where $i \neq j$, and with $b_{ii} = 1$ for all $i \in \{1, 2, \ldots, n\}$.

The proof goes by induction. First let $n = 2$ and denote the 2 symbols as p_1 and p_2. Suppose A contains m_1 permutations (p_1, p_2) and $m_2 = m - m_1$ permutations (p_2, p_1). We have $n_1^{p_1} = n_2^{p_2} = m_1$, $n_2^{p_1} = n_1^{p_2} = m_2$, $n_{p_2>p_1} = m_1$, $n_{p_1>p_2} = m_2$ and $m_1 + m_2 = m$. It then holds that

$$\sum_{i=1}^{2} i \cdot n_i^{p_1} = n_1^{p_1} + 2 \cdot n_2^{p_1} = m_1 + 2 \cdot m_2 = m + m_2$$

$$\sum_{i=1}^{2} i \cdot n_i^{p_2} = n_1^{p_2} + 2 \cdot n_2^{p_2} = m_2 + 2 \cdot m_1 = m + m_1$$

yielding

$$\sum_{i=1}^{2} i \cdot n_i^{p_1} = m + n_{p_1>p_2}$$

and

$$\sum_{i=1}^{2} i \cdot n_i^{p_2} = m + n_{p_2>p_1}.$$

Therefore identity (3) is satisfied for $n = 2$.

Now suppose (3) is satisfied for some $n \geq 2$. We will now prove that this is also the case for $n+1$. Let A be the given set of m permutations of $n+1$ symbols. Denote any of these symbols as x and the remaining symbols as $p_1, p_2, \ldots, p_n$. With each permutation in A, define a new permutation of n symbols $p_1, p_2, \ldots, p_n$ by leaving out symbol x. Denote by A' the list of m permutations obtained in this way. Let us take any symbol $p_l \in \{p_1, p_2, \ldots, p_n\}$. Denote the number of times p_l occurs at position i in A' as $n'^{p_l}_i$ and the number of times p_l occurs after p_j in A' as $n'_{p_l>p_j}$. In A', identity (3) holds:

$$\sum_{i=1}^{n} i \cdot n'^{p_l}_i = m + \sum_{j=1}^{n} n'_{p_l>p_j}\,.$$

We add to the permutations in A' the symbol x such as to retrieve the permutations in A. Suppose that p_l is at position i in a permutation in A', and p_l comes after x in the corresponding permutation in A, then we have one permutation less in A with p_l at position i and one permutation more with p_l at position $i+1$. Hence, it follows that

$$\sum_{i=1}^{n+1} i \cdot n_i^{p_l} = \sum_{i=1}^{n} i \cdot n_i'^{p_l} + n_{p_l>x}\,.$$

Moreover, $n'_{p_l>p_j} = n_{p_l>p_j}$ for all $p_j \neq p_l$ and $p_j \neq x$. Hence,

$$\sum_{i=1}^{n+1} i \cdot n_i^{p_l} = \sum_{i=1}^{n} i \cdot n_i'^{p_l} + n_{p_l>x} = m + \sum_{j=1}^{n} n_{p_l>p_j} + n_{p_l>x}\,.$$

Since x and p_l are any two symbols from the $n+1$ symbols, expression (3) is valid for $n+1$. □

As Theorem 1 points out, computing the average rank of an element $x \in P$ it is equivalent to computing 1 plus the sum of the mutual rank probabilities $\mathcal{P}(x > y)$ for all $y \in P$. On the basis of this relationship the average rank of x can be written as

$$\rho(x) = 1 + \sum_{y \in P} \mathcal{P}(x > y) = s(x) + 1 + \sum_{y \in I(x)} \mathcal{P}(x > y)\,. \tag{4}$$

Using an approximation for $\mathcal{P}(x > y)$, with $x \neq y$, suggested by Brüggemann *et al.* [4], namely

$$\hat{\mathcal{P}}(x > y) = \frac{[s(x)+1]\,[l(y)+1]}{[s(x)+1]\,[l(y)+1] + [l(x)+1]\,[s(y)+1]}\,, \tag{5}$$

one can therefore approximate the average rank of $x \in P$ as

$$\hat{\rho}_1(x) = s(x) + 1 + \sum_{y \in I(x)} \frac{[s(x)+1]\,[l(y)+1]}{[s(x)+1]\,[l(y)+1] + [l(x)+1]\,[s(y)+1]}\,. \tag{6}$$

Although $\hat{\rho}_1(x)$ turns out to approximate $\rho(x)$ better than the known formulae (see Section 4), only elements comparable to x or y are taken into account in the approximation of the mutual rank probabilities. One could expect to obtain a better approximation when incomparable elements would also be considered. The principal idea behind our enhanced formula is therefore to substitute $s(x)$ and $l(x)$ by quantities that incorporate incomparable elements. In order to account for these elements we will use the approximation of the mutual rank probabilities in (5). Hence, let us define for $x \in P$

$$\begin{aligned} \tilde{s}(x) &= s(x) + \sum_{z \in I(x)} \hat{\mathcal{P}}(x > z) \\ \tilde{l}(x) &= l(x) + \sum_{z \in I(x)} \hat{\mathcal{P}}(z > x)\,, \end{aligned}$$

and introduce the approximation

$$\hat{\rho}_2(x) = s(x) + 1 + \sum_{y \in I(x)} \frac{[\tilde{s}(x)+1][\tilde{l}(y)+1]}{[\tilde{s}(x)+1][\tilde{l}(y)+1] + [\tilde{l}(x)+1][\tilde{s}(y)+1]} \,. \tag{7}$$

Instead of counting the number of elements smaller than x, as $s(x)$ does, $\tilde{s}(x)$ approximates the total probability of elements to appear before x in a linear extension by using the approximated mutual rank probabilities in (5).

When we denote the size of the poset as n, this approximation has a time complexity of $\mathcal{O}(n^2)$, which is precisely the complexity of the known formulae.

4 Results

In order to compare the accuracy of the approximation formulae $\hat{\rho}_1$ and $\hat{\rho}_2$ with that of the known formulae based on the (extended) local partial order model, the following experiment is carried out. All posets of size $n \in \{4, \ldots, 10\}$ are enumerated using the algorithm of Brinkmann and McKay [2]. For each poset, all average ranks are computed using the exact algorithm developed by the present authors [12] and compared with the results obtained by the approximation formulae.

As an indication of the accuracy, the mean absolute errors of the approximate average ranks, averaged over all posets of size $n \in \{4, \ldots, 10\}$, are shown in Table 1. Additionally, in Table 2 the maximal absolute difference between the exact and approximate average ranks, averaged over all posets of size $n \in \{4, \ldots, 10\}$, is shown. Although for small n the formula $\hat{\rho}_E$ based on the extended local partial order model has the smallest mean absolute error, for $n \geq 9$ the mean and maximal absolute error of $\hat{\rho}_2$ become smaller. Although the exhaustive experiments only cover posets on up to 10 elements, a clear trend is visible in Table 1.

In Figure 1 a box plot is shown with the mean absolute errors of the approximate average ranks for all posets of size 10. Although the difference in accuracy between the new formulae and the formula $\hat{\rho}_E$ is still limited for this poset size, the plot clearly indicates there are less posets with a high mean absolute error with formula $\hat{\rho}_2$ in comparison with formula $\hat{\rho}_E$.

Table 1: The mean absolute errors of the approximate average ranks, averaged over all posets of size $n \in \{4, \ldots, 10\}$.

n	$\hat{\rho}_L$	$\hat{\rho}_E$	$\hat{\rho}_1$	$\hat{\rho}_2$
4	0.1026	**0.0057**	0.0115	0.0107
5	0.1736	**0.0195**	0.0308	0.0267
6	0.2441	**0.0393**	0.0536	0.0473
7	0.3170	**0.0647**	0.0780	0.0705
8	0.3917	**0.0944**	0.1029	0.0958
9	0.4690	0.1281	0.1281	**0.1226**
10	0.5497	0.1658	0.1533	**0.1505**

Table 2: The maximal absolute errors of the approximate average ranks, averaged over all posets of size $n \in \{4, \ldots, 10\}$.

n	$\hat{\rho}_L$	$\hat{\rho}_E$	$\hat{\rho}_1$	$\hat{\rho}_2$
4	0.1656	**0.0094**	0.0208	0.0182
5	0.3151	**0.0391**	0.0646	0.0514
6	0.4650	**0.0853**	0.1204	0.0992
7	0.6310	**0.1462**	0.1846	0.1545
8	0.7973	0.2158	0.2503	**0.2147**
9	0.9694	0.2927	0.3178	**0.2780**
10	1.1458	0.3755	0.3886	**0.3433**

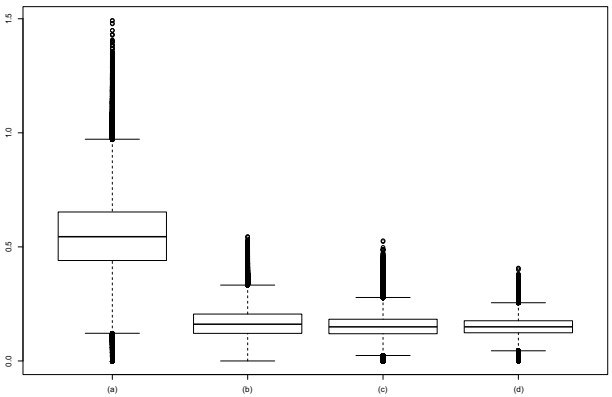

Figure 1: A box plot with the mean absolute errors of the approximate average ranks (a) $\hat{\rho}_L$, (b) $\hat{\rho}_E$, (c) $\hat{\rho}_1$ and (d) $\hat{\rho}_2$.

Table 3: The mean absolute errors of the approximate average ranks, averaged over 100 posets of size n with $n \in \{10, \ldots, 20\}$.

n	$\hat{\rho}_L$	$\hat{\rho}_E$	$\hat{\rho}_1$	$\hat{\rho}_2$
10	0.3397	0.1410	0.1674	**0.1302**
11	0.4043	0.1752	0.1970	**0.1537**
12	0.4750	0.2135	0.2259	**0.1825**
13	0.5600	0.2681	0.2775	**0.2294**
14	0.7437	0.3553	0.3152	**0.2567**
15	0.7851	0.3867	0.3725	**0.3019**
16	0.8787	0.4507	0.3587	**0.2954**
17	0.9629	0.4841	0.4474	**0.3673**
18	1.0576	0.5630	0.4582	**0.3688**
19	1.1706	0.6363	0.4995	**0.4090**
20	1.2300	0.6930	0.5287	**0.4221**

Furthermore, we generated 100 posets of size $n \in \{10, \ldots, 20\}$ by drawing attribute vectors uniformly at random from $\{1, \ldots, 20\}^4$ (see e.g. [10]). Although this procedure will not generate each poset with equal probability, one can expect the generated posets to be more representative for the posets induced by data sets encountered in practice. For each poset, the exact average ranks are computed and compared with the approximations. As Table 3 shows, formula $\hat{\rho}_2$ performs slightly better for $n = 10$ compared to the first experiment where all posets of size 10 are generated. Moreover, for larger n the difference in accuracy between $\hat{\rho}_2$ and $\hat{\rho}_E$ increases, in line with the trend visible in Table 1.

Finally, we consider six real-world data sets from literature (see Table 4) for which the average ranks of the poset can be computed exactly. From Table 5 it is clear that the best approximations are again obtained by formula $\hat{\rho}_2$, except for data set [5] where slightly better results are obtained by formula $\hat{\rho}_E$, and two data sets where formula $\hat{\rho}_1$ obtains the best results.

5 Conclusion

We established a new formula to approximate the average ranks of the elements of a poset based on an interesting relationship between the average ranks and the mutual rank probabilities. We verified the accuracy of the formula by carrying out exhaustive experiments on posets of size up to 10, by sampling posets of size up to 20 and by considering six real-

Table 4: The number and type of the objects, the number of criteria and the number of incomparable pairs, both absolute and relative to a poset of given size with only incomparable elements, for six real-world data sets from literature.

data set	objects	# crit.	# inc.	rel. inc.
[14]	12 high production volume chemicals	4	43	65,1%
[8]	15 online databases	5	68	64,8%
[13]	17 pesticides	4	116	85,3%
[5]	18 fish tests	6	98	64,1%
[16] case c	31 types of fruit	3	288	61,9%
[11]	33 regions in Baden-Württemberg	4	379	71,8%

Table 5: The mean absolute errors of the approximate average ranks of the six real-world data sets in Table 4.

data set	$\hat{\rho}_L$	$\hat{\rho}_E$	$\hat{\rho}_1$	$\hat{\rho}_2$
[14]	0.5266	0.2165	0.2497	**0.2128**
[8]	1.1386	0.3962	0.3507	**0.2757**
[13]	1.1182	0.6168	0.4118	**0.2194**
[5]	0.7988	**0.4100**	0.4830	0.5485
[16] case c	2.5389	1.4745	**0.6517**	0.9072
[11]	2.6908	1.3416	**1.0034**	1.2334

world data sets for which the approximations can be compared with the exact average ranks. For posets of size $n \geq 9$ our formula turns out to perform consistently better than presently known formulae without requiring additional computation time.

References

[1] European Commission. REACH in brief, `http://ecb.jrc.it/DOCUMENTS/REACH/REACH_in_brief_council_comm_pos_060905.pdf`, September 2006.

[2] G. Brinkmann, B. McKay, Posets on up to 16 points, *Order* **19** (2002) 147–179.

[3] R. Brüggemann, L. Carlsen, An improved estimation of averaged ranks of partial orders, *MATCH Commun. Math. Comput. Chem.* **65** (2011) 383–414.

[4] R. Brüggemann, D. Lerche, P. Sørensen, *First attempts to relate structures of Hasse diagrams with mutual probabilities*, Technical Report 479, The 5th workshop held at the National Environmental Research Institute (NERI), Roskilde, Denmark, 2003.

[5] R. Brüggemann, J. Schwaiger, R. Negele, Applying Hasse diagram technique for the evaluation of toxicological fish tests, *Chemosphere* **30** (1995) 1767–1780.

[6] R. Brüggemann, U. Simon, S. Mey, Estimation of averaged ranks by extended local partial order models, *MATCH Commun. Math. Comput. Chem.* **54** (2005) 489–518.

[7] R. Brüggemann, P. Sørensen, D. Lerche, L. Carlsen, Estimation of averaged ranks by a local partial order model, *J. Chem. Inf. Comp. Sci.* **44** (2004) 618–625.

[8] R. Brüggemann, K. Voigt, An evaluation of online databases by methods of lattice theory, *Chemosphere* **31** (1995) 3585–3594.

[9] R. Bubley, M. Dyer, Faster random generation of linear extensions, *Discr. Math.* **201** (1999) 81–88.

[10] K. De Loof, B. De Baets, H. De Meyer, On the random generation of monotone data sets, *Inform. Process. Lett.* **107** (2008) 216–220.

[11] K. De Loof, B. De Baets, H. De Meyer, R. Brüggemann, A hitchhiker's guide to poset ranking, *Comb. Chem. High T. Scr.* **11** (2008) 734–744.

[12] K. De Loof, H. De Meyer, B. De Baets, Exploiting the lattice of ideals representation of a poset, *Fundam. Inform.* **71** (2006) 309–321.

[13] K. De Loof, M. Rademaker, R. Bruggemann, H. De Meyer, G. Restrepo, B. De Baets, New tools in partial order theory for risk assessment of pesticides, *J. Chem. Inf. Model.*, submitted.

[14] D. Lerche, R. Brüggemann, P. Sørensen, L. Carlsen, O. Nielsen, A comparison of partial order technique with three methods of multi-criteria analysis for ranking of chemical substances, *J. Chem. Inf. Comp. Sci.* **42** (2002) 1086–1098.

[15] D. Lerche, P. Sørensen, Evaluation of the ranking probabilities for partial orders based on random linear extensions, *Chemosphere* **53** (2003) 981–992.

[16] M. Pavan, R. Todeschini, New indices for analysing partial ranking diagrams, *Anal. Chim. Acta* **515** (2004) 167–181.