



RJMCMCアルゴリズムを用いた 多次元一対比較データの解析法

宇佐美慧

(東京大学教育学研究科/日本学術振興会)

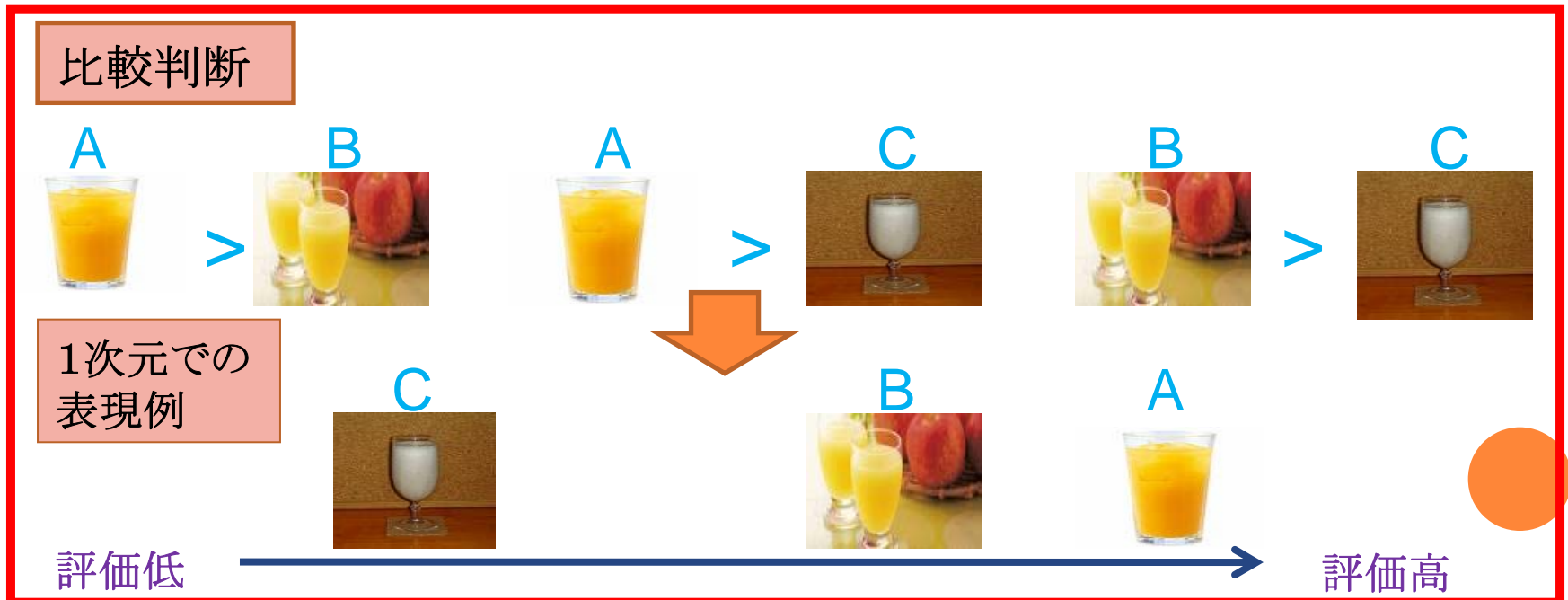
アウトライン

- 一対比較モデルの紹介と既存のモデルの問題点
- 個人差を考慮した多次元一対比較モデルの紹介
- 母数の推定手順 (**RJMCMC** アルゴリズム) の導出
- シミュレーションデータに基づくアルゴリズムの検証および、官能検査とスポーツデータを用いた分析例の紹介
- 総合考察
- **Appendix:** 推定の詳細及び、更なる応用モデルの紹介



一対比較法について

- **一対比較法**...2つの対刺激(対象・項目・選択肢)に対して何らかの観点に基づき比較判断を行う。その結果から、ある次元上に刺激を配置して、刺激間の特性(尺度値と呼ばれる)の評価を行う。
- **官能検査の例**: A・B・Cの3つのジュースの各対($3C_2=3$ 通り)において、美味しいと思ったジュースを選択させる。



一対比較法の利点と応用

精度と情報量の多さ

- 単一の刺激に対して評定するよりも、刺激間に存在する細かな差異を明瞭に検出できる。作業量に比して得られる情報量も多く、実施も容易 (e.g., Maydeu-Olivares & Bockenholt, 2005).

応用例 (心理学・行動科学・情報科学・工学)

- 心理学データ (表情認知, 図形認識, 意志決定)
- マーケティング (銘柄決定, 市場調査)
- ネットワークモデル (情報通信)
- 官能評価データ (飲料水, 嗜好品)
- スポーツデータについても、個人やチーム間の対戦結果をデータと見なして一対比較法を適用する例も多い. (e.g., Koehler & Ridpath, 1982 ; 竹内 & 藤野, 1988)

一対比較モデルの理論的拡張

- ・心理学・精神物理学におけるFechner(1860),Thurstone(1927)の研究を嚆矢として様々な理論的拡張を遂げており, 今日までに多くの応用モデルが提案されている.

拡張例

- ・比較判断に影響する諸要因(刺激の提示順序効果など)を考慮したモデリング(e.g., Davidson & Beaver, 1977; 宇佐美, 2009)
- ・評価の個人差を考慮したモデリング(e.g., 野澤・飯田他, 1995)
- ・多次元一対比較モデル(e.g., Hunter, 2004; Causeur & Husson, 2005)

他の統計モデルとの融合

- ・項目反応モデル(IRT)や構造方程式モデリング(SEM)など, 潜在変数モデルの文脈での発展(e.g., Maydeu-Olivares & Bockenholt, 2005)

→既存のソフトウェアを用いた分析アプローチも有効に.



既存の一对比較モデルの問題点

1, 多次元一对比較モデルにおける, 個人差の影響(および比較判断に影響する諸要因)を考慮した方法論の未開発.

2, 多次元一对比較モデルにおける, 最適な次元数を確率的にかつ短時間で決定する方法の未検討.

- 1, のモデリングそのものは容易であるが尤度関数が複雑.
- 2, においては異なる次元数のモデルを別々に分析する方法があるが, 確率的な判断はできず, しかも実行に時間を要する(モデルの次元数の決定は多くの分野が抱える困難な問題).



近年理論的な発展が盛んである, ベイズ法に基づくRJMC MC (Reversible Jump Markov Chain Monte Carlo)のような, サンプルングに基づく推定法が有用.

RJMCMCアルゴリズム

- MCMCを拡張したアルゴリズムで、(次元数の)異なるモデル間を超えてサンプリングし、統計的に最も整合的なモデルを確率的に評価する(Green, 1995 ; Richardson, Green, 1997).
- 身近な例として、因子分析における最適な因子数を決める問題を扱ったケースがある(Lopes & West, 2004).
- ベイズ法の枠組であるため、母数の事前情報を考慮することも可能であるが、ベイズ法を一対比較データに適用した例はほとんどない。



そこで本研究では、一対比較法の代表モデルである Bradley-Terry Model (Bradley & Terry, 1952)を拡張して

- (1) 多次元一対比較モデルにおいて個人差の影響を考慮し、
- (2) 次元数を推定する為にRJMCMCアルゴリズムを用いた、

応用的な多次元一対比較モデルの理論的検討を試みる。

モデル

モデル式

L 次元モデルにおける刺激 i と刺激 j の一対比較において、個人 h が刺激 i を選択する確率.

$$P_{hij/L} = \frac{\sum_{l=1}^L k_{hl}\theta_{il}}{\sum_{l=1}^L k_{hl}\theta_{il} + \sum_{l=1}^L k_{hl}\theta_{jl}}$$

i と j に関する
選好度合いの
比になる!

L ... モデルの次元数(未知).

k_{hl} ... 個人 h の $l(1, \dots, l, \dots, L)$ 番目の次元に対するウェイト.

θ_{il} ... 刺激 i の第 l 次元目における尺度値.

- θ_{il} が高いほど、刺激 i が次元 l において選好度合いが高いことを意味する.
- k_{hl} が高いほど、個人 h が判断の際に次元 l に関わる観点を重要視していることを意味する.

*判断に影響する諸要因を考慮した、より応用的なモデル構築も可能(Appendix参照.)

提案モデル (CONT.)

注意点

- k_{hl} は必ずしも個人差を意味する母数ではなく、あくまで全体のデータセットを層化すること (e.g., 個人毎に見ること) で、分析結果の解釈を補助する為の母数である。
- そのため, k_{hl} は例えば市場調査における地域差 (e.g., 地域によって製品に対し異なるニーズがある) や, スポーツデータにおける時期の差を表す場合 (e.g., 同じ選手でも時期によって能力に違いがある) でも利用可能である。

母数の制約

$$\sum_{i=1}^I \theta_{il} = S \quad (\forall l) \quad \sum_{l=1}^L k_{hl} = 1 \quad (\forall h) \quad \text{cov}(\theta_{il}, \theta_{il'}) = 0 \quad (\forall l, l')$$

S ... 各次元における尺度値の総和

I ... 刺激数

推定法

以下では、多次元一対比較モデルにおいて、次元数 L を1から L^+ の間の中で決定する問題を考える。

$L (L \in L^+)$ 次元モデルにおける尤度関数

$$L(K_L, \theta_L / Z, L) = \prod_h \prod_i \prod_j P_{hij/L}^{n_{hij}} (1 - P_{hij/L})^{N_{hij} - n_{hij}}$$

Z ... 得られたデータセット

K_L, θ_L ... それぞれサイズ $H \times L, I \times L$ の, K_{hl}, θ_{ik} に関する全ての要素を含んだ母数ベクトル

n_{hij} ... 刺激 i と刺激 j の一対比較において、個人 h が刺激 i を選択した総数. $N_{hij} = n_{hij} + n_{jih}$ が成り立つ.

事前分布

$$P\left(\frac{\theta_{il}}{S}\right) = \text{beta}(\alpha_0, \beta_0) \quad P(K_{hL}) = \text{dirichlet}(\alpha_{10}, \alpha_{20}, \dots, \alpha_{L0}) \quad P(L) = 1/L^+$$

因子数に関する事前分布を考える!

ここで $K_{hL} = (k_{h1}, k_{h2}, \dots, k_{hL})$ で, $\alpha_0, \beta_0, \alpha_{10}, \alpha_{20}, \dots, \alpha_{L0}$ は事前に設定する超母数.

推定法(CONT.)

L次元モデルにおける完全条件付事後分布

$$\begin{aligned} p(\boldsymbol{\theta}_L/\mathbf{K}_L, \mathbf{Z}, L) &= p(\mathbf{K}_L, \boldsymbol{\theta}_L/\mathbf{Z}, L)/p(\mathbf{K}_L) \propto L(\mathbf{K}_L, \boldsymbol{\theta}_L/\mathbf{Z}, L)p(\boldsymbol{\theta}_L) \\ p(\mathbf{K}_L/\boldsymbol{\theta}_L, \mathbf{Z}, L) &= p(\mathbf{K}_L, \boldsymbol{\theta}_L/\mathbf{Z}, L)/p(\boldsymbol{\theta}_L) \propto L(\mathbf{K}_L, \boldsymbol{\theta}_L/\mathbf{Z}, L)p(\mathbf{K}_L) \end{aligned} \quad (1)$$

ただし, $p(\boldsymbol{\theta}_L) = \prod_l \prod_i p(\theta_{il})$ $p(\mathbf{K}_L) = \prod_h p(\mathbf{k}_h)$

RJMCMCアルゴリズム (数学的な詳細はAppendix参照).

- 1、MCMCを用いて各 L ($L \in \mathcal{L}$) 次元のモデル内の母数を推定する.
- 2、特定の次元数 L' のモデルを選択するか否か決定する.
- 3、2、で選択されれば次元数 L' のモデルから、棄却されればもとの次元数のモデルから、MCMCを用いてサンプルを得る.
- 4、2, 3を母数が収束するまで繰り返し, 次元数に対する事後分布 $\hat{p}(L/\mathbf{Z})$ と, 各次元数のモデル内の母数 $\mathbf{K}_L, \boldsymbol{\theta}_L$ の推定値を得る.

(1)式の完全条件付事後分布は上記全ての段階で用いる.

シミュレーション

- アルゴリズムの妥当性を評価する上で、任意の条件での母数の推定精度を評価する方法が取られることが多い。
- RJMCMCアルゴリズムの場合、(1)真の次元数を正しく特定できるか(すなわち $\hat{p}(L/Z)$ が正しい L に対して高い事後確率を示しているか)という問題と、(2)各 L 次元モデルにおける母数 K_L, θ の推定精度に関する問題、の2点を検証する必要がある。

→シミュレーション①では(1)の問題を、シミュレーション②では(2)の問題を、それぞれ検証する。



シミュレーション① 真の次元数の推定精度

- 真の次元数を2として, 2次元モデルに基づいて母数の真値を定め, 一様乱数を用いてデータセット Z を作成.
- 項目数 = $I = 5, 10$, 被験者数 = $H = 5, 10, 20, 50, 100, 200$, 測定回数 = $N_{hij} = N = 1, 5, 10$ のもとで, 1次元モデルから3次元モデルにわたって $P^{\wedge}(L/Z)$ を得た.
- そして, $P^{\wedge}(L/Z)$ が $L = 2$ において高い事後確率を示しているか否かを検討する.

事前分布 ($I=5$ および 尺度値の総和 $S=100$ の場合)

$$\theta_{il}/100 \sim \text{Beta}(1, 4)$$

$$K_h \sim \text{dirichlet}(2, \dots, 2)$$

$$p(L) = 1/3$$

* 詳細は **Appendix** 参照.



シミュレーション① 結果

50回のシミュレーションにおける各次元の事後確率の平均(標準偏差).

I=10条件		L=1	L=2	L=3
N=1	H=5	0.0000(0.0000)	0.9040(0.2526)	0.0959(0.2526)
	H=10	0.0000(0.0001)	0.9450(0.1084)	0.0549(0.0043)
	H=20	0.0012(0.0033)	0.9954(0.0103)	0.0034(0.0101)
	H=50	0.0038(0.0073)	0.9956(0.0075)	0.0006(0.0015)
	H=100	0.0034(0.0078)	0.9952(0.0068)	0.0014(0.0023)
	H=200	0.0039(0.0093)	0.9924(0.0034)	0.0037(0.0043)
N=5	H=5	0.0008(0.0019)	0.9991(0.0019)	0.0000(0.0001)
	H=10	0.0135(0.0029)	0.9986(0.0029)	0.0000(0.0000)
	H=20	0.0001(0.0004)	0.9999(0.0003)	0.0000(0.0000)
	H=50	0.0000(0.0000)	0.9987(0.0037)	0.0012(0.0037)
	H=100	0.0000(0.0000)	0.9852(0.0186)	0.0148(0.0186)
	H=200	0.0000(0.0000)	0.9950(0.0055)	0.0050(0.0055)

- 一定の項目数や被験者数を確保すれば $L=2$ を採択する確率が最も高くなる。
- しかし、 I が小さい場合は、各項目が互いに因子を形成できず独立したために、結果として $L=3$ を支持する場合も見られた。

シミュレーション②

各モデル内の母数の推定精度

- 各モデル内でのMCMCに基づく母数の推定精度を検証するため、1次元モデルから3次元モデルにおける各モデル内の母数 K_L, θ_L の推定精度を検証する。
- 項目数, 被験者数, 測定回数, および事前分布はシミュレーション①と同様の条件の下で, 各モデル内で母数の真値を定めてMCMCシミュレーションを行い, 母数の復元精度について検討した。
- 紙面の都合から, ここでは2次元モデルにおける推定精度についての結果を報告する。



シミュレーション② 結果

- 50回のシミュレーションにおける各母数のBiasと、平均二乗誤差(RMSE)の平均を計算し、母数ベクトルごとにまとめた結果(L=2の場合)。

L=2		N=1				N=5			
		θ		K		θ		K	
		bias	RMSE	bias	RMSE	bias	RMSE	bias	RMSE
I=5	H=5	2.5972	5.5228	0.2220	0.3415	1.8394	5.1170	0.1515	0.2852
	H=10	2.5353	6.2627	0.1972	0.2936	2.0601	4.8397	0.1318	0.2801
	H=20	2.1180	5.2138	0.1999	0.3458	1.1891	3.4573	0.1299	0.2811
	H=50	1.6282	4.2305	0.1809	0.3359	0.6184	2.4873	0.1209	0.2763
	H=100	1.3304	3.4042	0.1471	0.3369	0.4546	1.7647	0.1221	0.2797
	H=200	1.0242	2.7200	0.1374	0.3136	0.4396	1.2988	0.1126	0.2721
I=10	H=5	1.3739	4.1557	0.1596	0.3215	1.2820	3.1260	0.0732	0.2632
	H=10	1.1733	3.6519	0.1537	0.3548	0.8568	2.3790	0.0926	0.2648
	H=20	0.9186	3.0785	0.1421	0.3242	0.3684	1.3594	0.0983	0.2176
	H=50	0.7014	2.0281	0.1481	0.3088	0.1987	0.9035	0.0812	0.2167
	H=100	0.5181	1.5584	0.1409	0.3037	0.1211	0.5965	0.0693	0.2146
	H=200	0.3414	1.1664	0.1265	0.2996	0.1321	0.4503	0.0629	0.2068

*表の「Bias」は各母数におけるbiasの推定値に関する絶対値の平均。

*表の「RMSE」は各母数におけるRMSEの平均値。

- MCMCに基づく推定法はおおむね良好であることがわかる。



分析例1 -官能検査の場合-

- 6種類のカルピス(ノーマル・ダイエット・バナナ・イチゴ・巨峰・りんご酢)の官能評価を通して,カルピスの順位付けと個人差の評価を同時に行う.
- 12名の被験者が,6種類のカルピスの全ての可能な組み合わせによる $6C_2=15$ 対のカルピスにおいて,それぞれ美味しいと思う方のカルピスを選択した.
- RJMCMCアルゴリズムを適用し,1次元モデルから4次元モデルのうちいずれが最も適合が良いかを検討した.



- $L=1,2,3,4$ における事後確率は以下のようになり, $L=2$ が最も高かった.

L=1	L=2	L=3	L=4
0.1853	0.7158	0.0634	0.0355



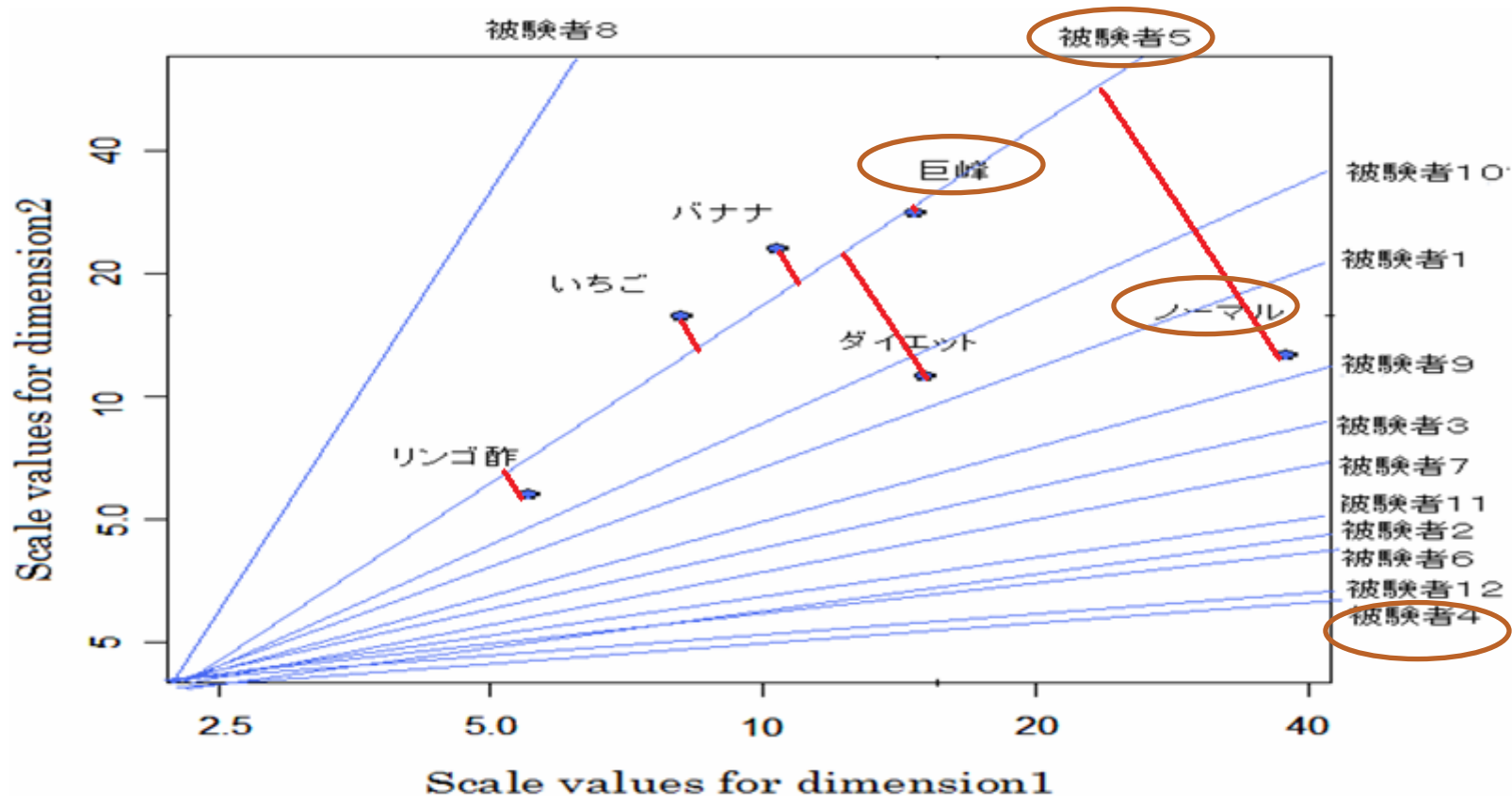
結果 (Kは一部のみ)

θ	l=1	l=2	K	l=1	l=2
ノーマル	42.63	14.16	被験者1	0.59	0.41
ダイエット	16.17	12.29	被験者2	0.81	0.19
バナナ	10.67	21.21	被験者3	0.72	0.28
イチゴ	8.59	18.81	被験者4	0.89	0.11
巨峰	16.24	28.46	被験者5	0.39	0.61
リンゴ酢	5.69	5.07	被験者6	0.84	0.16

- θ を見ると、 $l=1$ ではノーマルが最も好まれており、被験者に親しみのあると思われる種類のカルピスの推定値が高くなっている。 $\neq 2$ では巨峰やバナナなど、果物の味の強いカルピスにおいて値が高くなっている。
- K を見ると、多くの被験者が次元1の軸に基づいて味を評価しているが、たとえば被験者5は果物の味への志向性が高く、次元2の軸も重視して評価をしていたことがわかる。



母数の図式表現



*Kに関する直線は、次元1と次元2の重みづけの比を角度とした正接（タンジェント）の値に基づいて評価することで図示できる。



分析例2 -スポーツデータの場合-

- 相撲では、各力士の対戦結果(勝ちor負け)を一对比較データとしてみなすことができる.
- スポーツでは、一般にプレイヤーである力士(刺激)の能力(尺度値)が時系列的に変化していくため、 k_{hl} をこれまでのように個人差ではなく、例えば各年の力士の能力の変化を特徴づける母数として表わすことができる.
- 現在の横綱や大関などの、主要な力士10名に焦点を当てて、2005年から2009年までの対戦データの記録を収集し、データにRJCMCアルゴリズムを適用した.
- $L=1,2,3,4$ における事後確率は以下のようになり、 $L=2$ が最も高かった.

L=1	L=2	L=3	L=4
0.0449	0.9375	0.0111	0.0065



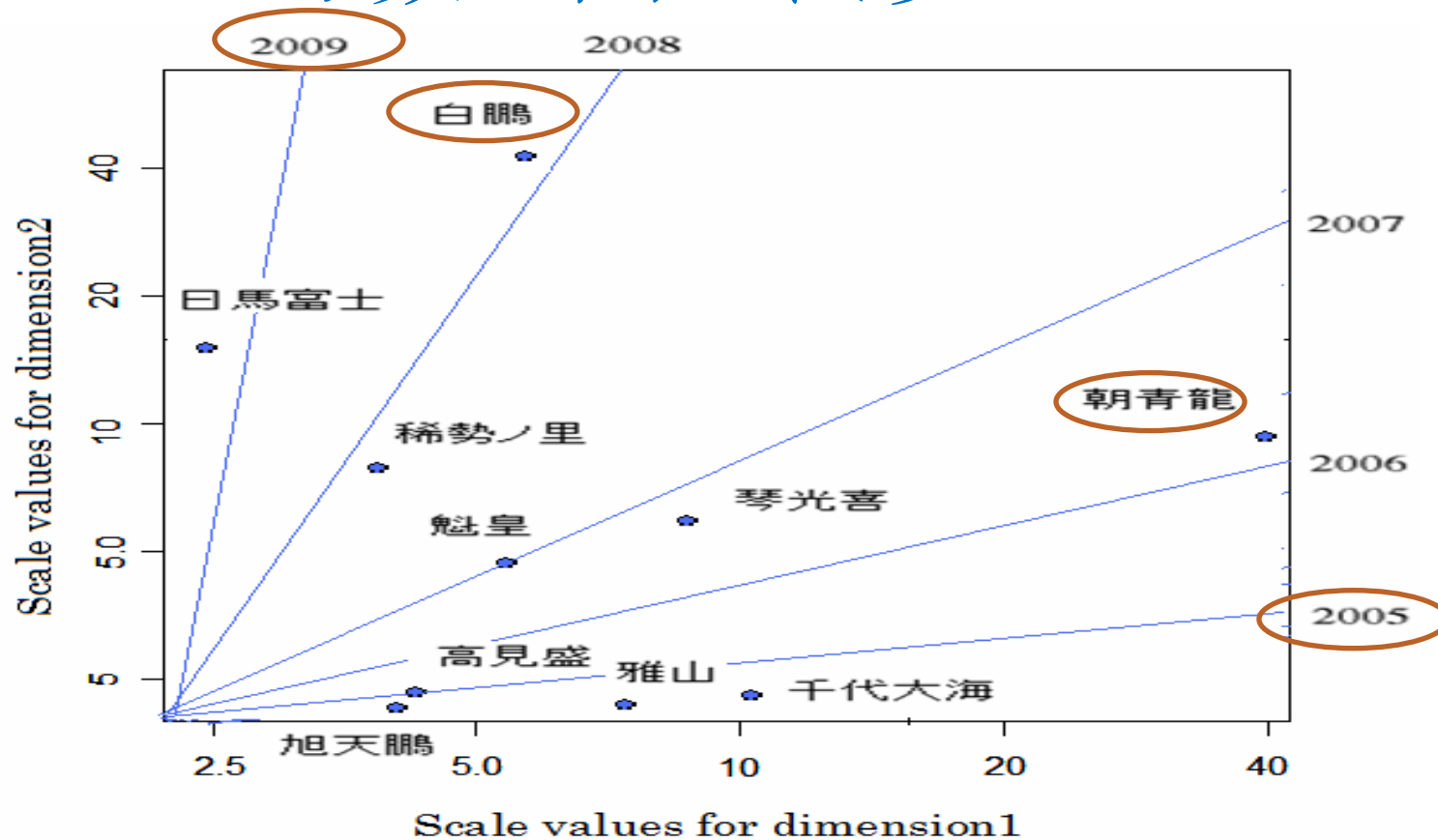
結果

		l=1	l=2
θ	朝青龍	43.82	11.39
	白鵬	6.86	40.11
	魁皇	6.84	5.32
	千代大海	10.72	3.14
	琴光喜	8.85	6.36
	高見盛	4.59	2.66
	日馬富士	2.24	16.11
	稀勢の里	3.99	8.89
	雅山	7.61	2.56
	旭天鵬	4.47	3.45
	K	2005	0.88
2006		0.73	0.27
2007		0.58	0.42
2008		0.19	0.81
2009		0.09	0.91

・Kの値は年度ごとに徐々に変化しており、l=1の推定値は旧時代(2005年度)の推定値に、逆にl=2の推定値は新時代(2009年度)の推定値に最も近い。

・θ (力士)に注目すると、例えば朝青龍は2005年度は全場所優勝をしていることからl=1において推定値が最も高い。しかし近年では推定値が低く見積もられ、白鵬の方が高い。このように年度間の力士の能力変化を追うことができる。

母数の図式表現



- 第1次元は(横軸)旧勢力で, 第2次元(縦軸)は新勢力.



総合考察

- 一対比較モデルは、心理学や行動科学をはじめ、マーケティングや情報科学、スポーツデータなど広く応用がなされている。
- 既存の多次元一対比較モデルでは、個人差を考慮するモデリングはなく、したとしても母数の評価は複雑であった。さらに次元数を確率的に決定する手立てがなかったという問題があった。
- 提案したモデルに基づいてRJMCMCアルゴリズムを利用することで、上記のいずれの問題をも解決することができる。しかもベイズ法を基礎と置いているので、分析者の持つ母数の事前情報を分析に反映させることも可能。



総合考察 (CONT.)

- 次元数や潜在クラス数の決定は多くの分野にとって頭の抱える問題であるが、RJMCMCアルゴリズムはS-PLUSなどのソフトウェアを用いた実装も容易であり、このようなサンプリング手法の応用可能性が期待できる。
- 異なる次元数のモデルに対してMCMCを別々に適用する場合に比べて、推定時間も短く、また次元数の決定に対して確率的な指針を与えてくれる。

今後の課題

- **欠損値の問題...**多くの要因を同時に扱う実験デザインの場合や、スポーツデータの適用の際は、データに無視できない欠損が存在する可能性が高い。これらの欠損メカニズムを考慮した推定手法の検討が必要。
- **多値型データへの拡張...**Bradley-Terry modelは2値データのための手法という制約がある。しかし多値データにおいては、一般化線形モデルなどの枠組みで扱うことが可能であり、拡張は容易であろう。

引用文献

- Augustin, T. (2005). An approach to combine the logistic threshold model of psychophysics with the Bradley-Terry-Luce models of choice theory. Journal of Mathematical Psychology, 49(1), 70-79.
- Bradley, R.A., & Terry, M.E. (1952). The rank analysis of incomplete block designs. : The method of paired comparisons. Biometrika, 39, 324-345.
- Causeur, D., & Husson, F. (2005). A 2-dimensional extension of the Bradley-Terry model for paired comparisons. Journal of Statistical Planning and Inference, 135(2), 245-259.
- Davidson, R.R., & Beaver, R.J. (1977). On extending the Bradley-Terry model to incorporate within-pair order effects. Biometrics, 33, 693-702.
- Fechner, G.T. (1860). *Elemente der Psychophysik*. Leipzig: Breitkopf und Hartel.Reprinted. Bristol : Thoemmes Press, 1999.
- Fruhworth-Schnatter, S. (2001). Markov Chain Monte Carlo estimation of classical and dynamic switching and mixture models. Journal of the American Statistical Association, 96, 194-209.
- Green, P. J., (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. Biometrika, 82(4), 711-732.



- Koehler, K.J., & Ridpath, H. (1982). An application of a biased version of the Bradley-Terry-Luce model to professional basketball results. Journal of Mathematical Psychology, 25, 187-205.
- Lopes, H.F. & West, M. (2004). Bayesian Model Assessment In Factor Analysis. Staistica Sinica, 14, 41-67.
- Maydeu-Olivares, A., & Bockenholt, U. (2005). Structural equation modeling of paired-comparison and ranking data. Psychological Methods, 10(3), 285-304.
- Richardson, S., & Green, P.J., (1997). On Bayesian analysis of mixtures with an unknown number of components. Journal of the Royal Statistical Society, B59, 731-792.
- 竹内啓・藤野和健 (1988). スポーツの数理科学. -もっと楽しむための数字の見方-. 共立出版.
- Thurstone, L., (1927). A law of comparative judgement. Psychological Review, 34, 273-286.
- 宇佐美慧. (2009). 比較判断に影響する複数の要因を考慮した一対比較データ分析法. 心理学研究, 79(6), 536-541.



Appendix

- 1、シミュレーション及び分析例における初期値と反復数
- 2、比較判断に影響する要因を考慮した多次元一対比較モデル
- 3、RJCMCMCアルゴリズムの詳細
- 4、母数の収束診断の問題と識別性の問題
- 5、分析アルゴリズム
- 6、ローデータ(相撲)



シミュレーション及び分析例における初期値と反復数

- K_L, θ_L の初期値については、それぞれ平均値を用いた。すなわち、

$$k_{hl} = 1/L, \quad \theta_{il} = S/I$$

- シミュレーション①・②, 分析例1, 2におけるBurn-inおよび全体の反復数は以下のとおりである。

	Burn-in	全体
シミュレーション①	40000	70000
シミュレーション②	30000	60000
分析例1	20000	60000
分析例2	50000	100000

*ただしシミュレーション②についてはMCMCアルゴリズムの、他のものについてはRJCMCアルゴリズムについての値である。



比較判断に影響する要因を考慮した 多次元一対比較モデル

- 複雑なモデリングに対して柔軟であるベイズ法を用いれば、本研究で検討したモデルを更に拡張することは容易である。
- 例えば、次元 l において刺激 i と j に総合的にそれぞれ γ_{lr} , $\gamma_{lr'}$ という形で複数の要因が影響を与える場合、提案モデルの式はより一般に γ_{lr} , $\gamma_{lr'}$ を含んだ以下の式で表わすことができる。

モデル式

$$P_{hijr'r'/L} = \frac{\sum_{l=1}^L \gamma_{lr} k_{hl} \theta_{il}}{\sum_{l=1}^L \gamma_{lr} k_{hl} \theta_{il} + \sum_{l=1}^L \gamma_{lr'} k_{hl} \theta_{jl}}$$

これまでとは
違い γ という
母数を含む！

$\gamma_{lr'}$ については例えば、提示順序効果やスポーツデータにおけるホーム・アウェイの効果、官能検査における刺激の濃度の違いなどが挙げられる。



RJMCMCアルゴリズムの詳細

• STEP0 -RJMCMC適用前の準備-

各 L ($L \in L^+$)次元モデルにおける母数の推定値を, MCMCアルゴリズムに基づいて得る.

• STEP1 -RJMCMCにおける初期値の設定-

次元数 L に関する初期値を任意に設定し, そこでの母数 K_L, θ_L の初期値を, STEP0の段階で得られた事後分布 $P(K_L, \theta_L / Z, L)$ を用いて設定する.

• STEP2 -モデル間移動ステップ-

•2.a 提案分布 $J(L \rightarrow L')$ を用いて, 以下の更新に対応する次元数 L' を決定する.

•2.b L' 次元モデルの母数全体に関する提案分布 $q_{L'}(\Theta_{L'}) = q_{L'}(K_{L'})q_{L'}(\theta_{L'})$ を用いて, 更新の候補値ベクトル $K_{L'}, \theta_{L'}$ がサンプリングされる(STEP0の解を利用).

•2.c 候補値を以下の確率で採択する. 棄却された場合には, 次元数は L を設定する.

$$\min \left[\frac{L(K_{L'}, \theta_{L'} / Z, L') p(K_{L'}, \theta_{L'} / L') p(L') J(L' \rightarrow L) q_L(\theta_L) q_L(K_L)}{L(K_L, \theta_L / Z, L) p(K_L, \theta_L / L) p(L) J(L \rightarrow L') q_{L'}(\theta_{L'}) q_{L'}(K_{L'})}, 1 \right]$$

ここで, $p(K_M, \theta_M / M)$ 次元数 M における母数全体の事前確率であり, $p(L)$ 次元数に関する事前分布である.



RJMCMCアルゴリズム (CONT.)

$p(L)$ および $J(L \rightarrow L')$ については、離散一様分布を用いている。すなわち、

$$p(L) = \frac{1}{L^+} \quad J(L \rightarrow L') = \frac{1}{L^+} \mathbf{1}_{L^+} \mathbf{1}'_{L^+}$$

また、 $p(K_M, \theta_M / M)$ については各母数に関する事前分布を掛け合わせた値となる。

・STEP3-モデル内移動ステップ-

STEP2で決定した次元数に基づいてMCMCを実行し、サンプルを得る。

・STEP4 -反復-

STEP2,3 推定値が収束するまで繰り返す。次元数の推定を以下のように行う。

$$\hat{P}(L/Z) = \frac{1}{T} \sum_{t=1}^T \mathbf{1}_L(L_t) \quad \mathbf{1}_L(L_t)=1 \text{ if } L = L_t \text{ and zero otherwise}$$

・このように、**STEP2**において異なる次元数のモデル間を行き来するところに、**RJMCMC**アルゴリズムの特徴がある(アルゴリズムについてのより詳細な部分や変更可能な部分についてはGreen, 1995 ; Lopes & West, 2004 を参照のこと)。

収束診断と識別性の問題について

収束診断の問題

- RJMCMCには手軽で実用的な収束診断方法が確立されていない (Castelloe & Zimmerman, 2002).
- 簡便な方法としては, MCMCにおける一般的な方法(複数の連鎖を用いる方法や, 単一の長めの連鎖を用いる方法)を利用することがあるが, この場合(特に事後確率の低い)各次元内の母数について解釈することが困難になる場合がある.
- このような制約を考慮しつつ(低い事後確率を示す次元における母数についての過度の解釈は控え), 本研究では簡便な方法を適用した.

Label-switchingの問題

- 異なる次元間の母数は互いに識別性がないために, 繰り返し計算の途中で各次元の母数が入れ替わってしまう現象.
- 本研究では推定の際に以下の制約をくわえ, 条件を満たさなかった場合, 各次元の母数を入れ替えて条件を満たすように調整する方法を採用した(e.g., Fruhwirth-Schnatter, 2001).

$$\theta_{ll} \leq \theta_{l'l'} \quad (l \leq l', l, l' \leq L)$$

分析アルゴリズム(一部)

#Program for analyzing paired comparison data by using RJMCMC (in case comparison of L=2 dimensional model and L=1 dimensional model)

Write the number of stimuli(I), respondents(or years, H), pairs for sumo data "Wsumo, which includes $n_{\{hij\}}$ "

```
I<-10
H<-5
C<-I*(I-1)/2
```

#Expand data "Wsumo" by "Data" for convenience of calculating likelihood

```
Data<-matrix(rep(0,H*C),H,C)
for(h in 1:H){
  for(i in 2:I){
    for(j in 1:(i-1)){
      Data[h,(10*i+j)]<-Wsumo[h,((i-1)*(i-2)/2+j)]
    }
  }
}
```

#Expand data "Nsumo (which includes $N_{\{hij\}}$)" by "NNN" for convenience of calculating likelihood

```
NNN<-matrix(rep(0,H*C),H,C)
for(h in 1:H){
  for(i in 2:I){
    for(j in 1:(i-1)){
      NNN[h,(10*i+j)]<-Nsumo[h,((i-1)*(i-2)/2+j)]
    }
  }
}
```

#Calculate log-likelihood (in case for L=2 dimensional paired comparison model : k_{21} , k_{22} , θ_{21} , θ_{22} are temporal samples associated with K and theta)

```
Loglikelihood2<-0
for(i in 2:I){
  for(j in 1:(i-1)){
    cc<-k21
    dd<-k22
    aa<- cc*theta21[i]+dd*theta22[i]
    bb<- cc*theta21[j]+dd*theta22[j]
    ee<-aa/(aa+bb)
    ff<-1-ee
    g<- Data[, (10*i+j)]
    Loglikelihood2<-Loglikelihood2+sum(log(ee^g))+sum(log(ff^(NNN[, (10*i+j)]-g)))
  }
}
```



分析アルゴリズム (CONT.)

```
#Define hyper-parameter for priors of theta, and calculate density (in case l=10)
```

```
alpha0<-1
beta0<-9
pt21<-0
for(i in 1:l){
  pt21<-pt21+log(dbeta(theta21[i]/100,alpha0,beta0))
}
pt22<-0
for(i in 1:l){
  pt22<-pt22+log(dbeta(theta22[i]/100,alpha0,beta0))
}
```

```
#Define hyper-parameter for priors of K, and calculate density
```

```
alpha10<-0.5
alpha20<-0.5
pk2<-0
for(i in 1:H){
  pk2<-pk2+log((gamma(alpha10+alpha20))/(gamma(alpha10)*gamma(alpha20))*k21[i]^(alpha10-1)*k22[i]^(alpha20-1))
}
```

```
#Define hyper-parameter for priors of L, and calculate density
```

```
pl2<-1/3
```

```
#Evaluate proposal distribution using values (mtheta21,mtheta22,,varmk22) evaluated from step0.("mdnorm" is original function used to evaluate density of multivariate normal distribution)
```

```
jt21<-log(mdnorm(theta21,mtheta21,varmtheta21))
jt22<-log(mdnorm(theta22,mtheta22,varmtheta22))
jk21<-log(mdnorm(k21,mk21,varmk21))
jk22<-log(mdnorm(k22,mk22,varmk22))
JJ2<-jt21+jt22+jk21+jk22
```

```
#Calculate values which is proportional to full conditional distribution for L=2 dimensional model
```

```
L2<-Loglikelihood2+pt21+pt22+pk2+pl2+JJ2
```

```
# L1 is values regarding L=1 dimensional model, calculated as L2
```

```
L1<- Loglikelihood1+pt11+pk1+pl1+JJ1
```

```
#Choose accept or reject about the number of dimension, and set log-likelihood L1 for STEP3 "within move step".
```

```
D<-(L2-L1)
SS<-exp(D)
MM<-pmin(SS,1)
NN<-runif(1)
KK<-sign(MM-NN)
if(KK==1){
  L1<-L2
}else{
  L1<-L1
}
}
```



実データ(相撲) 2005-2009年の一対比較データ

	朝青龍	白鵬	魁皇	千代大海	琴光喜	高見盛	日馬富士	稀勢ノ里	雅山	旭天鵬
朝青龍	0	11	14	16	21	1	16	11	19	19
白鵬	10	0	19	20	18	4	13	14	18	15
魁皇	3	2	0	9	9	0	8	7	12	14
千代大海	4	4	9	0	10	2	7	13	10	10
琴光喜	2	8	14	15	0	3	10	10	15	13
高見盛	0	0	1	0	1	0	4	2	3	1
日馬富士	4	6	12	12	14	3	0	12	9	10
稀勢ノ里	4	4	12	6	10	3	9	0	12	5
雅山	2	4	8	12	9	7	6	4	0	8
旭天鵬	0	2	2	6	7	7	4	8	8	0

*2009年は9月場所まで

*表は、行に示した力士が縦に示した力士に勝った総数を意味する。

