数理解析研究所講究録 1383 巻 2004 年 18-25

ランクつき投票モデルにおける類似度分析 —投票人数に関する考察—

大分大学 工学部 小畑 経史 (Tsuneshi OBATA)
Faculty of Engineering, Oita University
obata@csis.oita-u.ac.jp
大阪大学大学院 情報科学研究科 石井 博昭 (Hiroaki ISHII)
Graduate School of Information Science and Technology, Osaka University
ishii@ist.osaka-u.ac.jp

Abstract: 投票モデルにおいて単一投票にはいくつかの問題があることが知られており，そのため，それぞれの投票者が複数の票を持つような投票モデルが望ましい．このようなモデルによってえられる投票データには，同一の投票者によって同時に支持された候補者，という意味での，候補者どうしの類似性についての情報が含まれている．そこで，我々は，投票者が複数の候補にランクをつけて投票するランクつき投票モデルにおいて，この情報を利用して候補者の帰化的な位置関係や距離を評価する手法を提案した．本報告ではシミュレーション実験により，この手法が候補者の真の位置関係を再現できるかを，主に投票する候補者の人数による違いに注目して，調査する．

Keywords: ランクつき投票モデル，候補者の類似性，多次元尺度法 (Multidimensional Scaling, MDS)，データ包絡分析法 (Data Envelopment Analysis, DEA)

1 ランクつき投票モデル

複数の人々の意見を集約し，候補／選択肢から最も望ましいものを選択する，あるいはそれらを望ましい順に順位づけするために，しばしば投票という手段がとられる．その際に，おのおのの投票者が最も望ましい候補一人だけに票を投じる単一投票は，必ずしも妥当とはいえないことが知られており [9]，複数の候補に投票するモデルが望ましいとされる．

ここで取り扱うランクつき投票モデルとは，各投票者が全候補者の中から自分が望ましいと思うものから順に上位数名 (何人まで投票するかは選挙実施者によりあらかじめ決められているものとする) を順位つきで投票するモデルである．投票者は投票用紙に指定された人数だけ自分が望ましいと思う候補者を (順位つきで) 記入し，投票する．このときに得られる投票データは表 1 のようなものとなる．

このような投票モデルにおいて，得られた投票データは通常，各候補が獲得した票を順位ごとに集計した上で (たとえば表 2 のように)，当選者の決定，あるいは候補者の順位づけに利用される．このようなデータをもとに当選者を決定する／候補者を順位づけるには，ランクごとの得票数に何らかのウェイトをつけて集計したスコアにより各候補の選好度合いを数値化し，これを比較することが自然な方法である．

これ以降，候補者の人数を m，投票者の人数を n，各投票者が投票する候補者の人数を k (すなわち，各投票者は m 人の候補者の中から望ましい順に上位 k 人を選んで投票用紙に記入する) とし，第 i 候補を C_i，第 l 投票者を V_l と表すことにする．

表 1: ランクつき投票モデルにおける原データ

	投票者 1	投票者 2	...
1 位	候補者 A	候補者 C	...
2 位	候補者 B	候補者 D	...
⋮	⋮	⋮	

表 2: ランクごとに集計された投票データ

	1 位	2 位	...
候補者 A	32	10	...
候補者 B	28	20	...
⋮	⋮	⋮	

このとき上で述べた (第 i 候補 C_i にとっての) 選好スコアは次のように定義できる.

$$Z_i = \sum_{j=1}^{k} w_j v_{ij} \qquad i = 1, \ldots, m.$$

ここで v_{ij} は候補 C_i が得た第 j 位票の数, w_j は第 j 位票のウェイトである. このスコアを評価する際には, 当然ながら, ウェイト w_j の決め方が非常に重要となる. しばしば, ウェイト値をあらかじめ決定しておき (たとえば $w_1 = 10, w_2 = 5, \ldots$ のように), すべての候補のスコアをこのウェイトを用いて評価する方法が採られるが, ウェイトの値を変えると当選する候補も変わりうる. したがって誰からも文句の出ないウェイトをひとつ決定することは不可能と言ってよく, ウェイトの決定にはどうしても何らかの恣意性が含まれる. そのため DEA (data envelopment analysis, データ包絡分析法) をベースにして, 各候補にとって有利なウェイトで評価することのできる手法が提案されてきた [2, 3, 6]. DEA を利用することで, 各候補は自分にとって有利なウェイトを選好スコアの評価に用いることができる.

ところで, DEA においては, 類似したデータが存在するかどうかが, データの効率性評価に大きく影響する. いまのケースではデータが類似しているとは, 表 2 の得票数が似ていることをいい, 必ずしも候補がその政策や特徴の面で似ていることを意味しない. したがって, 候補者の類似性が結果として得られる順位づけにどのように関わってくるのかはわからない. 特に複数の (しかし少数の) 候補が当選するケースでは当選者が政策や特徴の面で互いに似ているかどうかは, 投票者の意思を広く反映するかどうかに大きく関わるため, 候補者の政策や特徴の面での類似性が評価できると, さらにそれを候補者の順位づけに利用できると非常に意義深いであろう.

そこで, 我々はランクごとに集計する前の投票データ (表 1) を用いて候補者の類似性を評価する手法を提案した [7, 8]. この手法は, 同じ投票者が同時に支持する候補は何らかの類似性を持っている, という考えに基づき, MDS (multidimensional scaling, 多次元尺度法) により候補者の空間的な位置関係を評価するものである.

2 候補者間の類似性評価手法

我々が提案した，ランクつき投票データから候補者間の類似性を評価する手法を紹介する．ここで，投票者 V_l により第 j 位にランクづけられた候補のインデックスを i_{lj} とあらわすことにする．したがって投票者 V_l は $C_{i_{l1}}, C_{i_{l2}}, \ldots, C_{I_{lk}}$ の順で投票を行う．

我々の基本的な考えは，同じ投票者が同時に支持した候補は何らかの類似性を持つ，という非常に単純なものである．さらに，ある候補の組に対して，彼らを同時に支持する投票者が多ければ，それらの候補者間の類似性が高いと判断してよいのではないかと考えた．

まず，$k=2$ の場合，すなわち，各投票者が上位二人を投票する場合を考える．このとき，投票者 V_l は $C_{i_{l1}}$ を 1 位，$C_{i_{l2}}$ を 2 位として票を投じる．ここで，C_i を 1 位に C_j を 2 位にランクづけた投票者の人数を s_{ij} とおく，すなわち，

$$s_{ij} = \#\{V_l | C_i = C_{i_{l1}} \text{ and } C_j = C_{i_{l2}}\}, \quad i,j = 1,\ldots,m,$$

ただし，記号 # は集合の要素数を意味する．もし候補者 C_i と C_j が似ていれば多くの投票者が C_i と C_j をともに支持し s_{ij} が大きくなると考えられる．また逆にこれらの候補が似ていなければ，どちらかの候補を支持する投票者が同時にもう一方の候補を支持するとは考えにくく，s_{ij} は小さな値となると思われる．したがって s_{ij} の大きさによって，C_i と C_j の間の類似性の高さを判断してよいだろう．

また，$k>2$ の場合には，s_{ij} を

$$s_{ij} = \sum_{q=1}^{k-1} s_{ij}^{(q)}, \quad i,j = 1,\ldots,m,$$

ただし

$$s_{ij}^{(q)} = \frac{1}{q}\#\{V_l | C_i = C_{i_{lq}} \text{ and } C_j = C_{i_{lq+1}}\}, \quad i,j = 1,\ldots,m; q = 1,\ldots,k-1,$$

と定める．これは C_i と C_j を隣り合った順位にランクづけた投票者の人数に，下位にいくほど軽くなる重みをかけて合計したものである．下位の候補，すなわちその投票者に好まれない候補，についても，ともに似ているために同じ様に下位にランクづけられた，という情報が含まれていると考えたためである．ただし，これには，似てはいないが好まれないというだけでたまたま隣り合った順位にランクづけられる，というケースも考えられるため，それを考慮して重みを軽くした[1]．

この s_{ij} に，対称化:

$$\bar{s}_{ij} = s_{ij} + s_{ji}, \quad i,j = 1,\ldots,m$$

と，基準化:

$$\bar{\bar{s}}_{ij} = \frac{\bar{s}_{ij}}{\bar{s}_{i+} + \bar{s}_{j+} - \bar{s}_{ij}}, \quad i,j = 1,\ldots,m,$$

ただし $\bar{s}_{i+} = \sum_k \bar{s}_{ik}$，を施した上で，類似性／非類似性の情報から対象間の距離や空間的な配置を分析するための手法である，MDS (multidimensional scaling，多次元尺度法) [4, 5, 10] を適用しようというのが，我々の提案した手法である [7, 8]．MDS を適用することにより，分析結果は候補者の空間的な配置として得られる．類似性を数値として表したい場合にはそれをもとに候補者間の距離を求め，それを用いればよい．

[1] 我々の最初の提案ではこのような重みを考えていない．

3 投票行動モデルと実験

我々の提案手法によってどの程度本来の候補者の類似性をとらえることが出来るかを調べるためにシミュレーション実験を行う．実験に先立ち，投票者の行動をシミュレートするためのモデルを提案する．このモデルは Gill and Gainous [1] が提案した投票空間モデルに似たものである．

我々の提案するモデルでは投票者 (および候補者) は以下のように行動するものとする．

1. おのおのの候補者は p 次元ユークリッド空間上の点として配置される．このとき点の座標はその候補の性格や政策によって定まる．
2. おのおのの投票者は同じ空間上の点として配置される．このとき点の座標はその投票者の理想とする性格や政策によって定まる．
3. おのおのの投票者は自分が理想とする点に近い候補にほど，より高い好感度を持つ．
4. おのおのの投票者は好感度の高い候補から順に候補者にランクをつけ，それにしたがってランクつき投票を行う．

今回の実験では，

1. 候補者と投票者は 2 次元空間に配置される (すなわち $p = 2$)，
2. すべての候補者と投票者は原点を中心とする 2 次元正規分布

$$N(\mu, \Sigma), \qquad \mu = (0,0)^{\mathrm{T}}, \Sigma = \mathrm{diag}(3,3)$$

にしたがって分布する，

ものとして，疑似的に候補者，投票者を生成することとする．さらに，MDS では通常，分析対象が分布する空間の次元について考慮する必要があるが，ここでは疑似的に生成する候補者の分布する空間を 2 次元としているため，MDS での分析でも求める候補者の配置の次元も 2 次元とする．

我々の手法では結果が候補者の空間的な配置として得られるが，これが最初に生成した候補者の本来の配置とくらべてどれくらいずれているかを計測したい．しかし，MDS で得られる配置は，回転，拡大縮小，平行移動，裏返しを施しても本質的な違いがない．そのため，得られた配置にこれらの変換を施した上で，本来の配置と比較する必要がある．そこで，対応する候補の座標間の距離の 2 乗和が最小になるように変換を施し，この和によって本来の配置とのずれを測ることにする．すなわち，最小化問題

$$r^2 = \min_T \frac{1}{m} \sum_{i=1}^{m} \|x_i - T(\bar{x}_i)\|^2$$

の解 r^2 を得られた配置のずれと考える．ここで，$x_i = (x_{i1}, x_{i2})^{\mathrm{T}}$ を第 i 候補の本来の座標，$\bar{x}_i = (\bar{x}_{i1}, \bar{x}_{i2})^{\mathrm{T}}$ を提案手法で得られた第 i 候補の座標，T を任意の回転，拡大・縮小，平行移動，裏返し変換とする．

2 次元の場合，$\bar{x}_i$ に対する回転，拡大・縮小，平行移動を施した点 $T^+(\bar{x}_i)$ は

$$T^+(\bar{x}_i) = \begin{pmatrix} a & b \\ -b & a \end{pmatrix} \bar{x}_i + \begin{pmatrix} s \\ t \end{pmatrix}$$

と表される．このような変換 T^+ のうち，ずれ r^2 を最小にするものは，線形方程式

$$\begin{pmatrix} A & 0 & B_1 & B_2 \\ 0 & A & B_2 & -B_1 \\ B_1 & B_2 & m & 0 \\ B_2 & -B_1 & 0 & m \end{pmatrix} \begin{pmatrix} a \\ b \\ s \\ t \end{pmatrix} = \begin{pmatrix} C_+ \\ C_- \\ D_1 \\ D_2 \end{pmatrix}$$

の解 a^*, b^*, s^*, t^* によって得られる．ここで，

$$A = \sum_{i=1}^{m} (\bar{x}_{i1}^2 + \bar{x}_{i2}^2), \qquad C_+ = \sum_{i=1}^{m} (x_{i1}\bar{x}_{i1} + x_{i2}\bar{x}_{i2}),$$

$$C_- = \sum_{i=1}^{m} (x_{i1}\bar{x}_{i2} - x_{i2}\bar{x}_{i1}),$$

$$B_1 = \sum_{i=1}^{m} \bar{x}_{i1}, \qquad D_1 = \sum_{i=1}^{m} x_{i1},$$

$$B_2 = \sum_{i=1}^{m} \bar{x}_{i2}, \qquad D_2 = \sum_{i=1}^{m} x_{i2}$$

である．裏返しに関しては，$\bar{x}_i$ を裏返した点 $\bar{x}_i^- = (\bar{x}_{i1}, -\bar{x}_{i2})^{\mathrm{T}}$ に対して上と同様にすればよい．そのようにして得られたずれのうち小さいものが求める最少のずれであり，それを与える変換を施した配置が我々の手法により評価された候補者の配置である．

図 1 に候補者数 $m = 10$ のときに得られた配置の例を示す．左から順に，疑似的に生成された候補者の配置，我々の手法で得られた配置，ずれが最小となるように変換した配置，である．ちなみにこのときの r^2 の値は 0.5170 である．

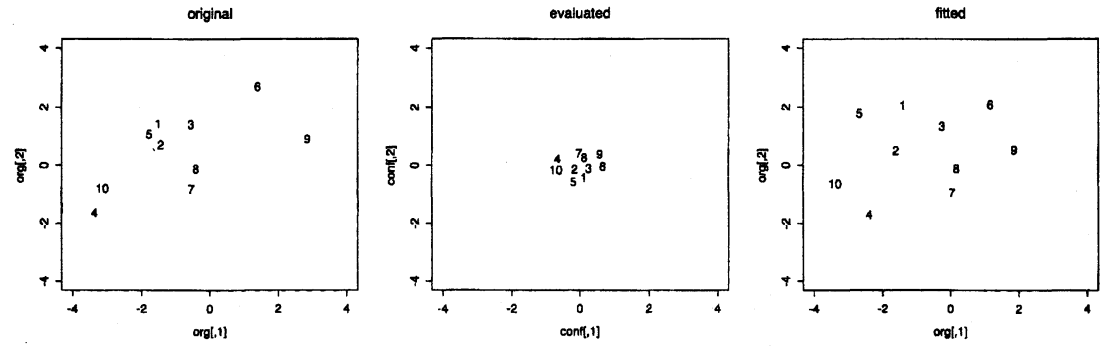

図 1: 得られた配置の例

[実験]

Step 1 m 人の候補者をランダムに生成

Step 2 n 人の投票者をランダムに生成

Step 3 おのおのの投票者から「近い」順に上位 k 人の候補者に投票

Step 4 我々の提案手法を用いて投票データを分析し，候補者の配置を評価

Step 5 得られた配置と本来の配置とのずれを測定

これを候補者数 $m = 5, 10$，投票者数 $n = 100, 1000$，投票人数 $k = 2, 3, \ldots, m$ について 1000 回の試行を行った．なお MDS の計算には統計解析環境 R の `isomds` 関数 (`pcurve` パッケージ) を使用した．

表 3, 4 が 1000 回の試行での r^2 の平均である．また，図 2, 3 に r^2 のボックスプロットを示す．図 2, 3 では，左が投票者数 $n = 100$，右が $n = 1000$ の場合の図である．

表 3: ずれ r^2 の平均，$m = 5$

$m = 5$	$k = 2$	3	4	5
$n = 100$	1.0778	0.75984	0.95755	1.3102
$n = 1000$	0.9976	0.68340	0.87052	1.1841

表 4: ずれ r^2 の平均，$m = 10$

$m = 10$	$k = 2$	3	4	5	6
$n = 100$	1.8987	1.1025	0.82607	0.74999	0.75911
$n = 1000$	1.5431	0.84429	0.63751	0.52661	0.5166
$m = 10$	$k = 7$	8	9	10	
$n = 100$	0.93647	1.2137	1.5561	1.9341	
$n = 1000$	0.61318	0.90346	1.3491	1.8274	

これらより，$m = 5, 10$ のいずれのケースでも，投票人数 k が増えるにつれ，いったん値が減少した後に再び増加する様子が見て取れる．すなわち k が多すぎても少なすぎても，もともとの候補者の配置の再現性が落ちてしまう．また，その傾向は値そのものだけでなく，値のばらつきについても言える．これは，k が小さい場合には投票データから十分な情報が得られないため，大きい場合には余計な情報を含んでしまうためではないかと考えられる．現実の投票では投票者にとっての負荷や集計の際の手間を考えると，k をあまり大きな数には出来ないことから，必ずしも k を大きくすればよいわけではないことが示されたことは意味がある．

投票者数 n による違いに着目すると，候補者数 $m = 5$ の場合には，$n = 100$ と $n = 1000$ とではそれほど大きな違いがないが，$m = 10$ の場合には，特にずれ r^2 のばらつきに多少の違いが見られる．これは，候補者数が 10 人の場合には，投票者数が 100 人では類似性の分析には不十分であることを示しているのではないだろうか．

4 おわりに

本稿では我々の提案した，ランクつき投票モデルのもとでの候補者の類似性評価手法について，シミュレーション実験により本来の候補者の配置の再現性について調査した．投票人数を多くしても必ずしも類似性の分析には有益ではないことがわかった．これは選挙実施者が投票人数を決定する際の参考となるであろう．

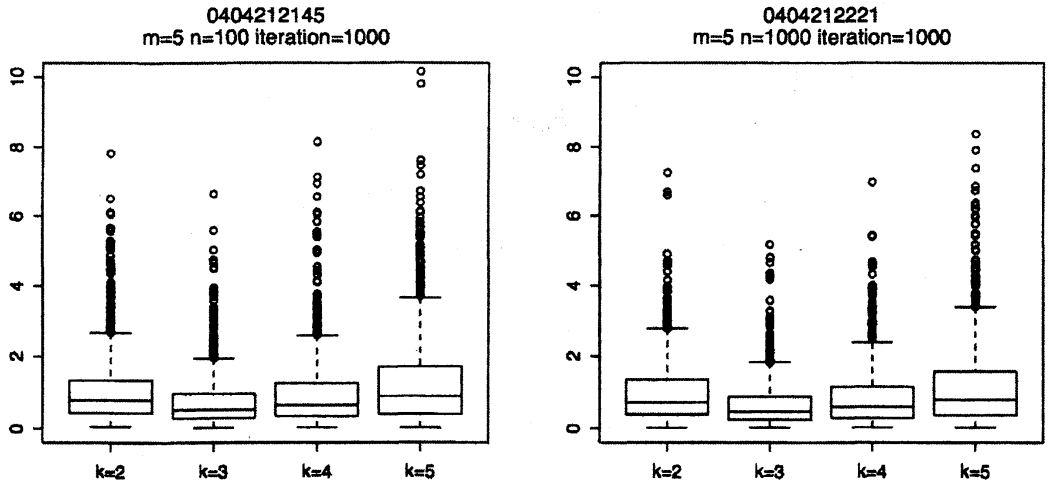

図 2: ずれ r^2 の box plot, $m = 5$

今回行った実験はあくまでも疑似的な人工データに対するものであり，実際の投票データに対して同様のことが言えるかどうかはわからない．そのためより現実に近い状況，すなわち，何らかのテーマで実際にランクつき投票を行ったデータを用いての検討が必要であろう．しかし，その場合は候補者の「真の」配置を知る術が無いため，投票と併せて投票者に類似性を評価してもらうなどの方法が必要となる．

現実の投票データの場合には，投票人数 k の影響についても注意が要る．疑似データでは，k が大きい場合でも各投票者は下位の候補に対する好感度の判断を正しく行うことが出来るが，実際の人間の判断では下位の候補に対するランク付けがおそらくいい加減なものになるであろう．したがって今回の実験以上に，下位の情報についての扱いを考慮しなければならない．

さらに，今回の疑似データと違い，現実の投票データでは，候補者の分布する空間の真の次元を知ることが出来ない．そのため，MDS 分析の際に得られるストレス値などを利用して，評価空間の次元を決定する必要がある．

我々の最終的な目的は単に候補者間の類似性を評価することではなく，適切な候補者の選択のために，得られた類似性を用いることにある．複数の候補が当選する選挙の場合，一緒に当選する候補同士が類似性の高い候補かどうかは大きな意味を持つであろう．一般に投票者の意思をより広く汲み上げるには類似性の低い候補が選ばれるほうがよいと考えられる．また逆に事業体の経営陣を選ぶようなケースではスムースな意思統一のために類似性の高いものが選ばれるほうがよいこともあるかもしれない．今後は今回提案した手法で得られた類似性の評価結果を，ランクつき投票モデルに対する候補者順位決定手法と組み合わせ，当選する候補の類似性の高低をコントロールする方法について検討したい．

参考文献

[1] J. Gill and J. Gainous, Why does voting get so complicated? A review of theories for analyzing democratic participation, Statistical Science 17 (2002), 383–404.

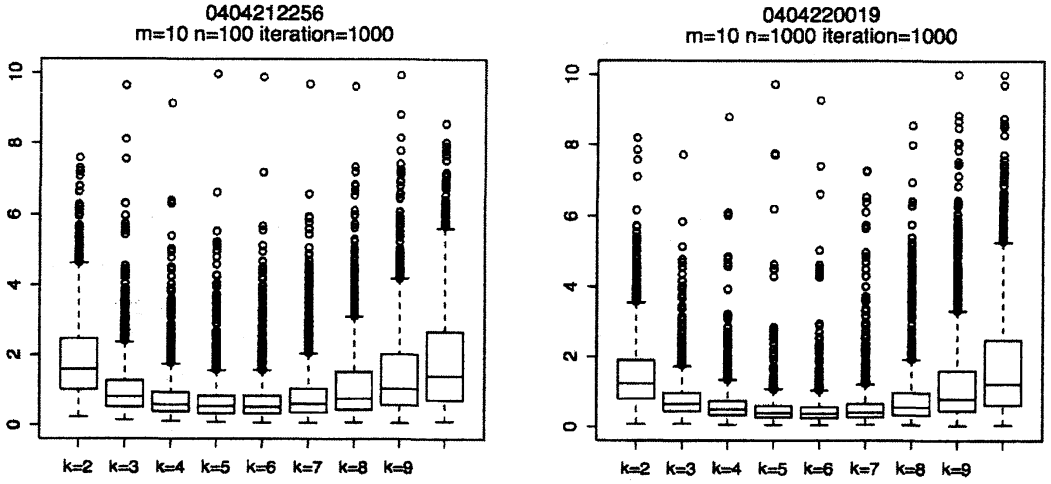

図 3: ずれ r^2 の box plot, $m = 10$

[2] R. H. Green, J. R. Doyle and W. D. Cook, Preference voting and project ranking using DEA and cross-evaluation, European Journal of Operational Research 90 (1996), 461–472.

[3] A. Hashimoto, A ranked voting system using a DEA/AR exclusion model: A note, European Journal of Operational Research 97 (1997), 600–604.

[4] J. B. Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, Psychometrika 29 (1964), 1–27.

[5] J. B. Kruskal, Nonmetric multidimensional scaling: A numerical method, Psychometrika 29 (1964), 115–129.

[6] T. Obata and H. Ishii, A method for discriminating efficient candidates with ranked voting data, European Journal of Operational Research 151 (2003), 233–237.

[7] T. Obata and H. Ishii, On the similarity evaluation of candidates in ranked voting model, in Proceedings of the Asia Pacific Management Conference, 2003, 707–714.

[8] 小畑経史・石井博昭, ランクつき投票モデルの多次元尺度法による類似度分析, 京都大学数理解析研究所講究録, 2004.

[9] 佐伯胖, 「きめ方」の論理, 東京大学出版会, 1980.

[10] 齋藤堯幸, 多次元尺度構成法, 朝倉書店, 1980.