ネットワークの中心性指標とレーティング手法について

2017SS099 横井将人
指導教員：塩濱敬之

1　はじめに

本研究では, 中心性指標, Massey のレーティングについて理解し, Massey のレーティング指標をネットワークの中心性指標とみたときに, その統計的な性質を調べることで, 中心性指標の仮説検定を可能にし, 統計的性質を調べるためにシミュレーションによる検証を行う. このような統計的な性質を得た後に, 都道府県間人口移動データの解析を行うことを目的としている.

2　Massey のレーティング

スポーツなどの試合結果に基づくレーティング手法として, Massey の手法が知られている [1]. Massey のレーティングでは, プレーヤー i と j によって行われた試合 k の得点差 y_k が, 両プレーヤーのレート r_i, r_j の差によって決まり, $y_k = r_i - r_j$ であることを仮定している. m プレーヤーによって n 試合行われた場合, 次式を得ることができる.

$$\boldsymbol{X}\boldsymbol{r} = \boldsymbol{y} + \boldsymbol{\varepsilon} \tag{1}$$

ここで, $\boldsymbol{X}$ は $n \times m$ の行列で, $X_{ki} = 1,\ X_{kj} = -1$ であり, その他の要素が 0 である.

また, レーティングベクトル $\boldsymbol{r}$ は正規方程式 $\hat{\boldsymbol{r}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ により推定することができる.

$\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{M},\ \boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{p}$ とすると, $\hat{\boldsymbol{r}} = \boldsymbol{M}^{-1}\boldsymbol{p}$ となる. 行列 $\boldsymbol{M}$ の対角成分は, 試合数に対応し, 非対角成分はどのチームと何回対戦したかを表している. またベクトル $\boldsymbol{p}$ は得失点差を表している. 実際には, $\boldsymbol{M}$ は正則行列ではないため上式の解は一意に定まらない. そこで, $\boldsymbol{M}$ の任意の 1 行を全て 1 にし, 対応する $\boldsymbol{p}$ の成分を 0 に修正したものを $\widetilde{\boldsymbol{M}}, \tilde{\boldsymbol{p}}$ とおけば, レーティングベクトルは, 次の正規方程式を解くことで得られる.

$$\hat{\boldsymbol{r}} = \widetilde{\boldsymbol{M}}^{-1}\tilde{\boldsymbol{p}} \tag{2}$$

3　レーティング差に関する統計的アプローチ

Massey のレーティング指標をネットワークの中心性指標とみたときに, その統計的な性質を調べることで, 中心性指標の仮説検定や信頼区間の構成が可能になる.

(2) 式における, ベクトル $\tilde{\boldsymbol{p}}$ は, 第 i 成分を 0 とおいたベクトルとする. そのとき, $\tilde{\boldsymbol{p}}$ は $\tilde{\boldsymbol{p}} = \tilde{\boldsymbol{X}}^T\boldsymbol{y}$ と表すことができる. ここで $\tilde{\boldsymbol{X}}$ はマッチング行列 $\boldsymbol{X}$ において, 第 i 列目のすべての成分を 0 とした $n \times m$ 行列である. したがって,

$$\hat{\boldsymbol{r}} = \widetilde{\boldsymbol{M}}^{-1}\tilde{\boldsymbol{p}} = \widetilde{\boldsymbol{M}}^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{y}$$

となる. 上式に (1) を代入すれば,

$$\hat{\boldsymbol{r}} = \widetilde{\boldsymbol{M}}^{-1}\tilde{\boldsymbol{X}}^T(\boldsymbol{X}\boldsymbol{r} - \boldsymbol{\varepsilon}) = \widetilde{\boldsymbol{M}}^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{X}\boldsymbol{r} - \widetilde{\boldsymbol{M}}^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{\varepsilon}$$

を得ることができる. 今, レーティング間には $\sum_{i=1}^{m}\hat{r}_i = 0$ の制約があるため, チーム数 m のレーティング計算における自由度は $m-1$ である. そのため, レーティングベクトルの標本分布は, チーム数 $m-1$ の下で求める必要がある. 行列 $\widetilde{\boldsymbol{M}}$ と $\tilde{\boldsymbol{X}}$ の行, 列の成分を 1 や 0 とおいた行, 列番目を i とし, 第 i 成分を除いたレーティングベクトルを

$$\hat{\boldsymbol{r}}_{-i} = (\hat{r}_1, \ldots, \hat{r}_{i-1}, \hat{r}_{i+1}, \ldots, \hat{r}_m)^T, \tag{3}$$

また, $\hat{r}_i = -\sum_{j=1, j\neq i}^{m}\hat{r}_j$ とすれば, (1) 式における誤差ベクトル $\boldsymbol{\varepsilon}$ に正規性と等分散性を仮定 ($\boldsymbol{\varepsilon}\ \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$) すると, $(m-1)$ チーム数の Massey のレーティングの標本分布は $\hat{\boldsymbol{r}}_{-i} \sim N(\boldsymbol{r}_{-i}, \sigma^2\boldsymbol{V}_{-i})$ となる. ここで, 平均ベクトル $\boldsymbol{r}_{-i}$ は (3) において, 推定値を真の値に置き換えたベクトルであり, 分散共分散行列 $\boldsymbol{V}_{-i}$ は

$$\boldsymbol{V}_{-i} = \left[\widetilde{\boldsymbol{M}}^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}(\widetilde{\boldsymbol{M}}^{-1})^T\right]_{j=1,\ldots,m,k=1,\ldots,m,(j,k)\neq i}$$

と表せる. また, $\hat{r}_i = -\sum_{j=1,j\neq i}^{m}\hat{r}_j$ の標本分布は

$$\hat{r}_i \sim N\left(\boldsymbol{r}_i, \sigma^2\boldsymbol{1}^T\boldsymbol{V}_{-i}\boldsymbol{1}\right)$$

となる. ここで $\boldsymbol{1} = (1, \ldots, 1)^T$ は, すべての成分が 1 のベクトルである. また, $\mathrm{Cov}(\hat{\boldsymbol{r}}_{-i}, \hat{r}_i) =: \sigma^2\boldsymbol{v}_{12}$ となる. これより, チーム数 m のレーティングベクトルの標本分布は, 次の正規分布に従う.

$$\hat{\boldsymbol{r}} = \begin{pmatrix}\hat{\boldsymbol{r}}_{-i}\\ \hat{r}_i\end{pmatrix} \sim N\left(\boldsymbol{r}, \sigma^2\begin{bmatrix}\boldsymbol{V}_{-i} & \boldsymbol{v}_{12}\\ \boldsymbol{v}_{12}^T & \boldsymbol{1}^T\boldsymbol{V}_{-i}\boldsymbol{1}\end{bmatrix}\right) = N(\boldsymbol{r}, \sigma^2\boldsymbol{\Sigma}_r)$$

レーティング指標の標本分布が得られたので, レーティング差の仮説検定を次のようにして実行することができる. 今, 第 i チームと第 j チームのレーティング差が 0 であるという帰無仮説の両側検定を考える. すなわち,

$$H_0 : r_i - r_j = 0, \quad H_1 : r_i - r_j \neq 0$$

の仮説検定を考える. i 番目の要素を 1, j 番目の要素を -1 とする m 次元ベクトルを $\boldsymbol{w}$ と定義すれば,

$$\hat{\boldsymbol{r}}^T\boldsymbol{w} \sim N(\boldsymbol{r}^T\boldsymbol{w}, \sigma^2\boldsymbol{w}^T\boldsymbol{\Sigma}_r\boldsymbol{w})$$

となることから, 検定統計量を以下のように定義すれば,

$$T = \frac{\hat{\boldsymbol{r}}^T\boldsymbol{w}}{SE(\hat{\boldsymbol{r}}^T\boldsymbol{w})}$$

ただし, $SE(\hat{\boldsymbol{r}}^T\boldsymbol{w}) = s\sqrt{\boldsymbol{w}^T\boldsymbol{\Sigma}_r\boldsymbol{w}}$ であり,

$$s^2 = \frac{1}{n-(m-1)}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{r}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{r}})$$

は, 誤差分散の不偏推定量である. 検定統計量 T は自由度 $n-(m-1)$ の t 分布に従う. これより, レーティング差の仮説検定が可能である.

4　シミュレーション

前章で得られた Massey のレーティング指標に関する統計的性質を調べるためにシミュレーションによる検証を行う．シミュレーションでは全 11 チーム ($m = 11$) からなる，総当たり戦 ($n = \binom{11}{2} = 55$) と，各チームが少なくとも一度は対戦するようにランダムに選択した $n = 25$ の場合を考える．各チームのレーティングベクトルを $\boldsymbol{r} = (r_1, r_2, r_3, \ldots, r_{11})^T$ とし，そのスコアを $\boldsymbol{r} = (5, 4, 3, 2, 1, 0, -1, -2, -3, -4, -5)^T$ とした．レート差による得失点差の誤差項には正規分布 $r_i - r_j \sim N(r_i - r_j, \sigma^2)$ を仮定した．ここで，誤差分散 σ^2 には $\sigma \in \{3, 5\}$ の 2 つの場合を検証した．各シミュレーションは 10,000 回繰り返し，シミュレーションによる Massey のレーティング指標の推定量の不偏性と，レーティング差の仮説検定における検出力を検証する．

図 1 は，$\sigma = 3, 5$ としたときの，有意水準 5% のレーティング差の両側検定における検出力を示している．図の曲線は，t 検定 (黒，赤) 及び，検出力関数を正規近似した曲線 (緑，青) を表し，図中の点は，10000 回の繰り返しにおいて，各レーティング差で帰無仮説が棄却された割合を示している．

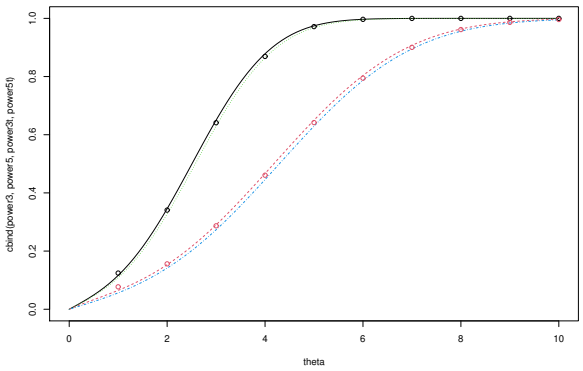

図 1　理論的な検出力関数 (曲線) とシミュレーションによる検出力 (点)

実際のレーティング計算において，各チームが全チームと少なくても 1 回以上試合をすることは現実的な仮定ではない．そこで，11 チームから 2 チームの選び方の全 55 通りの組み合わせから，ランダムに 25 試合を選択したときのシミュレーションによる検出力と理論的な検出力関数をプロットした．ただし，25 試合の選び方は，少なくとも各チームは 1 試合対戦し，その隣接行列は連結するように選択した．図 2 は，$\sigma = 3$ とした場合のプロットであり，図中の検出力関数は，マッチング行列 X に依存するため，25 組の一つの選び方から得られる検出力関数のバリエーションを示している．

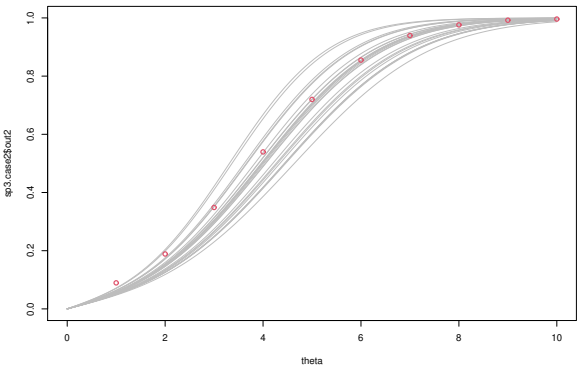

図 2　$n = 25$ のときの，理論的な検出力関数 (曲線) とシミュレーションによる検出力 (点)，$\sigma = 3$

5　都道府県間人口移動データの解析

日本では，進学や就職を機会とした，おもに若年層において，東京大阪への大都市圏への人口集中が進んだ．2008 年から総人口は減少しているにもかかわらず，大都市圏への人口集中のトレンドは継続している．人口移動の要因にはプッシュ要因とプル要因が知られていてプッシュ要因とは地域から人々が押し出される要因 (高い失業率など) プル要因とは，地域へ人々を引き寄せる要因 (高い賃金・所得等) を指す．本研究では，Massey のレーティングにおける攻撃力をプル指数，守備力をプッシュ指数として計算する．都道府県間人口移動データ解析をネットワークモデルから分析した研究はほどんど存在しない．2018 年と 2019 年の 2 カ年分の住民基本台帳人口移動報告から都道府県間人口移動データを用いた分析を行う．分析の結果，首都圏のレーティング値が他の都道府県に比べて高いことが分かり，人口減少社会にもかかわらず首都圏への人口集中が継続していることが分かった．首都圏の人口集中の要因は，プル指数よりもプッシュ指数が他都道府県に比べて著しく低いことが分かった．

6　おわりに

本研究では，Massey のレーティング指標をネットワークの中心性指標とみたときに，その統計的な性質を調べることで，Massey のレーティング指標の不偏性を確認し，レーティング差の仮説検定を可能にした．Massey のレーティング指標に関する統計的性質を調べるためにシミュレーションによる Massey のレーティング指標の推定量の不偏性と，レーティング差の仮説検定における検出力を確認した．都道府県間人口移動データの解析を行った結果，首都圏の人口集中について明らかになった．都道府県間についてデータ解析を行ったが，国同士についても行ってみたい．

参考文献

[1] Massey, K.: Statistical models applied to the rating of sports teams. Bluefield College, 1997.