S-PLUSを用いた ベイズ的多次元尺度構成法

岡田 謙介

東京大学大学院 博士課程3年

日本学術振興会 特別研究員

- 導入：多次元尺度構成法の紹介とベイズ推定の利点
- 研究1: ベイズ推定によるMDSの色覚データへの適用
- 研究2: ベイズ推定による多次元展開法の提案と評価・応用
- 研究3: ベイズ推定による確認的MDSの提案と評価・応用
- まとめと結論

多次元尺度構成法(MDS)とは

(Multidimensional Scaling)

- 観測された対象間の非類似度から、低次元空間での対象の布置座標を求める統計的手法.

観測非類似度行列 Δ　　真の距離行列 D　　布置座標 X

$$\delta_{ij} \simeq d_{ij} = \sqrt{\sum_k (x_{ik} - x_{jk})^2}$$

データ

推定

n

n

$\{\delta_{ij}\}$

$\{d_{ij}\}$

p

n

$\{x_{ik}\}$

MDSにおけるベイズ推定

- MDSの推定法としては次の3種類が知られている.
  - 最小二乗法 (Togerson, 1952)
  - 最尤法 (Ramsay, 1977)
  - ベイズ法 (Oh & Raftery, 2001)
- ベイズ推定は最も新しい方法論であり、多くの利点を持つことが報告されている：
  - 漸近論に依らずに推定誤差を評価できる.
  - 次元数選択のための指標MDSICが使える.
  - モデルの拡張性に優れる.
  - 既存の方法よりも推定誤差を小さくできる(次頁).

MDSにおけるベイズ推定(*cont'd*)

- ベイズMDS(BMDS)と既存の方法(CMDS,ALSCAL)の、二乗誤差(STRESS)の比較[1]

ベイズMDSが誤差最小

dim	CMDS STRESS	ALSCAL STRESS	BMDS STRESS
1	.6782	.4007	.3617
2	.4682	.1795	.1604
3	.3811	.0903	.0851
4	.4006	.0902	.0856
5	.4139	.0902	.0854

dim	CMDS STRESS	ALSCAL STRESS	BMDS STRESS
1	.6622	.5079	.4813
2	.4943	.3198	.3063
3	.3720	.2250	.2182
4	.2751	.1648	.1642
5	.2037	.1234	.1234
6	.1580	.0984	.0984
7	.1092	.0772	.0772
8	.0809	.0652	.0652
9	.0672	.0584	.0584
10	.0614	.0527	.0527
11	.0658	.0511	.0511
12	.0715	.0500	.0500
13	.0784	.0497	.0497
14	.0855	.0495	.0495

- しかし、これまでベイズMDSを応用ユーザーが手軽に実行できる方法はなかった.
- そのためもあり、ベイズMDSを用いた応用研究はこれまでに存在しなかった.

1. 出典：Oh & Raftery (2001), Table 1, 2

概要

- 以上を踏まえ、本発表では以下の研究を行う.

研究1:ベイズMDSの応用研究

- よい性質を持つとされるベイズMDSのモデルを用いて、古典的な色覚データの再分析を行う.

研究2:ベイズ推定による多次元展開法

- ベイズMDSのモデルを拡張し、Data augmentation法を利用したベイズ的多次元展開法モデルを提案する. 数値実験による手法の検証と、実データへの応用を行う.

研究3:ベイズ推定による確認的MDS

- ベイズMDSのモデルを拡張し、距離行列に構造のあるモデルに対する確認的MDSの推定法およびモデル評価法を提案する. 数値実験による検証と実データへの応用を行う.

- 分析と結果のプロットはS-PLUS 6.2を用いて行った.

ベイズMDS：モデル

- $\Delta = \{\delta_{ij}\} : (n \times n)$ 観測非類似度行列
- $D = \{d_{ij}\} : (n \times n)$ 真の(モデル)距離行列
- $X = \{x_{ik}\} : (n \times p)$ 布置座標行列

- $m = \dfrac{n(n-1)}{2}$ をユニークな観測非類似度数とする.

- このとき、ベイズMDSでは、観測非類似度 δ_{ij} が真の距離を平均パラメータとしてもつ切断正規分布にしたがうというモデルをおく.

$$\delta_{ij} \sim N(d_{ij}, \phi^2) I(\delta_{ij} > 0) \qquad (i \neq j, j = 1, ..., n)$$

ここで、$d_{ij} = \sqrt{\sum_k (x_{ik} - x_{jk})^2}$.

ベイズMDS：尤度と事前分布

- このとき、未知母数は X と ϕ^2 である.
- 尤度関数は、

$$l(X, \phi^2) \propto (\phi^2)^{-\frac{m}{2}} \exp\left[-\frac{1}{2\phi^2}\sigma_r - \sum_{i>j} \log \Phi\left(\frac{d_{ij}}{\phi}\right)\right].$$

ここで $\Phi(\cdot)$ は標準正規分布の累積分布関数.
また、$\sigma_r = \sum_{i>j}(\delta_{ij} - d_{ij})^2$ (粗ストレス).

- 未知母数に以下の事前分布を設定する.
  - $\mathbf{x}_i \sim N(0, \Lambda)$
    - ただし $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$ は対象iの位置ベクトル.
  - $\Lambda = \mathrm{Diag}(\lambda_1, ..., \lambda_p), \lambda_j \sim IG(\alpha, \beta_j)$
  - $\phi^2 \sim IG(a, b)$

ベイズMDS：事後分布

- 母数 $\boldsymbol{\theta} = (X, \phi^2, \Lambda)$ の事後分布は以下のようになる.

$$\pi(X, \phi^2, \Lambda | \Delta)$$
$$\propto (\phi^2)^{-(m/2+a+1)} \prod_{j=1}^{p} \lambda_j^{-n/2} \exp\left[-\frac{1}{2\phi^2}\sigma_r - \sum_{i>j} \log \Phi\left(\frac{d_{ij}}{\phi}\right) \right.$$
$$\left. -\frac{1}{2}\sum_{j=1}^{n} \mathbf{x}_j^t \Lambda^{-1} \mathbf{x}_j - \frac{b}{\phi^2} - \sum_{j=1}^{p} \frac{\beta_j}{\lambda_j} \right]$$

- この事後分布の形は複雑であり、事後推定値を得るためには数値的方法を用いなければならない.
- そこでMCMC法(Metropolis-Hastingsアルゴリズム)を用いて事後分布の構成とベイズ推定を行う

※Metropolis-HastingsアルゴリズムについてはAppendix参照

研究1 実データ分析1

- Ekman(1954)の色の類似度データを分析.
- 31人の正常な視力を持つ実験参加者に波長434nm～674nmの11種類の色を提示し、その類似度を評定させたもの.

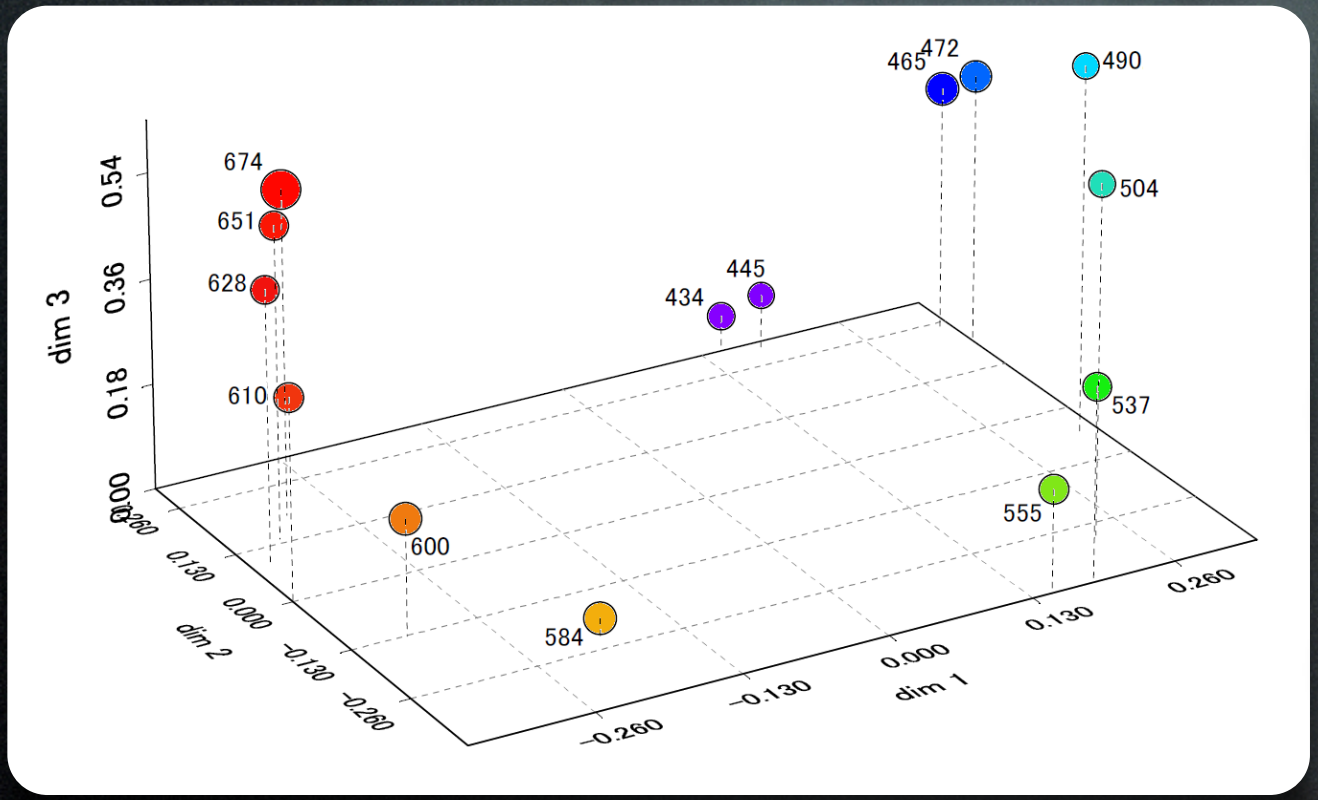
400nm　　500nm　　600nm

SIMILARITY MATRIX

Wave length	434	445	465	472	490	504	537	555	584	600	610	628	651	674
434		.86	.42	.42	.18	.06	.07	.04	.02	.07	.09	.12	.13	.16
445	.86		.50	.44	.22	.09	.07	.07	.02	.04	.07	.11	.13	.14
465	.42	.50		.81	.47	.17	.10	.08	.02	.01	.02	.01	.05	.03
472	.42	.44	.81		.54	.25	.10	.09	.02	.01	.00	.01	.02	.04
490	.18	.22	.47	.54		.61	.31	.26	.07	.02	.02	.01	.02	.00
504	.06	.09	.17	.25	.61		.62	.45	.14	.08	.02	.02	.02	.01
537	.07	.07	.10	.10	.31	.62		.73	.22	.14	.05	.02	.02	.00
555	.04	.07	.08	.09	.26	.45	.73		.33	.19	.04	.03	.02	.02
584	.02	.02	.02	.02	.07	.14	.22	.33		.58	.37	.27	.20	.23
600	.07	.04	.01	.01	.02	.08	.14	.19	.58		.74	.50	.41	.28
610	.09	.07	.02	.00	.02	.02	.05	.04	.37	.74		.76	.62	.55
628	.12	.11	.01	.01	.01	.02	.02	.03	.27	.50	.76		.85	.68
651	.13	.13	.05	.02	.02	.02	.02	.02	.20	.41	.62	.85		.76
674	.16	.14	.03	.04	.00	.01	.00	.02	.23	.28	.55	.68	.76	

実データ分析１ 結果

- ３次元での布置(色は波長に対応)



第1次元
赤ー青 の軸

第2次元
緑ー紫 の軸

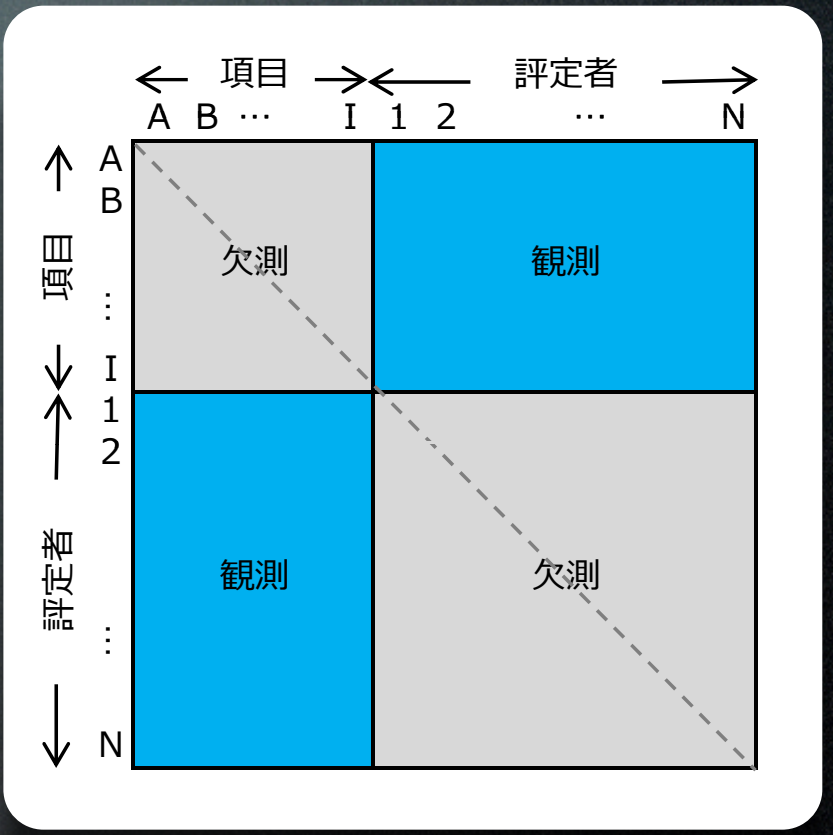
第3次元
鮮やかさ の軸

※丸の大きさは95%信用区間に比例

- 波長を短～長と物理的に操作したとき、色の主観的類似度が円環状に布置されることを実証的に示した.
- 赤の位置について若干ばらつきがあるようだ.

研究2 ベイズ推定による多次元展開法

- 複数の対象に対する評定データ、選好データは心理学・マーケティングで不可欠.
- 典型的には対象×評定者の形
- 選好や評定は、評定者と対象との一種の類似度／距離と捉えることができる.
- したがって、右図全体の距離行列のうち灰色部分が欠測したデータと考えればMDSの枠組みで扱える



ベイズMDSを拡張しData augmentation法を用いることで、このような欠測状況下での分析が可能になる

Data Augmentation法

- Tanner & Wong (1987)によって提案された、欠測を含む場合のベイズ推定法．I (Imputation) stepとP (Posterior) stepからなる．
- 今回の多次元展開法モデルでは、観測非類似度Δの一部が欠測になっている: $\Delta = \{\delta_{(obs)}, \delta_{(mis)}\}$
- したがって、t回めのMCMCで得られた母数ベクトルを$\theta^{(t)}$としたとき、次の2ステップを繰り返せばよい．

I step

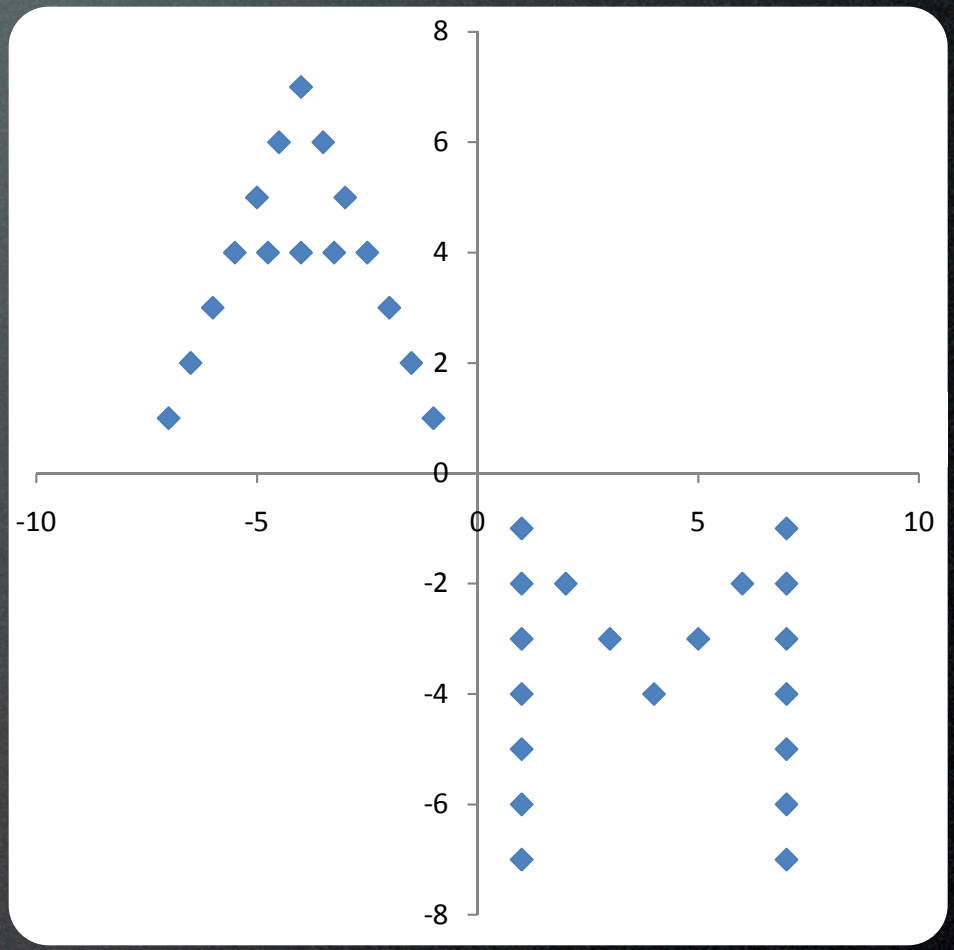
- $\delta_{(mis)}^{(t+1)}$ を $p(\delta_{(mis)}|\delta_{(obs)}, \theta^{(t)}) = N(d_{ij}^{(t)}, \phi^{2(t)})$から発生する

P step

- $\theta^{(t+1)}$ を $p(\theta|\delta_{(obs)}, \delta_{(mis)}^{(t+1)})$ から発生する(これは通常のベイズMDSのMCMCと同じ)

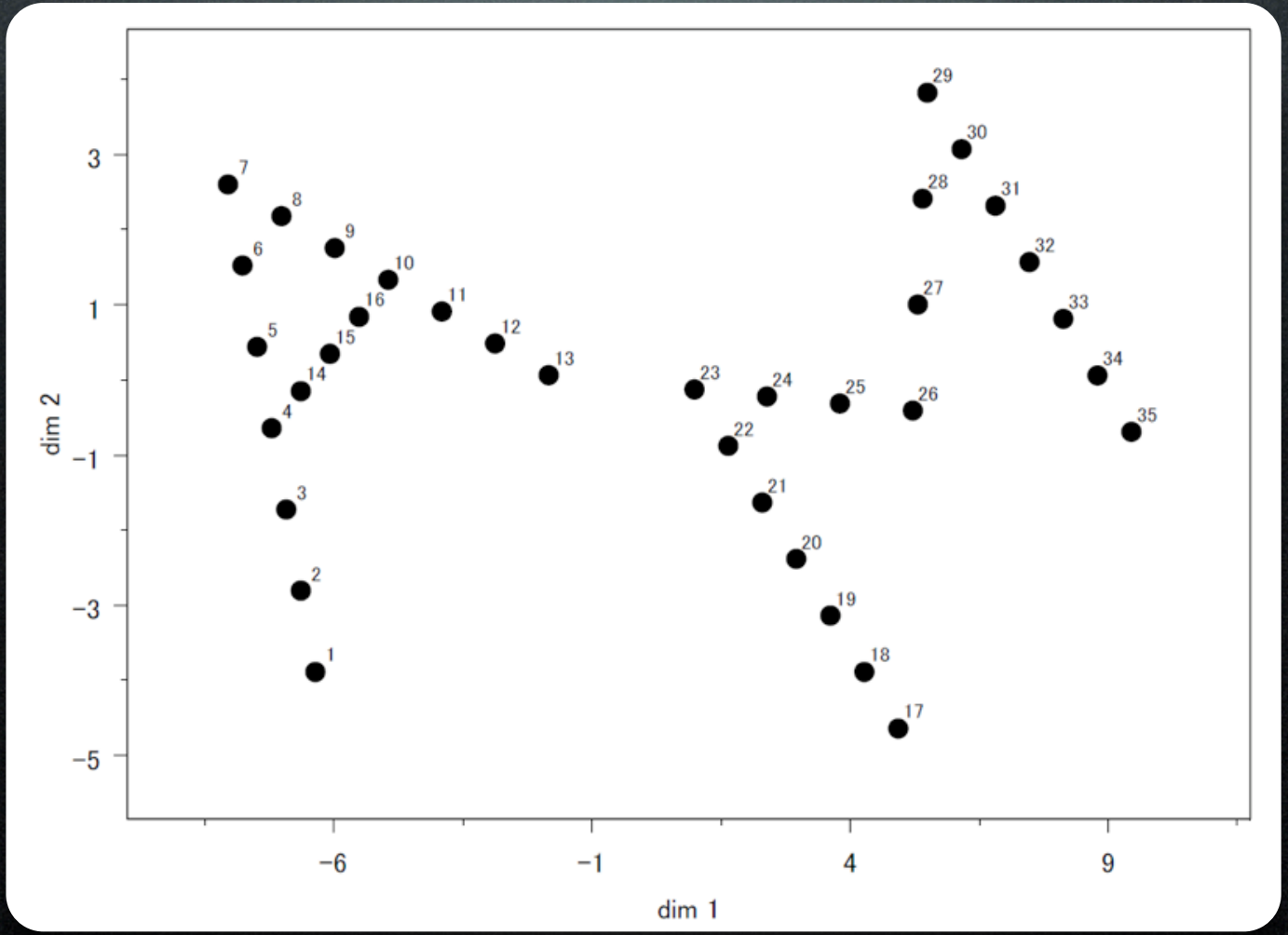
数値実験2

- 提案手法の検証のため、Green and Carmone(1970)が提案したAMデータを用いる.
- 右図のようなAMの２文字を構成する35点がある. Aを構成する点どうし、Mを構成する点どうしの距離は欠測とし、Aの点とMの点の間の距離のみを観測データとして用いる.
- このとき、提案手法により正しく布置を復元できるかを検証する.



数値実験2 結果

- ユークリッド距離が本質的にもつ回転・対称性の不定性を除いては、提案手法により正しい布置が復元できた.



実データ分析2

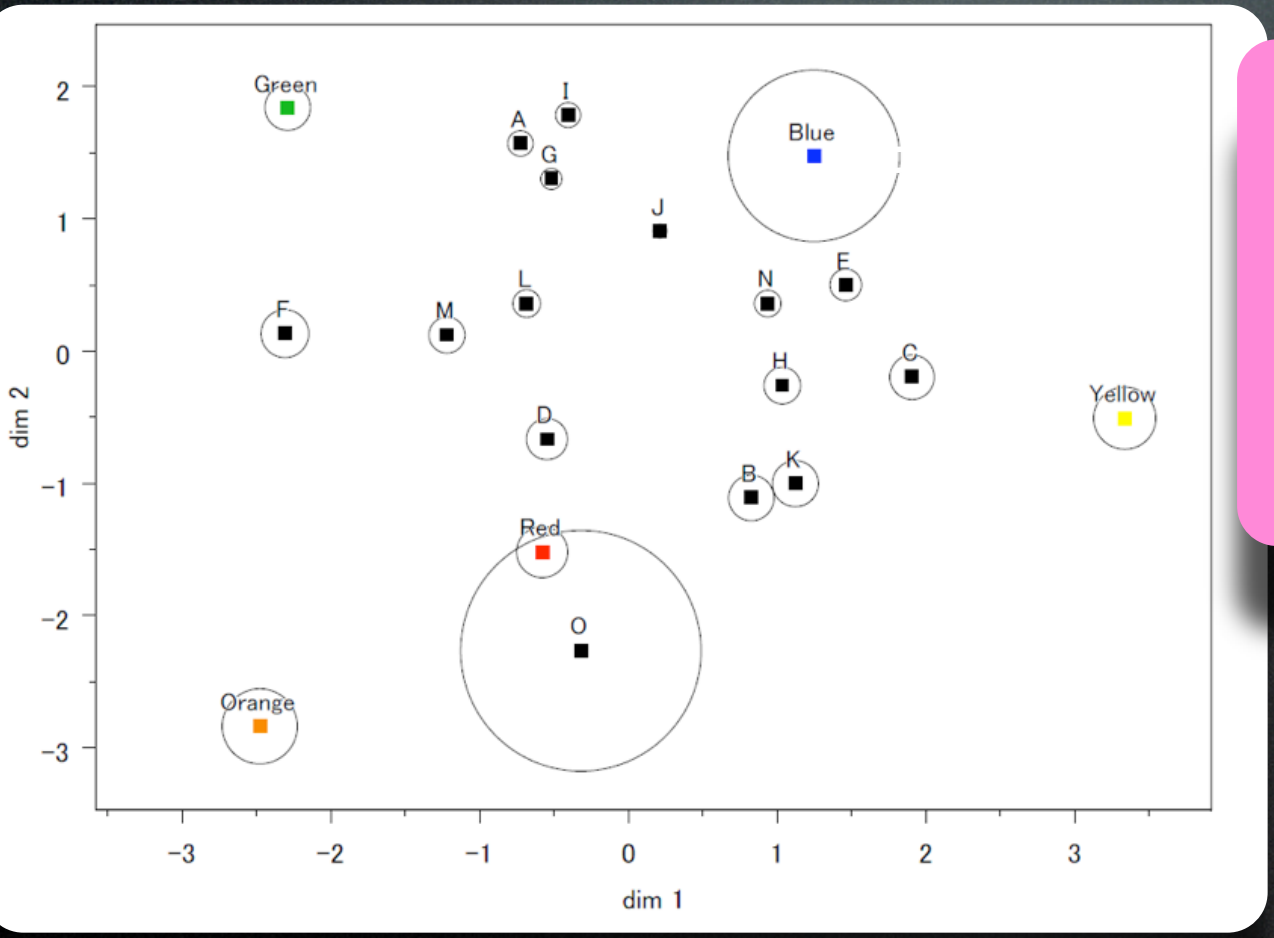
- Wilkinson(1996)の色の選好データ
- 15人の実験参加者に対して、5種類の色(赤, オレンジ, 黄, 緑, 青)に選好順位をつけてもらったデータ.
  - 1:最も好む ←→ 4: 最も好まない

	Person														
Color	A	B	C	D	E	F	G	H	I	J	L	M	N	O	P
Red	3	1	3	1	5	3	3	2	4	2	1	1	1	2	1
Orange	5	4	5	3	3	2	4	4	5	5	5	5	4	5	2
Yellow	4	3	1	5	2	5	5	3	3	4	2	4	5	3	3
Green	1	5	4	4	4	1	2	5	1	3	4	2	2	4	4
Blue	2	2	2	2	1	4	1	1	2	1	3	3	3	1	5

- 提案手法を用いて分析した結果、大まかには外側に色が布置され、内側に各参加者が布置された (次頁).

実データ分析2 結果

- ほとんどの人は好む最も好む色の近くに布置された.
- 本当は青を最も好むが、他の人との関係上青から遠くに布置せざるを得なかった参加者Oについては、推定誤差(点を囲む円で表される)が非常に大きくなった.



このように、推定誤差によって結果の信用できる度合いまでもわかるのがベイズ推定の大きな利点の一つである

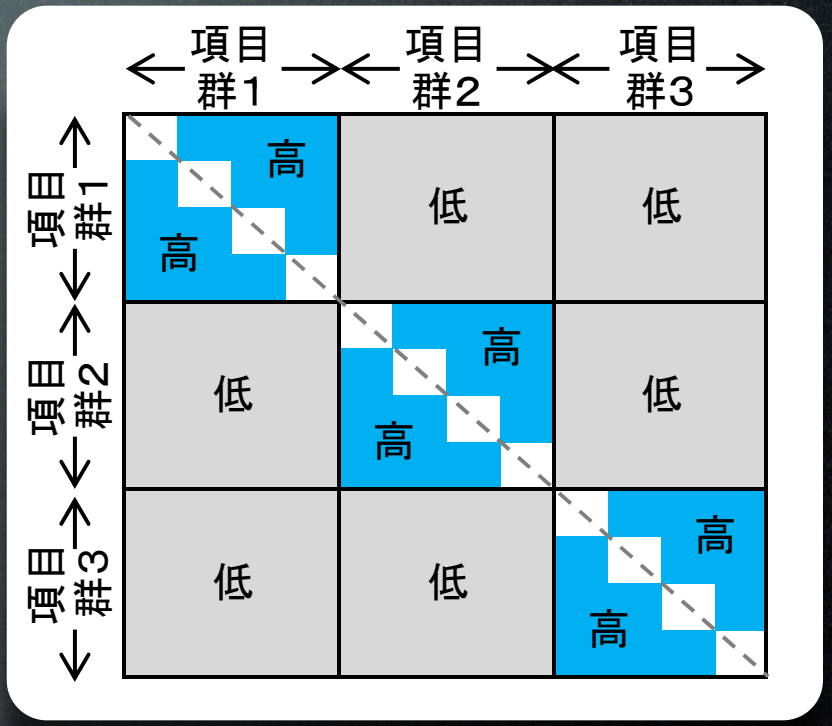
※円はベイズ推定の95%信用区間

研究3 ベイズ推定による確認的MDS

- データを分析するための統計的手法には、大別して探索的方法と確認的方法がある.
  - 探索的方法：母数に関して制約が(できるだけ)ないモデルを用いて、データに潜むパターンを分析する.
  - 確認的方法：仮説を母数に対する制約として導入し、モデル評価によって仮説を検証する.
- 科学的研究には両者がともに必要 (e.g. Tukey, 1980)
- 因子分析では双方がよく発展しているが、MDSでの確認的手法の研究・応用事例は相対的に乏しい.
- 本研究では、ベイズMDSを用いて、距離行列に構造を入れたMDSの推定方法およびモデル評価法を提案する

確認的MDS 方法

- 例えばデータ中に右図のような3項目群があり、各群中では類似度が他より高いことが予想されたとする.
- このような仮説は、距離パラメータへの不等式制約として表現することができる.
- この不等式制約を$\mathscr{C}$ で表す.



類似度の高低

- 制約$\mathscr{C}$ の元での条件付き推定を行うためには、MCMCにおいて次頁の手順を踏めばよい(条件付き推定におけるこの種の方法は, Gelfand & Smith, 1992; Holmes & Heard, 2003 で用いられている).

確認的MDS 方法 (*cont'd*)

1. パラメータ(X, σ^2, Λ)の各々を、通常のMCMCにより発生させる.
2. 対象間の距離
$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$$
を計算する
3. $\{d_{ij}\}$が制約 $\mathscr{C}$ を満たすかどうか判定する. $\mathscr{C}$ を満たさない場合はその回のパラメータ組を棄却、満たす場合は受容し、1へ戻る.

- 1〜3を十分多く繰り返し、受容したMCMC標本のみで事後分布を構成すれば$\mathscr{C}$で制約づけた推定ができる
- さらに、Bayes factorを計算することにより制約のないモデルと比較して制約のあるモデルを評価できる.

数値実験3

- 提案手法が正しい構造を正しいと評価できること、誤った構造を誤りと評価できることを確認する.
- 以下のセッティングよりデータを生成.

データ1

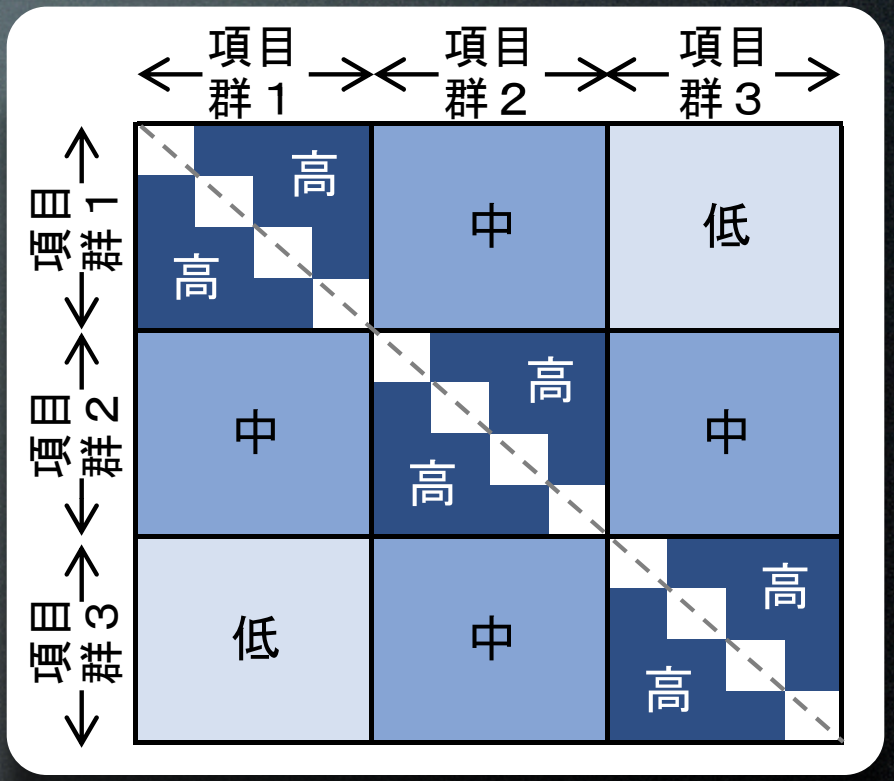
$$\Delta = \begin{pmatrix} 2 & & & & & & & \\ 2 & 2 & & & & & & \\ 3 & 3 & 3 & & & & & \\ 3 & 3 & 3 & 2 & & & & \\ 3 & 3 & 3 & 2 & 2 & & & \\ 4 & 4 & 4 & 3 & 3 & 3 & & \\ 4 & 4 & 4 & 3 & 3 & 3 & 2 & \\ 4 & 4 & 4 & 3 & 3 & 3 & 2 & 2 \end{pmatrix}_{sym.} + N(0, 1^2)$$

データ2

$$\Delta = \begin{pmatrix} 3 & & & & & & & \\ 3 & 3 & & & & & & \\ 3 & 3 & 3 & & & & & \\ 3 & 3 & 3 & 3 & & & & \\ 3 & 3 & 3 & 3 & 3 & & & \\ 3 & 3 & 3 & 3 & 3 & 3 & & \\ 3 & 3 & 3 & 3 & 3 & 3 & 3 & \\ 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 \end{pmatrix}_{sym.} + N(0, 1^2)$$

数値実験3 結果

- 右の構造に対応する不等式制約を$\mathscr{C}$としてデータ1, データ2を分析した.
- この構造はデータ1に対しては真、データ2に対しては偽である.
- 結果、Bayes factorは正しく真偽を判定した.



類似度の高低

データ1

事前確率	事後確率	Bayes Factor
0.2500	0.9486	55.3809

データ2

事前確率	事後確率	Bayes Factor
0.2500	0.1252	0.4292

※Bayes factorは制約のないモデルに比べ構造を入れたモデルがオッズ比の意味でどれだけ「よい」かを示す. 1以下だと制約のないモデルの方がよいことになる.

実データ分析3

- Palan(1998)の多特性・多評定者データ
- 家族コミュニケーションと子どもの消費活動の関係を調べるため、母親・父親・子どもの3評定者に対して家族コミュニケーションの質(CQ)、家族内消費教育(CI)と子どもの消費活動(CA)の3特性の関係を調べたデータ(N=234).
- ３種類(+帰無仮説=制約無し)の構造への仮説が考えうる.

評定者モデル

同一評定者の項目は類似度が高い

特性モデル

同一特性の項目は類似度が高い

双方モデル

上記両方の効果がある

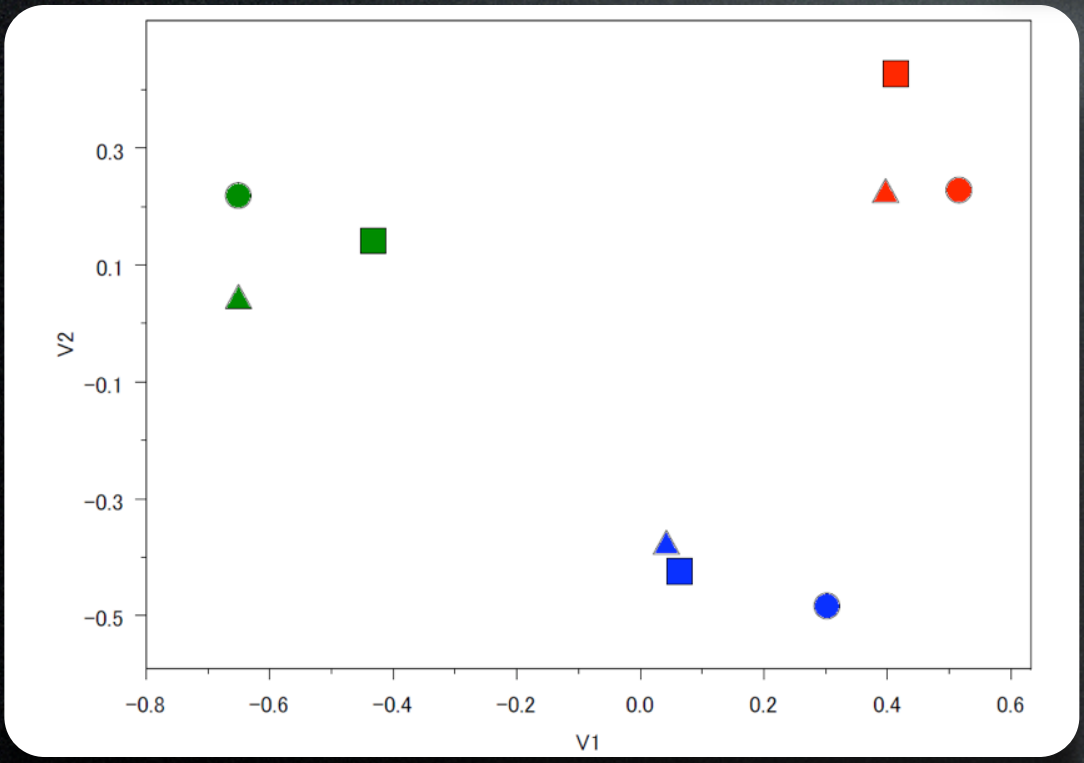
	Mothers			Fathers			Adolescents		
	CQ	CI	CA	CQ	CI	CA	CQ	CI	CA
Mothers									
CQ									
CI	-.01								
CA	-.01	.35							
Fathers									
CQ	.60	.00	-.01						
CI	-.04	.24	.04	.28					
CA	.01	.21	.43	.03	.36				
Adolescents									
CQ	.44	.11	.22	.45	.25	.27			
CI	.14	.25	.06	.22	.22	.20	.18		
CA	.08	-.09	.28	.00	.22	.47	.12	.14	

実データ分析3 結果

- 3つの仮説各々に対応するモデルで分析を行い、Bayes factorを比較した.

	事前確率	事後確率	Bayes Factor
評定者モデル	0.1250	0.073	0.5290
特性モデル	0.1250	0.3517	3.7970
双方モデル	0.0156	0.0343	2.2410

特性モデルの当てはまりが最もよいことがわかった



特性モデルのプロット．評定者によらず特性の効果が強く、家族間で消費行動関連の意識が共有されていることがわかる

緑:CQ 青: CI 赤: CA
●:母 ▲:父 ■:子

まとめと結論

- ベイズ推定を用いた多次元尺度構成法の、
  - 色覚データへの応用研究を行い、先行研究と一致する布置と散布度に関する新たな知見を得た.
  - 方法論的発展として多次元展開法モデルを提案し、その評価・応用を行った.
  - 方法論的発展として確認的MDSモデルを提案し、その評価・応用を行った.
- MDSにおけるベイズ推定はこれまであまり研究されてこなかったが、その有用性が示された.
- S-PLUSを用いてこれらの推定やわかりやすい結果の図示を行うことができた.
- ベイズMDSのS-PLUSコードサンプルを以下で公開
  http://bayes.c.u-tokyo.ac.jp/~ken/BMDSPLUS/

参考文献

- Gelfand, A. E., Smith, A. F., & Lee, T-M. (1992) Bayesian analysis of constrained parameter and truncated data problem using Gibbs sampling. *Journal of American Statistical Association*, **87**, 523-532.
- Green, P. E. & Carmone, F. J. (1970). *Multidimensional scaling and related techniques in marketing analysis*. Boston, MA: Allyn & Bacon.
- Holmes, C. C. & Heard, N. A. (2003). Generalized monotonic regression using random change points. *Statistics in Medicine*, **22**, 623-638.
- Oh, M-S. & Raftery, A.E. (2001). Bayesian multidimensional scaling and choice of dimension. *Journal of American Statistical Association*, **96**, 1031-1044.
- Palan, K.M. (1998). Relationships between family communication and consumer activities of adolescents: an exploratory study. *Journal of the Academy of Marketing Science*, **26**, 338-349.
- Ramsay, J. O. (1977). Maximum likelihood estimation in multidimensional scaling. *Psychometrika*, **42**, 241-266.
- Tanner, M. A. & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with duscussion). *Journal of American Statistical Association*, **82**, 528-550.
- Togerson, W. S. (1952). Multidimensional Scaling: I. Theory and method. *Psychometrika*, **17**, 401-419.
- Tukey, J.W. (1980).We need both exploratory and confirmatory. *The American Statistician*, **34**, 23-25.
- Wilkinson, L. (1996). Multidimensional scaling. In L. Wilkinson (Ed.), *Systat 6.0 for Windows: Statistics* (pp.573-606). Chicago, IL: SPSS Inc.

Appendix: 分析のセッティング

- 分析に利用したMCMCの繰り返し回数は以下の通りであった.

	burn-in	MCMC
実データ分析1	3,000	10,000
数値実験2	500	5,000
実データ分析2	3,000	10,000
数値実験3	5,000	200,000
実データ分析3	5,000	200,000

- 各母数の初期値、および超母数($\alpha, \beta_j, a, b, \gamma$)はOh & Raftery(2001)にしたがって設定した.

Appendix: Metropolis-Hastings

- 提案分布 $q(\theta)$ と棄却サンプリングを用いるMCMCの一種
- t+1回目のパラメータ候補θ' について、$u \sim U(0,1)$ が

$$u < \frac{\pi(\theta')q(\theta^{(t)}|\theta')}{\pi(\theta^{(t)})q(\theta'|\theta^{(t)})}$$

を満たせば $\theta^{(t+1)} = \theta'$とする．満たさなければθ' は棄却され、パラメータの値は$\theta^{(t)}$のままである.

- の提案分布は

$$q(\mathbf{x}_i'|\mathbf{x}_i^{(t)}) = N\left(\mathbf{x}_i^{(t)}, \gamma\frac{\phi^2}{n-1}\right)$$

$$q(\phi^{2\prime}|\phi^{2(t)}) = N\left(\phi^{2(t)}, \gamma\frac{\frac{\sigma_r}{2}+b}{(\frac{m}{2}+a-1)^2(\frac{m}{2}+a-2)}\right)$$

- λ については完全条件付き事後分布が求まり、

$$\pi(\lambda_j|X,\phi^2) = IG(\alpha+\frac{n}{2}, \beta_j+\omega_j)$$

- ω_j は X のj番目の次元の標本分散