疑似希望順位データ作成の説明書

2021 年 5 月 5 日

筑波大学社会工学類マッチング委員

1　はじめに

この説明書では、教員と学生の二部マッチングのシミュレーション研究などに利用可能な疑似希望順位データ作成の方法と付属のサンプルデータについて説明する。疑似データ作成には、順位ベクトルを確率変数とする確率モデルを利用する。そのようなモデルの代表的なものに Mallows (1957) および Luce (1959) がある。2 つのモデルについては、Qin et al. (2010) の 1 節及び 2 節において簡潔な説明があり、この文書内の説明もおおむね Qin et al. (2010) に従っている。この文書では、Mallows (1957) に従った疑似希望順位データ生成方法を説明する。

まず、Mallows (1957) モデルについて簡単に説明する。以下では順位付けの対象を教員と呼ぶ。対象となる教員の集合を $N = \{1, \ldots, n\}$ で表す。順位付け π を N から N への写像として定義する（$\pi : N \to N$）。この時、$\pi(i)$ は教員 i の順位を表す。表記上の混乱がない場合は、π は教員につけられた順位のベクトルを表す、つまり $\pi = \langle \pi(1), \pi(2), \ldots, \pi(n) \rangle$。同様に π^{-1} は順位の順に並んだ教員のベクトル、$\pi^{-1} = \langle \pi^{-1}(1), \pi^{-1}(2), \ldots, \pi^{-1}(n) \rangle$ を表す。

順位のベクトルの集合を S_n とする。$\sigma \in S_n$ を基準となる順位とする。また、関数 $d : S_n \times S_n \to \mathbb{R}$ を 2 つの順位付けの距離を表す距離関数とし、$\theta \in \mathbb{R}$ を距離関数に重み付けをするパラメータとする。ある順位付け $\pi \in S_n$ の確率 $P(\pi|\theta, \sigma)$ は以下の式で与えられる。

$$P(\pi|\theta, \sigma) = \frac{\exp(-\theta d(\pi, \sigma))}{Z(\theta, \sigma)}, \tag{1}$$

$$Z(\theta, \sigma) = \sum_{\hat{\pi} \in S_n} \exp(-\theta d(\hat{\pi}, \sigma)). \tag{2}$$

上式に示されるように $\theta > 0$ であれば基準となる順位ベクトル σ に距離の近い順位ベクトルがより大きい確率を持つ。いくつかの距離関数を紹介する。代表的な距離関数には Spearman の順位相関 (Spearman rank correlation) がある。これは以下の式で計算される。

$$d(\pi, \sigma) = \sum_{i \in N} (\pi(i) - \sigma(i))^2, \qquad \forall \pi, \sigma \in S_n \tag{3}$$

他の距離関数には Kendall の距離関数 (Kendall's tau) や Spearman の物差し (Spearman's footrule) がある。Kendall の距離関数は、２つの順位ベクトルで相対的な順位が異なる教員ペアの数で計算される。つまり、以下のように表現される距離関数である。

$$d(\pi, \sigma) = \sum_{i \in N} \sum_{j > i} \mathbb{1}[(\pi(i) > \pi(j) \Rightarrow \sigma(i) < \sigma(j)) \text{ or } (\pi(i) < \pi(j) \Rightarrow \sigma(i) > \sigma(j))] \qquad \forall \pi, \sigma \in S_n \tag{4}$$

Spearman の物差し (Spearman's footrule) は、Spearman の順位相関と同様であるが、各要素の距離に L1 距離を用いる。

$$d(\pi, \sigma) = \sum_{i \in N} |\pi(i) - \sigma(i)|, \qquad \forall \pi, \sigma \in S_n \tag{5}$$

また、Langville and Meyer (2012)(日本語訳は Langville et al. (2015)) では、2 つのランキングリストについて「リスト間の

うえのほうの相違は、下のほうの相違よりもより重んじられる」距離を与える Spearman の重み付け物差しを提案している。

$$d(\pi,\sigma)=\sum_{i\in N}\frac{|\pi(i)-\sigma(i)|}{\min\{\pi(i),\sigma(i)\}},\qquad\forall\pi,\sigma\in S_n \tag{6}$$

これらの距離関数をより詳しく知りたい場合は、Langville and Meyer (2012)(日本語訳は Langville et al. (2015)) の 16 章が参考となる。

2　準備

疑似データ作成にあたって、以下のことを事前に決める。

- 教員の数
- 基準となる順位
- 距離関数
- 分散パラメータ θ

基準となる順位は、$\theta>0$ であれば生成される確率が常に最も高い。例えば教員と学生の 2 部マッチング用のデータ作成であれば、最も確率が高そうな（最も多くの学生が持っていそうな）教員に対する順位を基準の順位として選ぶのが適当である。距離関数は、学生の選好の違いを適切に表現可能なものを選ぶ。一般的な順位ベクトルの相違には、前述の Kendall の距離関数や Spearman の順位相関が用いられる。分散パラメータは生成されたデータの順位相関に影響を与えるが、ある程度トライアル&エラーのプロセスなどを経て決める必要がある。

3　データ生成

データ生成は本文書に付属する Python プログラム (pseudo_ranking.py) にて行うことが可能である。生成可能なデータについては使用 PC の性能により教員数に制限を置かなくてはならない場合がある。この点については、5 節を参照すること。pseudo_ranking.py の冒頭部分で以下の変数を適宜変更することで、データ生成に用いるパラメータ制御が可能である。

変数名 (pseudo_ranking.py ファイル内)	変数の説明
num_faculty	教員の数を指定
num_students	生成されるデータ個数を指定
distfunc	1〜3 を指定
theta	パラメータ θ を指定

距離関数は、Spearman の順位相関、Spearman の物差し、Spearman の重み付け物差しが実装されている。「distfunc」変数に、1 を指定すると Spearman の順位相関、2 を指定すると Spearman の物差し、3 を指定すると Spearman の重み付け物差しが距離関数として使用される。基準となる順位ベクトルは、教員 1 を第一位、教員 2 を第二位、. . . とするベクトルと設定されている。これはプログラム内の「rankbase」変数を変更することで、変更可能である。

4　サンプルデータ

付属の各サンプルデータで使用したパラメータなどを以下の表に示す。基準となる順位はすべて教員 1 を第一位、教員 2 を第二位、. . . として生成した。

ファイル名 (csv ファイル)	教員数	距離関数	θ	データ個数
Sample_N10_S100_theta0.1_spearmanRankCorr	10	Spearman の順位相関	0.1	100
Sample_N10_S100_theta1_spearmanRankCorr	10	Spearman の順位相関	1.0	100
Sample_N10_S100_theta0.1_spearmanFootrule	10	Spearman の物差し	0.1	100
Sample_N10_S100_theta1_spearmanFootrule	10	Spearman の物差し	1.0	100
Sample_N10_S100_theta0.1_spearmanWeighted	10	Spearman の重み付け物差し	0.1	100
Sample_N10_S100_theta1_spearmanWeighted	10	Spearman の重み付け物差し	1.0	100
Sample_N12_S100_theta0.1_Spearmanweighted	12	Spearman の重み付け物差し	0.1	100

サンプルデータの様式を以下に示す。n を教員の数、m を作成されたデータ個数とする。

- 各 $i \in \{1, \ldots, m\}$ 行に確率モデルに従い生成された順位ベクトル 1 個を格納。
- 1 列目に順位ベクトルの識別子を格納。識別子は生成された順に $s1, \ldots, sm$ とした。
- 各 $j \in \{2, \ldots, n+1\}$ 列に順位ベクトルを格納。j 列の数値は、教員 $j-1$ の順位を示す。

5 留意点

データ生成には、n の要素の順列の組み合わせ（$n!$ 通り）を考慮する必要がある。n の値によって計算に多大なメモリと時間が必要となる場合がある。Windows 10, Core i7, 96GB メモリの PC を使用して、$n = 12$ に対するデータ 100 個生成（教員 12 人に対する学生の選好を 100 人分作成に相当）に約 40 分程度が必要であった。$n \geq 13$ ではメモリ不足でエラーとなった。Mallows (1957) モデルの計算量については Qin et al. (2010) の 2.2 節でも膨大になりえると指摘されており、プログラムのメモリ使用の効率化は検討事項である。

参考文献

Langville, A. N. and Meyer, C. D. (2012). *Who's #1?: The Science of Rating and Ranking*. Princeton University Press.

Langville, A. N., Meyer, C. D., 岩野和生 (翻訳), 中村英史 (翻訳), and 清水咲里 (翻訳) (2015). **レイティング・ランキングの数理**—*No.1* **は誰か？**—. 共立出版.

Luce, R. D. (1959). *Individual Choice Behavior*. John Wiley.

Mallows, C. L. (1957). NON-NULL RANKING MODELS. I. *Biometrika*, 44(1-2):114–130.

Qin, T., Geng, X., and Liu, T.-y. (2010). A new probabilistic model for rank aggregation. In Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.