Cluster Analysis of Heterogeneous Rank Data

L.M.Busse, Peter Orbanz, J.M.Buhmann

発表者：瀬々　潤

なぜこの論文を選んだか

- Rank Dataに興味がある
  - Google の様なランキングとか，時系列の並びとか
  - 定性的でも，定量的でもないケース
  - グラフでもない
- クラスタリングに興味がある

想定しているケース

- この論文で考えているのは，次のような投票のケース
  - 会員の中から，評議員を選出したい
  - 投票者は上位5名を記名できる
  - 候補は会員全員（会員数＞＞5名）
  - 投票者の投票傾向をクラスタ化して知りたい
- 問題点は
  - 投票者は全ての会員に順序を付けるわけではない
  - 各投票者が投じる人は，ばらつきがある
  - 投票したい人の数は3名かもしれない
    - 5名記名欄があっても，全部を埋めてくれるとは限らない
- 投票者間の距離を示すmetric spaceが問題

論文の流れ

- 背景
  - 完全に順序のついたランクデータの扱い
  - Mallow’s Model
  - 部分的にしかランクのないデータの扱い
- モデル化
  - 確率モデルを作成し,
  - EMでパラメータを決定する
- クラスタリング, はオマケ程度

全ての順序がある場合

- アイテム（候補者）数: r, 投票者: n
- i番目の投票者がアイテムmをj番目にランクした時, $\pi_i(m) = j$
- 距離関数 ：ランクデータに対する「距離」モデル

$$P(\pi \mid \lambda, \sigma) := \frac{1}{Z(\lambda)} \exp(-\lambda d(\pi, \sigma))$$

$$Z(\lambda) := \sum_{\pi \in S(r)} \exp(-\lambda d(\pi, \sigma)) \qquad \sigma \in S(r), \lambda \in \mathbf{R}$$

- 距離 $d : \mathbf{S}(r) \times \mathbf{S}(r) \to \mathbf{R}_{\geq 0}$
  - ここではKendall distanceを用いる.
    - この時, 上記の距離関数はMallows’ ϕ モデルと呼ばれる
    - Hamming, Cayley, Ulamなど様々な距離があるが・・・
- Kendall distance: $d_\tau(\pi, \sigma)$
  - π を σ に変換するのに必要な隣接順序の交換回数
  - A>B>C>D → A>C>B>D : 距離1

Mallows' modelを用いたクラスタリング

- K個のグループに分ける
- 各グループに中心順序 σ_k があるとする
- 各グループは，Mallows' 分布を用い，次式でモデル化

$$P(\pi \mid \lambda_k, \sigma_k) \quad := \quad \frac{1}{Z(\lambda_k)} \exp(-\lambda_k d(\pi, \sigma_k))$$

- 重み $(c_1, ..., c_K)$ を付けた混合分布を考えて

$$Q(\pi) := \sum_{k=1}^{K} c_k P_k(\pi \mid \lambda_k, \sigma_k)$$

- モデルパラメータ $(\mathbf{c}, \lambda, \sigma)$ をEMやら，SAやらで求める.

部分的なランキング

- r個のアイテム(候補)から，t個をランキング
- 置換πのinverse π^{-1}を考える
- Top-t ランキングは，次のように表される

$$\pi^{-1} = (\pi^{-1}(1), \cdots, \pi^{-1}(t), *, \cdots, *)$$

- Top-tのランキングの置換をC(π)と表す

$$C(\pi) := \{\overline{\pi} \in \mathbf{S}(r) \mid \overline{\pi}(j) = \pi(j), j = 1, ..., t\}$$

  - $\overline{\pi}$ は，完全なランキングをした場合の置換

- 置換の和集合の様な形で距離を拡張：Critchlow, 1985
- Mallows' model を使って，補完：Beckett, 1993

異種混合データのモデル化

- Kendall 距離をベースに考える
- 確率モデル
  - エントロピー最大化を行う
- $P(\pi)$をTop-tでの確率$P^t(\pi)$に一般化したとすると

$$P^t(\pi) := P(C(\pi)) = \sum_{\overline{\pi} \in C(\pi)} P(\overline{\pi})$$

- 各順位 j に対して，次の統計量を導入する

$$\overline{s}_j(\pi) := \sum_{l=j+1}^{r} I\{\pi^{-1}(j) > \pi^{-1}(l)\}$$

  - Iはindicator function
  - 順位がjより大きい中で，jより，会員番号が小さい会員数

異種混合データのモデル化(続)

- $d_\tau(\pi_1\pi_3, \pi_2\pi_3) = d_\tau(\pi_1, \pi_2)$ が成り立つので

  $$d_\tau(\pi, \sigma) = d_\tau(\pi\sigma^{-1}, \mathrm{Id}_{\mathrm{S}(r)}) = \sum_{j=1}^{r-1} s_j(\pi\sigma^{-1})$$

- 右辺の計算が大変なので,

$$\overline{s}_j(\pi^{-1}) := \pi(j) - \sum_{l=1}^{j} I\{\pi^{-1}(j) \geq \pi^{-1}(l)\}$$

$$\begin{aligned} d_\tau(\pi, \sigma) &:= \sum_{j=1}^{r} s_j(\pi^{-1}\sigma) \\ &= \sum_{j=1}^{t} s_j(\pi^{-1}\sigma) + \sum_{j=t+1}^{r} s_j(\pi^{-1}\sigma) \\ &= \mathrm{s}^t(\pi^{-1}\sigma) + \mathrm{s}^{\mathrm{empty}}(\pi^{-1}\sigma) \end{aligned}$$

$$
\begin{aligned}
P(C(\pi) \mid \lambda, \sigma) &= \frac{1}{Z(\lambda)} \sum_{\overline{\pi} \in C(\pi)} \exp(-\lambda d_\tau(\overline{\pi}, \sigma)) \\
&= \frac{\exp(-\lambda \mathrm{s}^t(\pi^{-1}\sigma))}{Z(\lambda)} \sum_{\overline{pi} \in C(\pi)} \exp(-\lambda \mathrm{s}^{\mathrm{empty}}(\overline{\pi}^{-1}\sigma)) \\
&= \frac{1}{Z^t(\lambda)} \exp(-\lambda \mathrm{s}^t(\pi^{-1}\sigma))
\end{aligned}
$$

$$
Z^t(\lambda) := \prod_{j=1}^{t} \frac{1 - e^{-\lambda(r-j+1)}}{1 - e^{-\lambda}}
$$

$$
Q(\pi \mid \mathbf{c}, \lambda, \sigma) := \sum_{k=1}^{K} \frac{c_k}{Z^{t(\pi)}(\lambda_k)} e^{-\lambda_k \mathbf{s}^{t(\pi)}(\pi^{-1}\sigma_k)}
$$

EMで解く

- $\mathbf{M}_i := (M_{i1}, ..., M_{iK})$
  - π_i がクラスタkに属する時, $M_{ik} = 1$
  - $q_{ik} := \mathrm{E}[M_{ik}]$ とすると, $\sum_k q_{ik} = 1$ を満たす
- E step
  - 推定量 c_k, λ_k, σ_k が与えられたとして，qik を次式で推定

$$q_{ik} = \frac{c_k P^t(\pi_i \mid \lambda_k, \sigma_k)}{\sum_{l=1}^{K} c_l P^t(\pi_i \mid \lambda_l, \sigma_l)}$$

- M step
  - $c_k = \frac{1}{n} \sum_{i=1}^{n} q_{ik}$
  - σ_k の後, λ_k を求める（λ_kが σ_k に依存するため）

$$\hat{\sigma} = \arg\max_{\sigma_k} \log \prod_{i=1}^{n} P(\pi_i^t \mid \lambda_k, \sigma_k)^{q_{ik}}$$

$$= \arg\min_{\sigma_k} \sum_{i=1}^{n} q_{ik} \sum_{j=1}^{t(\pi_i)} s_j(\pi_i^{-1}\sigma_k)$$

  - 全探索はせず，ローカルサーチだけしている

- λ について確率を最大化する

$$\log \prod_{i=1}^{n} P(\pi_i \mid \lambda_k, \sigma_k)$$

$$= \sum_{t=1}^{r} \sum_{i \in I_t} \log P(\pi_i \mid \lambda_k, \sigma_k)$$

$$= -\sum_{t=1}^{r} |I_t| \log(Z^t(\lambda_k)) - \sum_{t=1}^{r} \sum_{i \in I_t} \lambda_k \sum_{j=1}^{t} s_j(\pi_i^{-1}\sigma_k)$$

- 極大点では微分が0になるので,

$$-\sum_{t=1}^{r} |I_t| \frac{\partial}{\partial \lambda_k} \log(Z^t(\lambda_k)) = \sum_{i=1}^{n} \sum_{j=1}^{t(\pi_i)} s_j(\pi_i^{-1}\sigma_k)$$

- logの部分は下記の式で簡単にでき，方程式を解けば λ が求まる

$$\frac{\partial}{\partial \lambda_k} \log(Z^t(\lambda_k)) = \sum_{j=r-t+1}^{r} \frac{j}{e^{j\lambda_k} - 1} - \frac{t}{e^{\lambda_k} - 1}$$

実験結果(1)

Table 1. Estimation errors on artificial data of sample size $n = 300$, with $K = 3$ clusters. For uniform c, all clusters have equal size. For non-uniform c, cluster sizes differ.

Settings			Results		
c	d	λ	$\hat{K}$	error $\hat{c}$	error $\hat{\lambda}$
uniform	[2, 9, 9]	0.50	1	0.033	0.086
		1.00	3	0.007	0.056
		1.50	3	0.027	0.151
	[8, 6, 6]	0.50	1	0.155	0.274
		1.00	3	0.029	0.094
		1.50	3	0.016	0.050
non-uniform	[2, 9, 9]	0.50	1	0.248	0.324
		1.00	3	0.013	0.032
		1.50	3	0.001	0.048
	[8, 6, 6]	0.50	1	0.189	0.331
		1.00	3	0.047	0.144
		1.50	3	0.013	0.057

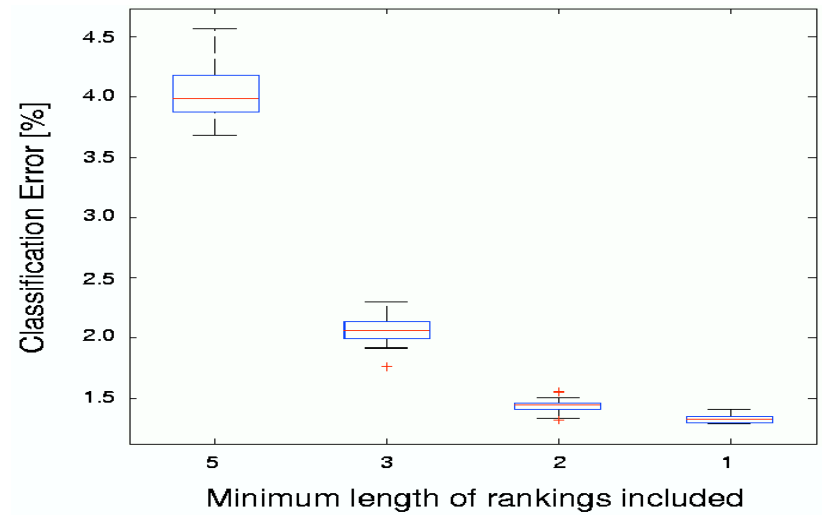

Figure 1. Full versus restricted data set: Average estimation error for cluster assignments (vertical) versus the number of ranking types present in the data set (horizontal).

Table 2. Long rankings: Estimation error comparison for ranking length $r = 20$, with $K = 10$ clusters and $n = 1000$ samples (uniform over partial lengths).

Method	error $\hat{\sigma}_k$	error $\hat{\lambda}_k$
Maximum Entropy	0	0.06 ± 0.01
Beckett's completion	1.52 ± 0.57	0.11 ± 0.02

実験結果(2)

- American Psychological Association
- 15,000の候補者から5人投票

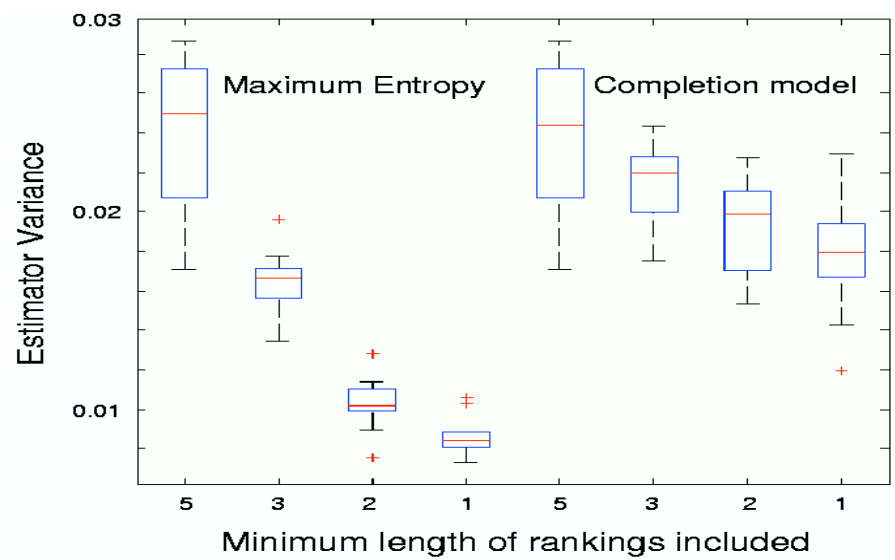

Figure 2. APA data set: Variance of dispersion estimates (vertical) versus number of ranking types present in the data set (horizontal), for our method (left) and Beckett's completion model (right). Minimum length 5 corresponds to the subset of complete rankings, 1 to the whole data set. The variance is computed over 20 bootstrap samples.

まとめ

- 部分的にランキングされてデータをクラスタリングした
- ランクとランクの間の距離はKendall距離を利用した
- 各クラスタを確率分布で表した．その分布をEMで求めた
- EMの計算は，解析的に計算できる部分があり，高速化できることがわかった
- もう少し，ちゃんとしたクラスタリング結果が欲しいなぁ