集約的シンボリックデータの階層的クラスタリング

清水　信夫　　データ科学研究系　助教

何故シンボリックデータ解析か？
- 解析対象とする多変量データが大規模化かつ多様化
- それらを記述する上で柔軟なデータ構造を定義した枠組みが必要
  ⇒Didayにより<u>シンボリックデータ</u>(SD)が定義され、
  これらを解析する枠組みとして<u>シンボリックデータ解析</u>(SDA)が出現

シンボリックデータの例としてどのようなものがあるのか？
- 1個の量的データ
- 1個の質的データ
- 複数の量的データもしくは質的データの集合
- 区間データ
- 上記の各種データに重みがついたデータの集合（ヒストグラムデータなど）
- 何らかの関係に基づく依存構造をもつデータ集合
  など

集約的シンボリックデータとは何か？
- SDAにおいてデータ集合に対しいくつかのグループ分けが
  既に行われている場合にオリジナルデータではなく
  <u>各々のグループについての情報</u>に興味がある場合が存在
  ⇒それらのグループの特徴を<u>分布として表現</u>し、
  それを近似的に表現した統計量をデータと考えたものを
  <u>集約的シンボリックデータ</u>と呼ぶ

従来のシンボリックデータ解析の特徴と問題点は？
- 主に多変量区間データを考慮し、それらに既存の各種統計手法を拡張
  ⇒各変数ごとの平均および分散の情報（＝周辺分布の情報）のみを用いた解析
- 分散と同様に2次のモーメントである各変数間の相関係数は解析に利用されず
  ⇒相関係数の情報も用いてより詳細な解析を行えないだろうか？

本報告における提案
- 周辺分布の情報に加え相関係数の情報も利用した集約的SD間の非類似度
  およびそれを用いた階層的クラスタリング手法の提案

集約的シンボリックデータ間の非類似度の定義

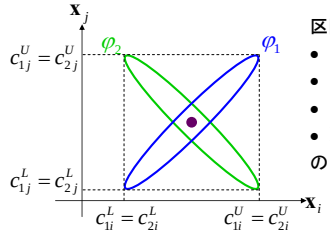
$\varphi_1,\cdots,\varphi_m$：$m$個の集約的SDの確率密度関数（<u>正規分布</u>を仮定）
$\boldsymbol{\mu}_i$：φ_i の平均　　$\boldsymbol{\Sigma}_i$：φ_i の分散共分散行列
p_i：φ_i に含まれるオリジナルデータ数の総数に対する比率

$$d_{(ij)} = p_i p_j \iint (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)\varphi_i(\mathbf{x}_i)\varphi_j(\mathbf{x}_j)d\mathbf{x}_i d\mathbf{x}_j$$

（階層的クラスタリングの群平均法における 2乗距離を用いた場合の定義と同等）

$$\Rightarrow d_{(ij)} = p_i p_j\{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)'(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + trace(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j)\}$$

従来の区間データ間の距離（非類似度）規準との違い



区間データ間の距離規準としてよく利用される
- Hausdorff 距離
- Euclidean Hausdorff 距離
- Gowda-Diday 非類似度
- Ichino-Yaguchi 非類似度

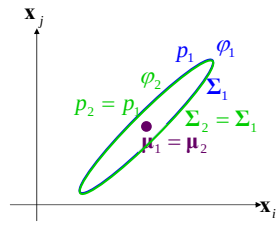
のいずれの場合でも $d_{(12)} = 0$

集約的SDの定義を用いると
$d_{(12)} = p_1 p_2 trace(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$**と表せる**

今後の課題

$$d_{(ij)} = p_i p_j\{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)'(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + trace(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j)\}$$

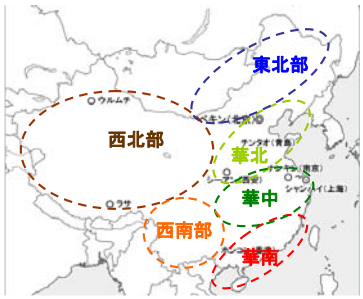
⇒ **結果的に変数間の相関が反映されていない**



$\Rightarrow d_{(12)} = p_1 p_2 trace(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2) \neq 0$

同一の集約的SD間の非類似度を0としつつ変数間の相関が考慮されるような
統計学的に合理的な非類似度をどう定義するかが課題である。

中国60都市の月別平均気温データ(1988年)の階層的クラスタリング（最長距離法を利用）

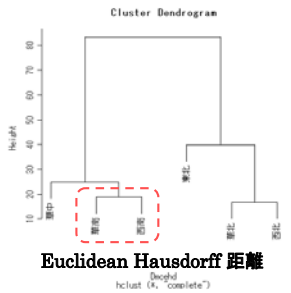


http://dss.ucar.edu/datasets/ds578.5/

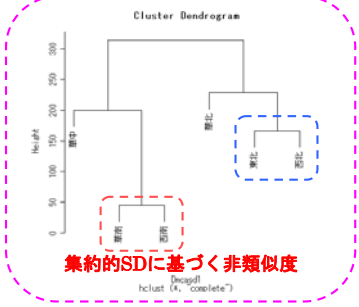
Meteorological Station	January	February	...	December
AnQing (安慶) 華北	[1.8, 7.1]	[5.2, 11.2]	...	[4.3, 11.8]
BaoDing (保定) 華北	[-7.1, 1.7]	[-5.3, 4.8]	...	[-3.9, 5.2]
BeiJing (北京) 華北	[-7.2, 2.1]	[-5.3, 4.8]	...	[-4.4, 4.7]
...	...	...	...	...
ChangChun (長春) 東北部	[-16.9, -6.7]	[-17.6, -6.8]	...	[-15.9, -7.2]
ChanSha (長沙) 華中	[2.7, 7.4]	[3.1, 7.7]	...	[4.1, 13.3]
ZhiJiang (枝江) 華中	[2.7, 8.2]	[2.7, 8.7]	...	[5.1, 13.3]

地域	都市数	January	February	...	December
東北部	8	[-28.4, -6.1]	[-29.6, -1.5]	...	[-26.1, -2.5]
華北	12	[-11.9, 7.0]	[-12.8, 7.1]	...	[-9.9, 8.3]
華中	14	[-2.7, 12.9]	[-2.0, 10.8]	...	[-1.1, 16.6]
華南	10	[8.6, 23.0]	[7.3, 23.0]	...	[4.7, 22.4]
西南部	5	[2.2, 18.5]	[3.4, 22.1]	...	[3.8, 18.0]
西北部	11	[-15.6, 9.0]	[-19.4, 11.2]	...	[-12.7, 8.8]

Hausdorff 距離



Euclidean Hausdorff 距離



集約的SDに基づく非類似度

⇒ **他の距離規準と比べて隣接地域間の近さの特徴をより捉えた結果が出ている**