2022 12th International Congress on Advanced Applied Informatics (IIAI-AAI) | 978-1-6654-9755-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/IIAIAAI55812.2022.00076

# A computationally efficient approach for solving RBSC-based formulation of the subset selection problem

1st Kohei Furuya
*Department of Computer Science*
*Okayama University*
Okayama, Japan
p95q8dp3@s.okayama-u.ac.jp

2nd Zeynep Yücel
*Department of Computer Science*
*Okayama University*
Okayama, Japan
zeynep@okayama-u.ac.jp

3th Parisa Supitayakul
*Department of Computer Science*
*Okayama University*
Okayama, Japan
pgw45ydd@s.okayama-u.ac.jp

4th Akito Monden
*Department of Computer Science*
*Okayama University*
Okayama, Japan
monden@okayama-u.ac.jp

***Abstract*—This study focuses on a specific type of subset selection problem, which is constrained in terms of the rank bi-serial correlation (RBSC) coefficient of the outputs. For solving such problems, we propose an approach with several advantages such as (i) providing a clear insight into the feasibility of the problem with respect to the hyper-parameters, (ii) being non-iterative, (iii) having a foreseeable running time, and (iv) with the potential to yield non-deterministic (diverse) outputs. In particular, the proposed approach is based on starting from a composition of subsets with an extreme value of the RBSC coefficient (e.g. $\rho = 1$) and swapping certain elements of the subsets in order to adjust $\rho$ into the desired range. The proposed method is superior to the previously proposed `RBSC-SubGen`, which attempts to solve the problem before confirming its feasibility, taking random steps, and has unforeseeable running times and saturation issues.**

***Index Terms*—ranking-and-selection problem, rank bi-serial correlation, subset selection**

## I. Introduction and motivation

There is a certain degree of randomness in many real-life processes, e.g. meteorological events, malfunctioning of machinery in industrial plants etc. In order to make predictions on future states of those systems or develop a management strategy, it is common to model them with stochastic models treating the observations as random variables. However, in certain cases, data collection and model building may be difficult, e.g. when there is too much noise on the observations or there is not enough computational power to reflect all factors in a stochastic simulation model.

In such cases, choosing a subset, is a convenient approach. The research problem of *Ranking-and-selection* (R&S) of computational statistics focuses on such processes with random factors [1]. Specifically, R&S problem addresses the choosing of the best (e.g. most representative) of two (or more) processes (or items) which bear a certain degree of randomness, according to a given performance measure.

In solving the R&S problems, due to the stochasticity of the system in question, numerous real-world experiments need to be carried out. However, in certain cases (e.g. nuclear power plant accidents), it may not possible to make experiments. In addition, in certain other cases (e.g. pharmacological studies), physical experiments might cost significant time and budget.

In that respect, computer simulations are considered to be a suitable tool to expand the data set [2]. However, although computer simulations are easier to perform than physical experiments, since the stochastic models corresponding to the system in focus tend to be highly complex, they may take a long time and require high computational resources.

In that respect, R&S algorithms search for solutions in an efficient way, namely, with a decent number of simulations. To that end, certain R&S algorithms choose the (single) best process out of two or more processes according to a certain performance measure [1]. On the other hand, other R&S algorithms search for a subset of processes rather than a single one, which are termed as *subset selection* problems [3].

In this article, we focus on the subset selection problem conditioned on a rank bi-serial correlation (RBSC) relation between the subsets. We propose an alternative method to our previous work [4], [5], which remedies its several shortcomings. Namely, in our previous study [4], we proposed the algorithm `RBSC-SubGen` for solving the subset generation problem. This method is shown to yield outcomes with a diverse composition, but suffer from high computational load and saturation [5]. Therefore, in this study we try to assess the feasibility of the problem before attempting to solve it to save any pointless efforts. Once the existence of a feasible solution is confirmed, we generate the subsets without any iterations, which enable anticipation of running time. In this manner, we develop a remedy for the most serious issues of `RBSC-SubGen`. In addition, due to non-deterministic nature of the proposed method, the diversity in the outcomes is expected to be sustained.

978-1-6654-9755-8/22/$31.00 ©2022 IEEE
DOI 10.1109/IIAIAAI55812.2022.00076

## II. Related work

In literature, R&S problems are often deployed in agricultural and clinical sciences (e.g. grain yields, drug treatments) [6]. Neverthless, recently the application area R&S problems has expanded to involve various other feilds such as judicial system [7], vehicular safety [8], composition of photo albums [9] etc.

Although a large part of the R&S algorithms target choosing the (single) best process [1], [10], there are also certain others, which search for a subset of processes [11], [12], [3].

To this day, numerous R&S algorithms and subset selection approaches have been proposed. A few well-know schemes include indifference-zone approach (IZ), maximization of the probability of correct selection (PCS), optimal computing budget allocation (OCBA) and expected value of information (EVI) [13], [14], [10]. As R&S requires pairwise comparison of performance metrics [15], there is a group of studies, which aim building lightweight solutions by focusing on this aspect. Namely, they make an attempt to optimize R&S by decreasing the number of pairwise comparisons by eliminating unprofitable ones [16].

On the other hand, subset selection is an important tool in stochastic simulation. For instance, it is a common approach to test several competing models and choose the best performing model to base future decisions on. However, since such stochastic simulations concerning systems with a high degree of freedom are quite time consuming and demand significant computational resources, it is plausible approach to first eliminate non-competitive designs and then study the remaining ones in detail [17]. Subset selection is a convenient tool for such sort of approach.

In addition to eliminating non-competitive designs, subset selection problem may be formulated in various other settings for diverse range of purposes. For instance, if the observatiosn collecte from the systemin focus are noisy, it is not certain that the single best performing process is actually the true best solution. In that resepct, in order to avoid the effect of noise, it is better to choose a subset of models and then study further in more detail [18].

Post-hoc analysis is another potential application domain for subset selection problem. For example, if the system in focus can be broken down into two (or moe) sub-systems, a follow-up simulation of the subsequent process can be tested in conjunction with a subset of solutions (of the preceding process). This serves for achieving more accurate and comprehensive results.

The subsets are often required to attain the maximum (or mimnimum) score in accordance to a pre-determined performance metric. In addition, they are required to be as homogenous as possible (to avoid outliers) [17]. Moreover, in relation to the specificities of the problem, various other conditions can be imposed, such as uniformity and diversity [4] or inclusiveness and fairness [19].

## III. Background on RBSC-SubGen

`RBSC-SubGen` is originally designed for collating a desired number of vocabulary decks (out of a large corpus) with a desired level of word frequency relation. In that respect, the problem in the focus of `RBSC-SubGen` shares many common aspects with the generic subset selection problem, where each item is associated with a *score*. Therefore, `RBSC-SubGen` can potentially be applied on other scenarios relating to a large variety of subset selection problems.

In what follows, avoiding any specifics on the items in the data set and how they are ranked, we will give a brief outline of the subset selection problem, overview the previously proposed `RBSC-SubGen` and point out to its advantages and disadvantages.

### A. Subset selection problem defined in terms of ranking relations

Suppose a data set with a sufficient amount of items (henceforth, referred to as the universal set), is provided and that each item is associated with a (quantitative) score. Let $U$ denote the universal set and lets its size be $L$.

Consider that two subsets are to be selected out of the universal set with a *desired ranking relation* between their elements. Suppose that the subsets are denoted with $A$, $B$ and that they have a size of $\ell$. Let us denote the scores of the items in $A$ and $B$ with $a_i$ and $b_i$, respectively, where $1 \leq i, j, \leq \ell$.

In particular, according to `RBSC-SubGen`, the ranking relation between $A$ and $B$ is defined in terms of the RBSC coefficient, denoted with $\rho$ [20].

The RBSC coefficient $\rho$ measures the correlation between a dichotomous variable and a ranking variable (see [5] for an example problem). In this case, the two subsets represent the dichotomy and the scores of their items define the ranking.

Without loss of generality, let $A$ represent the subset with relatively lower scores and $B$ represent the subset with relatively higher scores. Let $S$ stand for the number of evidence *supporting* this relation, (i.e. $a_i < b_j \ \forall 1 \leq i, j \leq \ell$).

$$S = 0.5 \cdot \sum_{\forall i,j} \left(\operatorname{sign}\left(b_j - a_i\right) + 1\right), \tag{1}$$

where $\operatorname{sign}(\cdot)$ denotes signum. Similarly, let $C$ represent the number of evidence *contradicting* with this relation, (i.e. $a_i > b_j \ \forall 1 \leq i, j \leq \ell$). The number of contradicting evidence can be obtained by replacing $\operatorname{sign}\left(b_j - a_i\right)$ with $\operatorname{sign}\left(a_i - b_j\right)$ in Equation 1.

Given $S$ and $C$, we can quantify how much the subsets $A$ and $B$ agree with the indicated relation, in terms of the RBSC coefficient as follows,

$$\rho = \frac{S - C}{S + C}.$$

Note that $\rho \in [-1, 1]$. If all pairwise relations agree with the indicated condition, then $\rho$ is exactly 1. On the contrary, they all disagree with the indicated condition, $\rho$ is -1, whereas a correlation of 0 indicates equal amount of agreeing and disagreeing evidence.

### B. Overview of RBSC-SubGen

RBSC-Subgen firstly builds a pair of initial (random) subsets out of $U$. If the RBSC coefficient $\rho$ associated with these subsets is within $\rho^* \pm \epsilon$, the algorithm returns the subsets and terminates. However, achieving such subsets is extremely rare. Often the RBSC coefficient $\rho$ is not within $\rho^* \pm \epsilon$. In that case, the kind of necessary update on each set is determined. In particular, if $\rho$ is too low, then the scores of the items in $A$ need to be decreased and those of $B$ need to be increased. On the other hand, if $\rho$ is too high, the scores of the items in $A$ need to be increased and those of $B$ need to be decreased.

This information is provided to the update routine as an input, together with the current subsets $A$, $B$ and the set of available items (the items of $U$ which are not yet disposed). The update is realized by inserting/removing a single arbitrarily chosen element to/from each subset. Since the item to be inserted or removed is chosen in an arbitrary way, it is not clear how much impact it will do on $\rho$. Thus, the new value of RBSC coefficient $\rho$ needs to computed by checking all evidence once again and inserting the terms in Equation 4.

The three-step procedure composed of (i) computation of $\rho$, (ii) determining the sort of update and (iii) realizing the update, is repeated until $\rho \in [\rho^* - \epsilon, \rho^* + \epsilon]$.

### C. Advantages of RBSC-SubGen

According to [4], RBSC-SubGen yields outcomes, which are uniform (within the subsets) and diverse (across the subsets). Namely, items in the same subset are similar enough in terms of their ranks (uniform), such that they can be grouped together, and there is enough difference between different subsets such that they can be considered distinct.

### D. Disadvantages of RBSC-SubGen

As mentioned above, the main routine of RBSC-Subgen is the random iterative update. This part of the algorithm is quite brute-force and, thus, amount to a high computational cost (i.e. increasing the running time and using large memory space). An upper bound for computational cost cannot be determined[1], but it is still possible to comment on a lower bound. Namely, since the computation of RBSC coefficient $\rho$ requires checking all pairwise relations, computational cost cannot be lower than $\mathcal{O}(\ell^2)$, which considered to be heavy for running time.

In addition, if we attempt to keep the computational cost under control by imposing an upper limit in the number of iterations, the number of saturated cases increase.

In order to overcome such issues, in Section V, we will introduce a new approach. However, before that, in Section IV we will make some remarks on the feasibility of the problem and introduce several relations and terms, which help to present the proposed approach in a simple and clear manner.

## IV. Restructuring the problem

### A. Further remarks on fesibility of the problem

Assume that $A$ is required to have lower ranks than $B$, where the ranking relation is governed by the RBSC coefficient $\rho$ and the desired value of RBSC coefficient is given as $\rho^*$ with tolerance of $\epsilon$,

$$\rho \in [\rho^* - \epsilon, \rho^* + \epsilon]. \tag{2}$$

There is obviously no solution to the problem, if there are not enough elements in the universal set $U$ to build the two subsets

$$L < 2\ell. \tag{3}$$

Note that if $L = 2\ell$, it does not necessarily mean that the problem is infeasible, but it certainly is much stricter.

In addition, there is the following relation between $\ell$ and $\epsilon$, which helps in assessing the amount of flexibility of the solution. In computing RBSC coefficient relating to sets $A$ and $B$, we consider the ranking relation between every pair of elements of $A$ and $B$, i.e. $a_i$ and $b_j$ for $\forall i, j \in [1, \ell]$. Thus, the number of possible pairs is $\ell^2$. In that respect, the RBSC coefficient $\rho$ may change at discrete intervals of $1/\ell^2$. If $\epsilon$ is less than $1/\ell^2$, although this does not mean that there is no solution, since the interval given in Equation 2 is reduced to contain a single value, the solution is less flexible, i.e. the exact value of $\rho^*$ needs to achieved.

### B. A preface on the relation between RBSC coefficient and number of supporting evidence

We first point out to some trivial but useful relations between the RBSC coefficient $\rho$ and the number of supporting evidence $S$.

Remember that the RBSC coefficient relating to sets $A$ and $B$, where $A$ is hypothesized to have lower ranks that $B$, is computed based on the simple difference formula of Kerby [20]. Let $S$ stand for the number of evidence supporting the hypothesis (i.e. $a_i$ is smaller than $b_j$). Similarly, assume that $C$ represents the number of evidence contradicting the hypothesis (i.e. $a_i$ is larger than $b_j$). Then, according to the simple difference formula, $\rho$ is computed as

$$\rho = \frac{S - C}{S + C}. \tag{4}$$

As mentioned in Section IV-A, the number of all possible pairs is $\ell^2$. Thus, the total number of evidence (supporting or contradicting the hypothesis) is $\ell^2$,

$$S + C = \ell^2. \tag{5}$$

So, the number of evidence contradicting the hypothesis can be written in terms of $\ell$ and $S$ as $C = \ell^2 - S$. Thus, $\rho$ given in Equation 4 can be rewritten as below

$$\begin{aligned} \rho &= \frac{S - (\ell^2 - S)}{\ell^2}, \\ &= \frac{2S - \ell^2}{\ell^2}. \end{aligned} \tag{6}$$

Assuming that $\epsilon > 1/\ell^2$, any one of the multiple values of $\rho$ in the range $[\rho^* - \epsilon, \rho^* + \epsilon]$ will satisfy our requirements. Suppose that the maximum and minimum values of $\rho$, which

[1] unless there is a limit on number of iterations

satisfy the requirements are denoted with $\rho_{\max}$ and $\rho_{\min}$, respectively. Specifically,

$$\begin{aligned} \rho_{\max} &= \rho^* + \epsilon, \\ \rho_{\min} &= \rho^* - \epsilon. \end{aligned} \tag{7}$$

## V. Proposed approach

In this section, we propose an approach based on constructing two subsets with an extreme ranking relation (i.e. with $\rho = 1$) and then exchanging a set of elements such that the ranking relation is adjusted to the desired range.

We implement this approach with (i) a block-swapping strategy and (ii) an alternating-swapping strategy. As a matter of fact, the simple block-swapping strategy introduced in Section V-A is mainly aimed at explaining the gist of the proposed approach. In that respect, it is focused primarily on satisfying the condition on $\rho$, without actually caring for the distributions of the output subsets. As a second step, we make a simple modification on block-swapping and introduce the alternating-swapping strategy in Section V-E, which provides a remedy to the shortcomings of block-swapping to a certain degree.

### A. Block-swapping method

In explaining the block-swapping method, we assume $\rho^*$ is defined as positive and start from two subsets $A$ and $B$ with the extreme value of $\rho = 1$. In other words, we adjust the elements of $A$ and $B$ such that there is only supporting evidence and no contradicting evidence.

*1) Initializing the subsets:* The initial subsets $A$ and $B$ are chosen out of the universal set $U$ such that $a_i < b_j$ for $\forall i, j \in [1, \ell]$. To that end, we first pick $2\ell$ elements out of $U$ randomly and sort them. Let us denote this sorted subset of $U$ with $U_s$. We then assign the first half of $U_s$ to $A$ and the second half of $U_s$ to $B$.

Note that since we already sort $U_s$, also the elements of $A$ and $B$ are sorted. Also, since $a_i < b_j$ for every $1 \leq i, j \leq \ell$, all evidence supports the hypothesis. Thus, $S = \ell^2$ and $C = 0$, which gives $\rho = 1$. We will now modify $A$ and $B$ as explained in Section V-A2 to introduce some contradicting evidence into these sets.

*2) Modifying the subsets:* For introducing contradicting evidence, we propose swapping certain elements of $A$ and $B$. Suppose that $A$ and $B$ look as given in Equation 8 before swapping,

$$\begin{aligned} A &= [a_1 \dots a_\ell], \\ B &= [b_1 \dots b_\ell]. \end{aligned} \tag{8}$$

Lets say we take two blocks, namely take last $\alpha$ elements of $A$ and first $\alpha$ elements of $B$, and swap them. After swapping, we obtain new sets $A'$ and $B'$ as given in Equation 9,

$$\begin{aligned} A' &= [\overbrace{a_1 \dots a_{\ell-\alpha}}^{\ell-\alpha \text{ elements}} \overbrace{b_1 \dots b_\alpha}^{\alpha \text{ elements}}], \\ B' &= [\overbrace{a_{\ell-\alpha+1} \dots a_\ell}^{\alpha \text{ elements}} \overbrace{b_{\alpha+1} \dots b_\ell}^{\ell-\alpha \text{ elements}}]. \end{aligned} \tag{9}$$

This swapping introduces the necessary contradicting evidence. To be specific, the amount of contradicting evidence is equal to the number of pairs constructed from the second part of $A'$, i.e. $b_1 \dots b_\alpha$, and the first part of $B'$, $a_{\ell-\alpha+1} \dots a_\ell$.

Note that there are $\alpha$ elements in each of those pieces and, thus, the number of associated pairs is $\alpha^2$, i.e. $C = \alpha^2$. Since there are $\ell^2$ pairs in total and $\alpha^2$ of them are contradicting, the number of supporting evidence concerning $A'$ and $B'$ is

$$S' = \ell^2 - \alpha^2. \tag{10}$$

We can now replace $S'$ in Equation 6 to find the number of elements to be replaced corresponding each bounding permissible value $\rho_{\max}$ and $\rho_{\min}$.

Let us denote the number of elements to be replaced for obtaining $\rho_{\min}$ and $\rho_{\max}$ as $\alpha_{\max}$ and $\alpha_{\min}$, respectively. Note that $\rho_{\max}$ is closer to the initial value of $\rho = 1$, which suggests that a relatively less number of elements need to be swapped as compared to $\rho_{\min}$. Thus, $\alpha_{\min}$ corresponds to $\rho_{\max}$ and $\alpha_{\max}$ corresponds to $\rho_{\min}$.

Without loss of generality, let us start from $\rho_{\max}$. Replacing Equation 10 in Equation 6 and inserting $\rho_{\max}$ and $\alpha_{\min}$, we get

$$\rho_{\max} = \frac{2(\ell^2 - \alpha_{\min}^2) - \ell^2}{\ell^2}. \tag{11}$$

Arranging the terms, we get $\alpha_{\min}$ in terms of $\rho_{\max}$ as follows,

$$\alpha_{\min}^2 = \ell^2 \left( \frac{1 - \rho_{\max}}{2} \right). \tag{12}$$

Note that since $\alpha_{\min}$ is the number of elements to be swapped, it has to be an integer. Therefore, after taking the square root of Equation 12, we also need to take the ceiling to find the value of $\alpha_{\min}$,

$$\alpha_{\min} = \ell \left\lceil \sqrt{\left( \frac{1 - \rho_{\max}}{2} \right)} \right\rceil. \tag{13}$$

Using a similar logic, $\alpha_{\max}$ can be found as

$$\alpha_{\max} = \ell \left\lfloor \sqrt{\left( \frac{1 - \rho_{\min}}{2} \right)} \right\rfloor. \tag{14}$$

### B. An ensuing remark on feasibility of block-swapping

Note that as explain in Section IV-A, if $\epsilon < 1/\ell^2$, then $\rho^* = \rho_{\max} = \rho_{\min}$. In that case, also $\alpha = \alpha_{\max} = \alpha_{\min}$ and the above swapping strategy still applies. However, in that case, if the square root of the term in Equation 12 is not an integer, i.e.

$$\epsilon < \frac{1}{\ell^2} \ \wedge \ \sqrt{\left( \frac{1 - \rho^*}{2} \right)} \notin \mathbb{Z}, \tag{15}$$

then there is no solution to the problem.

### C. Advantages of the block-swapping method

As discussed in detail in [5], the performance of `RBSC-Subgen` is highly sensitive to the hyper-parameters, which may often result in saturation. For that reason, it is not possible to accurately determine the running time. Nevertheless, due to confirmation of RBSC coefficient at every step, one can at the very least say that it is no lower than quadratic time, i.e. $\geq \mathcal{O}(\ell^2)$.

However, unlike `RBSC-Subgen`, the computational load of block-swapping strategy can be determined accurately. To be specific, the only part of the method, which relies on the hyper-parameters of the problem, is the sorting of $U_s$, whereas the other parts require simple arithmetic operations, which can all be realized in linear time, i.e. $\mathcal{O}(1)$, making them negligible. In that respect, computational complexity due to the sorting of $U_s$ is what determines the running time of the proposed method, which can be implemented with an efficient sorting algorithm. For instance, if this part is implemented with `QuickSort`, the running time of block-swapping will be log-linear, i.e. $\mathcal{O}(\ell \log \ell)$.

In addition, the block-swapping strategy does not suffer from the saturation problem of `RBSC-Subgen`, since it does not employ iterations. It also enables an accurate assessment of feasibility through Equations 3 and 15. This saves the effort to attempt for a solution, which would be in vain with a brute-force manner.

Note also that the block-swapping strategy sustains the capability of `RBSC-Subgen` of producing outcomes with a certain level of diversity. Since (i) the initial subsets $A$ and $B$ are built out of a randomly sampled subset of the universal set $U$ and (ii) the number of elements to be swapped $\alpha$ is chosen arbitrarily (should that be possible), block-swapping strategy does not yield deterministic outcomes.

TABLE I
HYPER-PARAMETERS USED FOR THE SAMPLE IMPLEMENTATIONS.

| Hyper-parameter | Value set-1 | Value set-2 |
|---|---|---|
| Distribution | Normal | Normal |
| $\|U\|$ | 30000 | 3000 |
| $\ell$ | 1000 | 1000 |
| $\rho^*$ | 0.8 | 0.9 |
| $\epsilon$ | 0.02 | 0.002 |

### D. Disadvantages of the block-swapping method

Although the block-swapping strategy satisfies the conditions imposed on $\rho^*$, it does not provide continuous distributions. Namely, since a certain continuous piece of the subsets are cropped and swapped, one may see two "holes" in the distributions of the raking variables.

As an example see Figure 1, which is obtained setting the hyper-parameters as given in Table I. Note that in Table I hyper-parameter value set-1 describes a less strict problem than value set-2, since it has a larger universal set, a lower value of $\rho^*$ and a larger margin on permissible error.

In Figures 1-(a) and (b), one may clearly see that that $A$ has mostly the items with negative scores, but there is a discontinuity around the mean score (i.e. 0). Similarly, $B$ has mostly the items with positive scores but also several items with scores just below the mean, whereas no items with scores slightly over the mean. This indicates the fact that while the subsets satisfy the quantitative conditions, they are discontinuous and qualitatively quite artificial.

Depending on the application, this discontinuity may lead to serious problems, since the bulk of the distributions are concentrated around the mean and destabilizing this range may affect the behavior of the distributions in a significant way.

### E. Alternating-swapping approach

The discontinuity issue can be overcome by swapping *some elements over a certain range*, such that there is no abrupt change in the assignment of elements.

As an example, consider that every second element among the largest $2\beta$ elements of $A$ and every second element in the smallest $2\beta$ elements of $B$ are moved to one another [2] This will result in swapping half of the elements in these ranges, where the other half is retained, and the elements out of these ranges are simply not-touched.

An important point in resolving $\beta$ is that, if every other element in the smallest/largest $2\beta$ elements are swapped, then the number of contradicting evidence $C$ is no more directly equal to the square of the number of displaced elements $\beta$. Let us now derive the relation between $\beta$ and $S$.

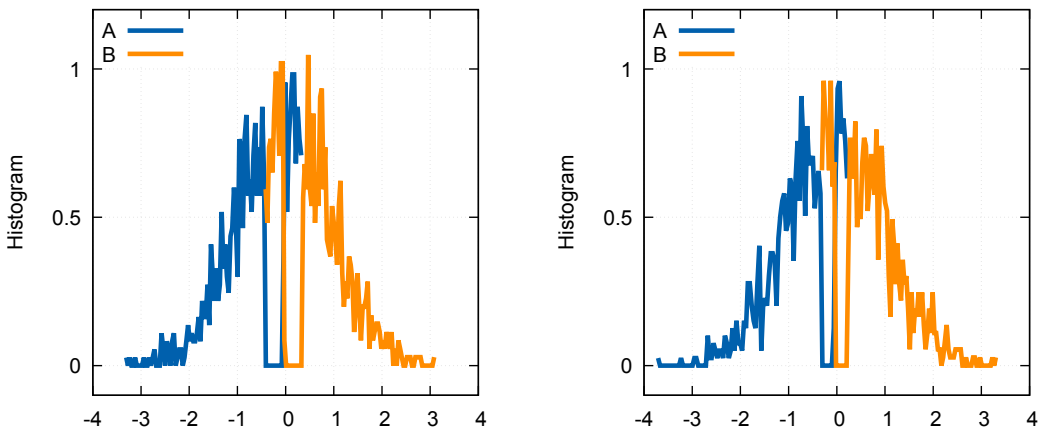
Regarding each element that is retained in $A$, i.e. $a_i$ for $i < \ell - 2\beta$, there will only be supporting evidence. Namely, for $i < \ell - 2\beta$, the number of supporting evidence associated with $a_i$ will be $S(i) = \ell$, which implies that the number of contradicting evidence associated with $a_i$ will be $C(i) = 0$. On the other hand, for each element that is retained in $A$ with $i \geq \ell - 2\beta$, we will get $\ell - i$ supporting evidence and $i$ contradicting evidence. Namely, the number of supporting and contradicting evidence associated with $a_i$ for $i \geq \ell - 2\beta$ will be $S(i) = \ell - i$ and $C(i) = i$, respectively. Thus, the total number of contradicting evidence associated with those retained elements will be

$$\sum_{i=0}^{\beta-1} i. \tag{16}$$

On the other hand, there will be also be some contradicting evidence associated with the elements which are moved from $B$ to $A$. Specifically,

$$\sum_{i=\beta}^{2\beta-1} i. \tag{17}$$

[2]In the following formulation, we consider moving $a_{\ell-2\beta+2}$ and every second element succeeding it until $a_{\ell-2\beta+2}$ concerning $A$. We consider a similar way concerning $B$. However, one may opt to move $a_\ell$ and every second element preceding it concerning $A$, and $b_{2\beta}$ and every second element preceding it until $b_{2\beta-1}$ concerning $B$. The particulars of computation are slightly different and in this article we address those relating to the first approach.




(a) (b)

Fig. 1. Histograms of item scores belonging to subsets $A$ and $B$ with block-swapping and (a) hyper-parameter value set-1 and (b) hyper-parameter value set-2.

Eventually, the total number of contradicting evidence is obtained as the sum of Equations 16 and 17 as

$$C = \sum_{i=0}^{2\beta-1} i, \tag{18}$$

which can be simplified as

$$C = (2\beta - 1)\beta. \tag{19}$$

Using Equation 5, the number of supporting evidence $S$ is obtained as

$$S = \ell^2 - (2\beta - 1)\beta, \tag{20}$$

Replacing Equation 20 in Equation 6, the RBSC coefficient$\rho$ is obtained as follows

$$\rho = \frac{\ell^2 - (2\beta - 1)\beta - (2\beta - 1)\beta}{\ell^2}. \tag{21}$$

Organizing the terms of Equation 21, we get the following quadratic polynomial

$$4\beta^2 - 2\beta + \rho\ell^2 - \ell^2 = 0. \tag{22}$$

We denote the lowest and highest possible values of $\beta$ satisfying the constrains with $\beta_{\min}$ and $\beta_{\max}$. We can solve for $\beta_{\min}$ (corresponding to the highest permissible value of $\rho_{\max}$) as the smallest integer larger than the positive root of the quadratic Equation 21, replacing $\rho$ with $\rho_{\max} = \rho^* + \epsilon$. This yields

$$\beta_{\min} = \left\lceil \frac{2 + \sqrt{4 - 4 \cdot 4 \cdot ((\rho^* + \epsilon)\,\ell^2 - \ell^2)}}{8} \right\rceil. \tag{23}$$

Similarly, $\beta_{\max}$, can be solved as

$$\beta_{\max} = \left\lfloor \frac{2 + \sqrt{4 - 4 \cdot 4 \cdot ((\rho^* - \epsilon)\,\ell^2 - \ell^2)}}{8} \right\rfloor. \tag{24}$$

In Figures 2-(a) and (b), we present the distributions of subsets $A$ and $B$ constructed with the alternating-swapping approach. One may see in these figures that although $A$ has mostly the items with negative scores, but there are also some items with slightly higher scores than the mean. Similarly, $B$ has mostly the items with positive scores but there are also some items in $B$ with scores a little below the mean. Therefore, the distributions of the subsets are no longer discontinuous and they diffuse into each other to a certain extent.

### *F. An ensuing remark on feasibility of alternating-swapping*

According to the alternating-swapping method, we can pick a $\beta$ in the range $[\beta_{\min}, \beta_{\max}]$ and carry out the alternating-swapping accordingly. Note that the limiting condition for obtaining subsets $A$ and $B$ with $\rho(A, B) \in [\rho^* - \epsilon, \rho^* + \epsilon]$ based on this revised method, is to achieve a $\beta$ such that $2\beta < \ell$.

## VI. Conclusions

In this study, we focus on the development of an algorithm for solving the subset selection problem with predefined constraints on the ranking relation between the subsets. The proposed method is is an alternative the the previously proposed `RBSC-SubGen`, which is computationally expensive, has unforeseeable running time and saturation issues. The idea underlying the proposed approach is to firstly initialize the subsets with an extreme value of RBSC coefficient. e.g. $\rho = 1$, and then update them by introducing a certain amount of supporting/contradicting evidence such that $\rho$ is eventually confined into the permissible range defined by the user. We implemented this approach initially with swapping of a block of (sequential) elements between the subsets and then by swapping alternating elements over a certain segment of the subsets. The block-swapping implementation is considered as

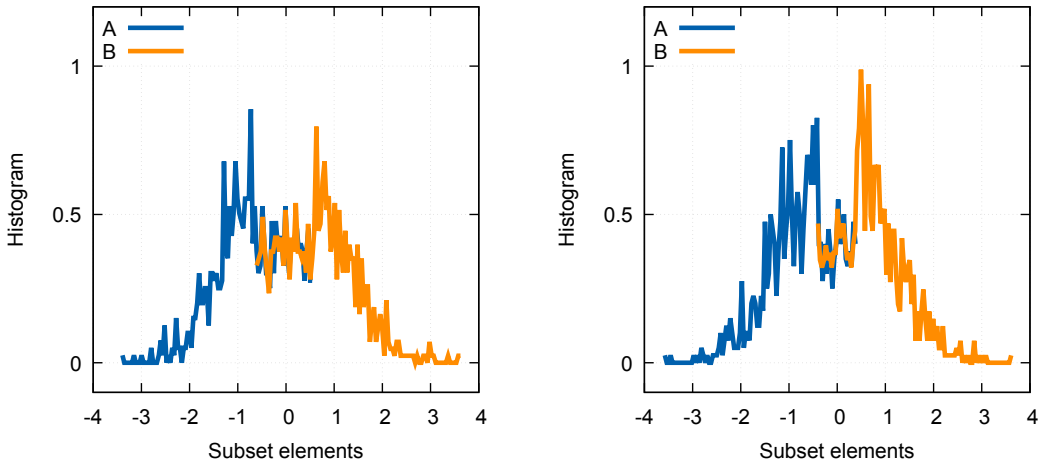



Fig. 2. Histograms of item scores belonging to subsets $A$ and $B$ with alternating-swapping and (a) hyper-parameter value set-1 and (b) hyper-parameter value set-2.

a convenient tool for a simple explanation of the proposed approach. By running it on a hypothetical set, we illustrate that it attains the desired constraints on $\rho$, but yields subsets with discontinuous distributions. As a remedy to this issue, we offer alternating-swapping, which yields subsets with diffusing distributions and solves the discontinuity issue to a certain degree. The proposed method has the advantage that it provides a clear insight into the feasibility of the problem such that superfluous computations can be avoided. Also unlike `RBSC-SubGen`, we can also comment on the computational complexity in a reliable manner. In the future, we are planning to apply the proposed method and previous alternatives on real-world data and demonstrate its efficacy and practical uses.

## REFERENCES


[1] L. J. Hong, W. Fan, and J. Luo, "Review on ranking and selection: A new perspective," *arXiv preprint arXiv:2008.00249*, 2020.

[2] T. Foss, E. Stensrud, B. Kitchenham, and I. Myrtveit, "A simulation study of the model evaluation criterion MMRE," *IEEE Tran. Software Engineering*, vol. 29, no. 11, pp. 985–995, 2003.

[3] D. J. Eckman, M. Plumlee, and B. L. Nelson, "Revisiting subset selection," in *Proc. Winter Simulation Conf.*, pp. 2972–2983, IEEE, 2020.

[4] Z. Yücel, P. Supitayakul, A. Monden, and P. Leelaprute, "An algorithm for automatic collation of vocabulary decks based on word frequency," *IEICE Tran. Information and Systems*, vol. 103, no. 8, pp. 1865–1874, 2020.

[5] K. Furuya, Z. Yücel, P. Supitayakul, A. Monden, and P. Leelaprute, "Exploring thelimitsofanrbsc-basedapproachinsolvingthesubsetselectionproblem," in *Proceedings of International Conference on Smart Computing and Artificial Intelligence*, vol. 81 of *EPiC Series in Computing*, pp. 1–9, EasyChair, 2021.

[6] S. S. Gupta, "On some multiple decision (selection and ranking) rules," *Technometrics*, vol. 7, no. 2, pp. 225–245, 1965.

[7] L. Huang, J. Wei, and E. Celis, "Towards just, fair and interpretable methods for judicial subset selection," in *Proc. AAAI/ACM Conf. AI, Ethics, and Society*, pp. 293–299, 2020.

[8] G. C. McDonald, "Applications of subset selection procedures and Bayesian ranking methods in analysis of traffic fatality data," *Computational Statistics*, vol. 8, no. 6, pp. 222–237, 2016.

[9] C.-H. Yeh, B. A. Barsky, and M. Ouhyoung, "Personalized photograph ranking and selection system considering positive and negative user feedback," *ACM Tran. Multimedia Computing, Communications, and Applications*, vol. 10, no. 4, pp. 1–20, 2014.

[10] C.-H. Chen, S. E. Chick, L. H. Lee, and N. A. Pujowidianto, "Ranking and selection: Efficient simulation budget allocation," *Handbook of Simulation Optimization*, pp. 45–80, 2015.

[11] S. H. Choi and T. G. Kim, "A heuristic approach for selecting best-subset including ranking within the subset," *IEEE Tran. Systems, Man, and Cybernetics-A*, vol. 50, no. 10, pp. 3852–3862, 2018.

[12] M. H. Alrefaei and M. Almomani, "Subset selection of best simulated systems," *Journal of the Franklin Institute*, vol. 344, no. 5, pp. 495–506, 2007.

[13] S. Gao and W. Chen, "A note on the subset selection for simulation optimization," in *Proc. Winter Simulation Conf.*, pp. 3768–3776, IEEE, 2015.

[14] C.-H. Chen, D. He, M. Fu, and L. H. Lee, "Efficient simulation budget allocation for selecting an optimal subset," *INFORMS Journal on Computing*, vol. 20, no. 4, pp. 579–595, 2008.

[15] M. J. Groves, *Efficient Pairwise Information Collection for Subset Selection*. PhD thesis, University of Warwick, 2020.

[16] L. J. Hong, J. Luo, and Y. Zhong, "Speeding up pairwise comparisons for large scale ranking and selection," in *Proc. Winter Simulation Conf.*, pp. 749–757, IEEE, 2016.

[17] Y. Wang, L. Luangkesorn, and L. J. Shuman, "Best-subset selection procedure," in *Proc. Winter Simulation Conf.*, pp. 4310–4318, IEEE, 2011.

[18] S. Gao, H. Xiao, E. Zhou, and W. Chen, "Robust ranking and selection with optimal computing budget allocation," *Automatica*, vol. 81, pp. 30–36, 2017.

[19] M. Mitchell, D. Baker, N. Moorosi, E. Denton, B. Hutchinson, A. Hanna, T. Gebru, and J. Morgenstern, "Diversity and inclusion metrics in subset selection," in *Proc. AAAI/ACM Conf. AI, Ethics, and Society*, pp. 117–123, 2020.

[20] D. S. Kerby, "The simple difference formula: An approach to teaching nonparametric correlation," *Comprehensive Psychology*, vol. 3, pp. 11–IT, 2014.