２段階推定法による階層潜在クラスモデルの推定

-階層クラスタリング法と情報ボトルネックEM法を用いて-

東京工業大学社会理工学研究科
人間行動システム専攻 前川研究室
M2 岡田棟大

潜在クラスモデルとは

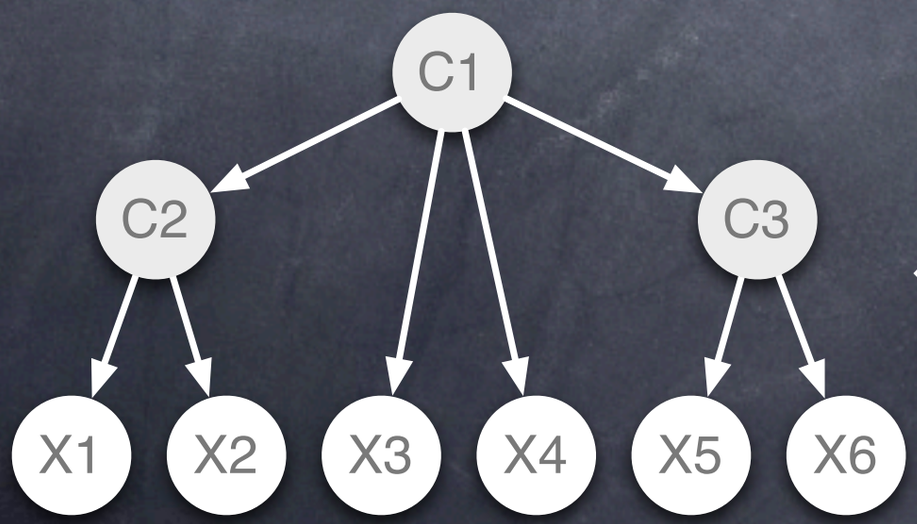
- 潜在的なクラスを表す変数を与える事で,
  各観測変数が局所独立となると仮定する確率モデル

$$P(C, X_1, \ldots, X_n) = P(C) \prod_{i=1}^{n} P(X_i|C)$$

- 潜在クラスモデルを用いた分析で可能な事
  - 各観測変数の背後に潜むクラスの特徴の分析
  - 各観測個体のソフトクラスタリング
  - しかし，局所独立の仮定はかなり強い仮定

階層潜在クラスモデルとは

- ベイジアンネットワーク[Perl'88]を用いて,
  潜在変数を階層化し潜在クラスモデルの局所独立の仮定を緩和したモデル
  - [Zhang et al.'03]では次のように定義
    - モデルの構造は木構造
    - 葉ノードは全て観測変数



階層潜在クラスモデルの例
潜在変数C1を所与としても,
C2とC3に属す観測変数間は
局所独立とならない

階層潜在クラスモデルの推定法

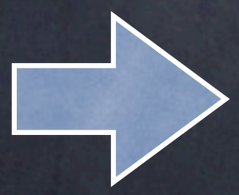
- 階層潜在クラスモデルの推定法として，次の２つの手法が提案されている

	CPT推定	クラス数推定	階層構造推定
Zhang el al. ’03,’04	EM法	探索的手法	探索的手法
Elidan&Friedman ’05	IB-EM法+確定的アニーリング(DA)	IB-EM法+DAを利用した手法	IB-EM法+DAを利用した手法

* CPT：条件付確率表（Conditional Probability Table）
* IB-EM法[Elidan&Friedman’03], 確定的アニーリング[Rose’98]

先行研究の問題点と研究目的
-２段階推定法の提案-

- [Zhang et al.’03,’04]の問題点
  - ✓ アルゴリズムが複雑，計算量が多い
- [Elidan&Friedman’05]の利点と問題点
  - ✓ Zhangらの手法よりシンプルで効率的
  - ✓ 階層構造の推定が頑健でない
- [Elidan&Friedman’05]の階層構造推定法を改良し，より良い推定法を開発したい



そこで，本研究では２段階推定により
効率的かつ頑健な推定法を提案する

本研究で提案する階層潜在クラスモデルの２段階推定法

- 提案する推定法の概要

  - 第１段階
    “潜在変数間の階層構造のトポロジー推定”
    相互情報量に基づく距離を用いた階層クラスタリング法による推定

  - 第２段階
    “トポロジー情報を利用した階層潜在クラスモデルの推定”
    [Elidan&Friedman’05]の階層構造の推定法を，第１段階で求めたトポロジー情報を利用するように改良

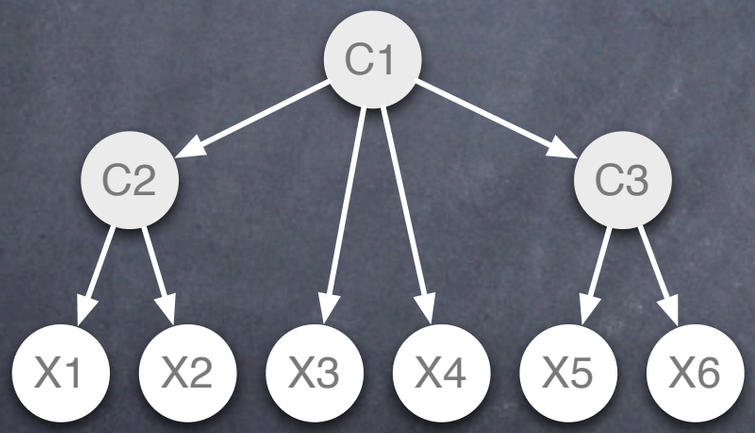
階層クラスタリングによる潜在変数間の階層構造のトポロジー推定

第１段階

- 潜在変数が階層構造を持つ場合，次の関係が予想される

$$I_{brother} \geq I_{relative}$$

  - $I_{brother}$：同じ親を持つ観測変数間の相互情報量
  - $I_{relative}$：親の違う観測変数間の相互情報量



左に示すモデルの場合

$$I(X_1;X_2) \geq I(X_1;X_3)$$
$$I(X_1;X_3) \geq I(X_1;X_5)$$

上の不等式が成り立つと仮定

→ 相互情報量による階層クラスタリングにより，潜在変数間の階層構造のトポロジーを推定できる

人工データを用いた階層クラスタリングの有効性の検証

第１段階

- 人工データを用いて，
  階層クラスタリングによる潜在変数間の階層構造の
  トポロジー推定が有効であるかを検証する

- S-PLUSで提供されている以下の手法について比較する
  - 群平均法，最長距離法，ウォード法
  - 観測変数間の距離には相互情報量に基づく次の距離
    [Li et al.’02]を用いる

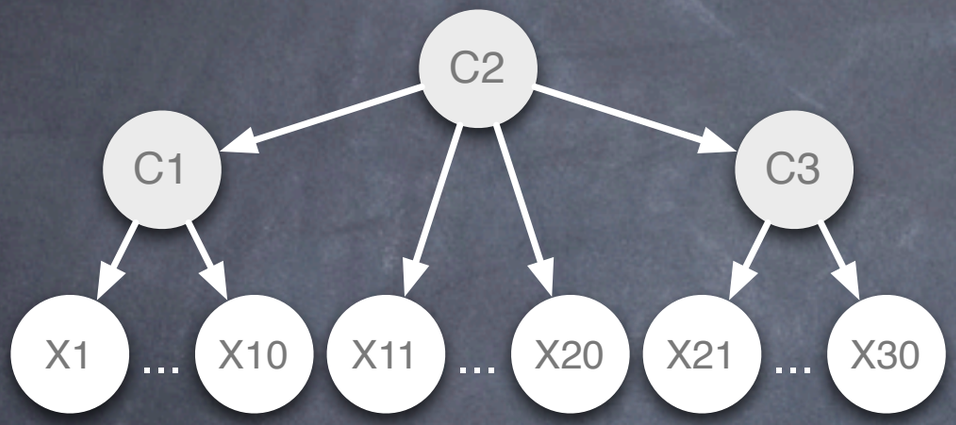
$$D(X,Y) = 1 - \frac{I(X;Y)}{H(X,Y)}$$

I(X;Y)：XとYの相互情報量
H(X,Y)：XとYの結合エントロピー

人工データでの数値実験

第１段階

- 人工データの生成モデルには次のモデルを用いる



- 潜在変数：C1,C2,C3（2値）
- 観測変数：X1~X30（２値）

	親変数 = 0	親変数 = 1
変数 = 1	0.2	0.8
変数 = 0	0.8	0.2

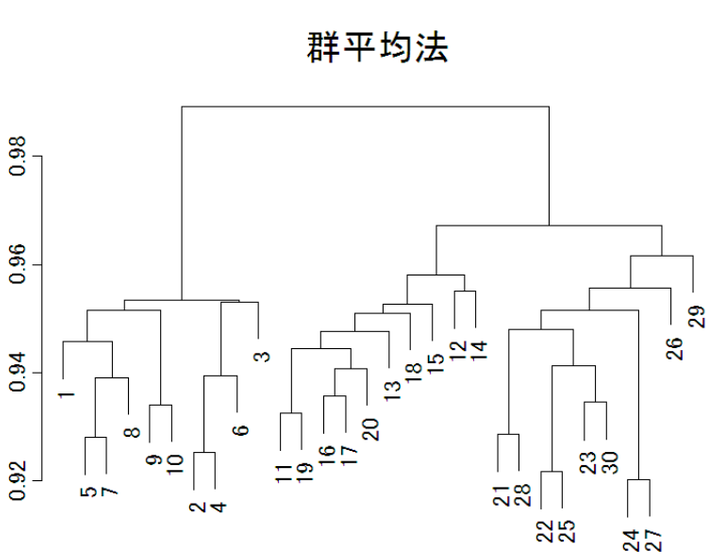
C2以外の変数のCPT (C2は一様分布)

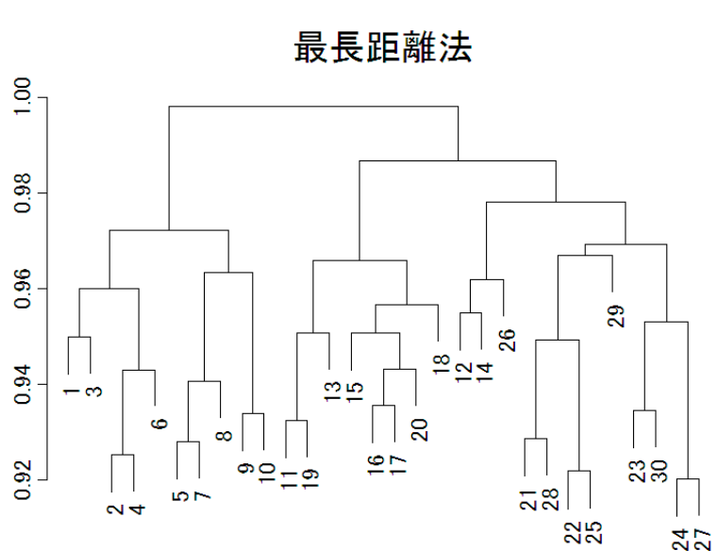
各変数間には強い
確率依存関係を設定

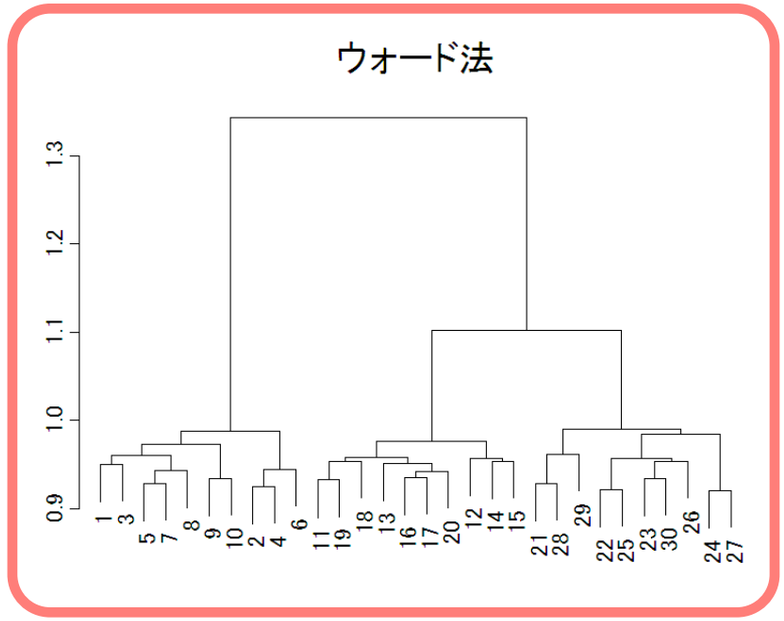
- ギブスサンプリングにより500のサンプルを生成

人工データでの数値実験の結果

第１段階





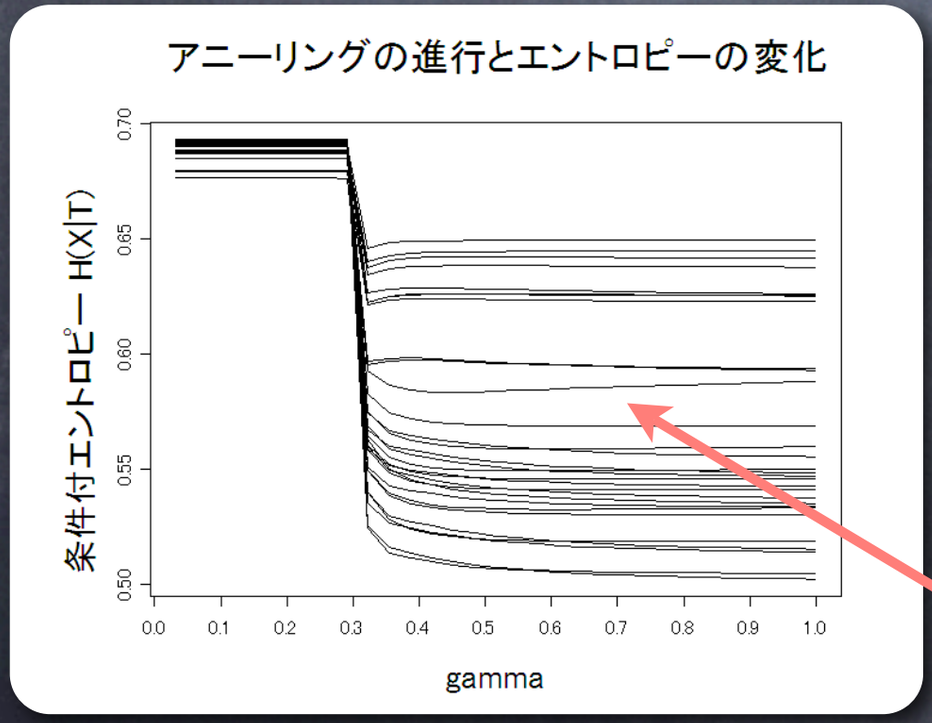


- ウォード法が元のモデルの階層構造のトポロジーを最も良く見つけ出していることが分かる
  - ウォード法による潜在変数間の階層構造のトポロジー推定が有効であるといえる

階層潜在クラスモデルの推定

第2段階

- IB-EM法 + 確定的アニーリング：
  潜在変数と観測変数間の相互情報量を０から徐々に増加
  (条件付エントロピーは減少) するようにCPTを推定していく手法
  → モデルを徐々に”informative”になるように推定



条件付エントロピーの減少の様子
(横軸はアニーリングの進行度合い)

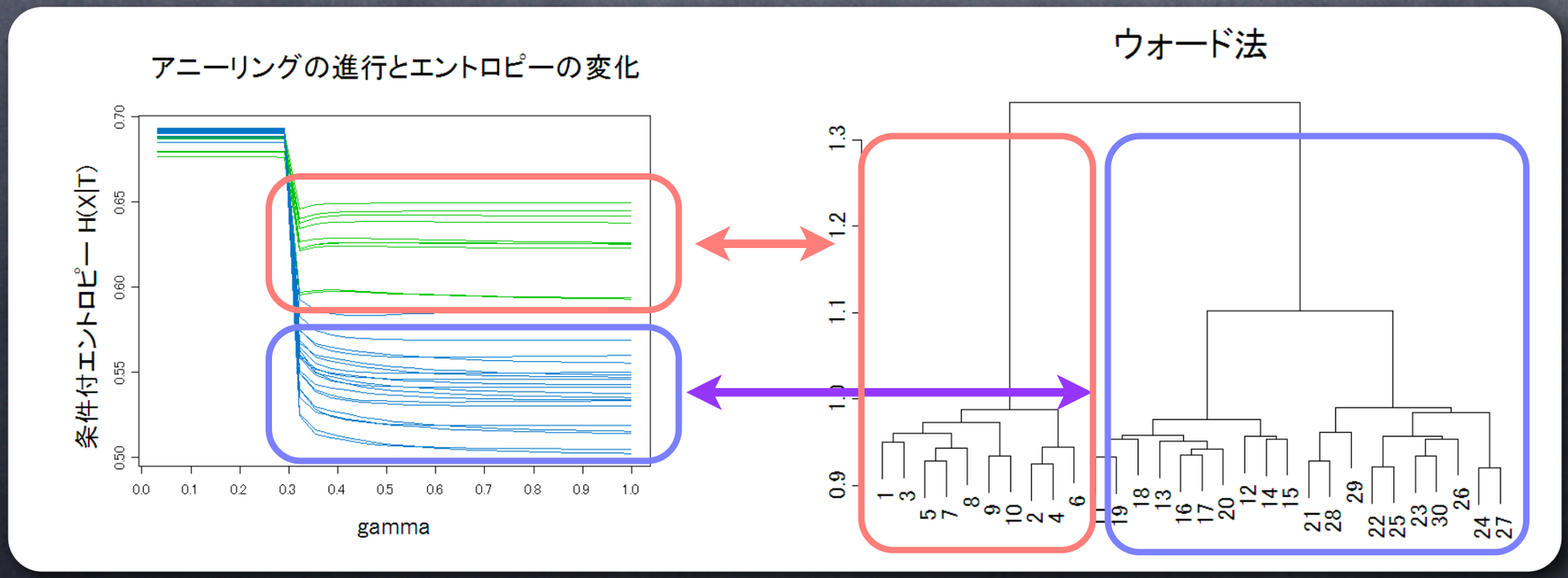
[Elidan&Friedman’05]では，
条件付エントロピーの高い変数群に対して逐次新しい潜在変数を追加してくことで階層構造を推定

“エントロピーの高い変数”の基準は分析者が決めなければならない
→ この基準は推定される階層構造に大きく影響する（頑健でない）

エントロピーとトポロジー情報を利用した階層構造の推定法の提案

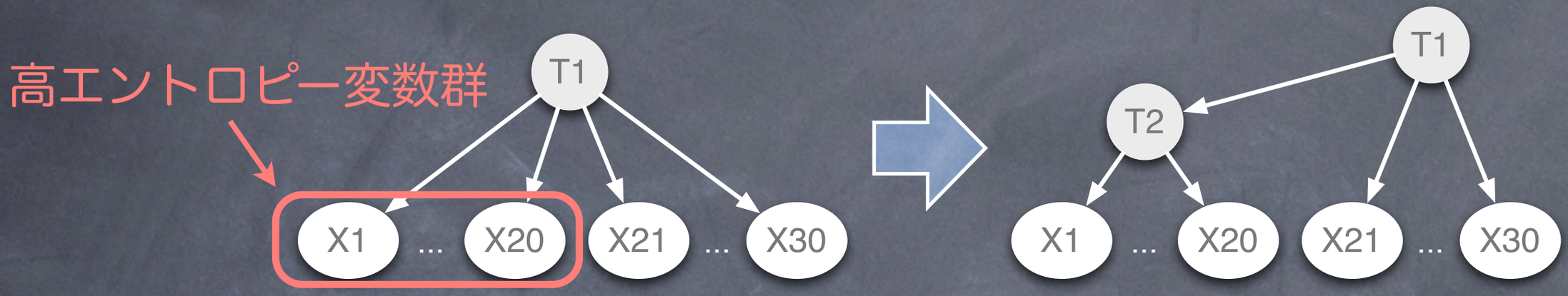
第2段階

- トポロジー情報を利用して，エントロピーの高い変数群とそうでない変数群を判別する手法を提案する
- トポロジー情報(２分木)をトップダウンに見ていくと...
  → ２分される各クラスタは高エントロピーの変数群と低エントロピーの変数群にきれいに分けられている
  → この性質を利用して高エントロピーの変数群を判別する



エントロピーとトポロジー情報を利用した階層構造の推定法の提案

第2段階



- トポロジー情報が示す２分すべき各クラスターの
  エントロピーの平均に差が生じたとき
  - 高エントロピーのクラスターの親に新しい潜在変数
    を追加していくことで階層構造を推定する
- また，パラメタkで許容する階層構造の深さを調整する
  - k：潜在変数の追加対象となるクラスターの数
  - 最適なkの値の決定
    → CSスコア[Cheeseman&Stutz'88]を用いる

提案手法を用いた数値実験と結果

- 先の実験で用いた人工データに提案手法を適用
  - kの値が1~5のときのCSスコアを比較
    (各kの値について3回の推定でのベストスコアを比較)

実験結果

- パラメタkに対してCSスコアは単峰に増減している
  - CSスコアがピークのときの推定結果は，元のモデルの構造(k=3)を完全に推定できている

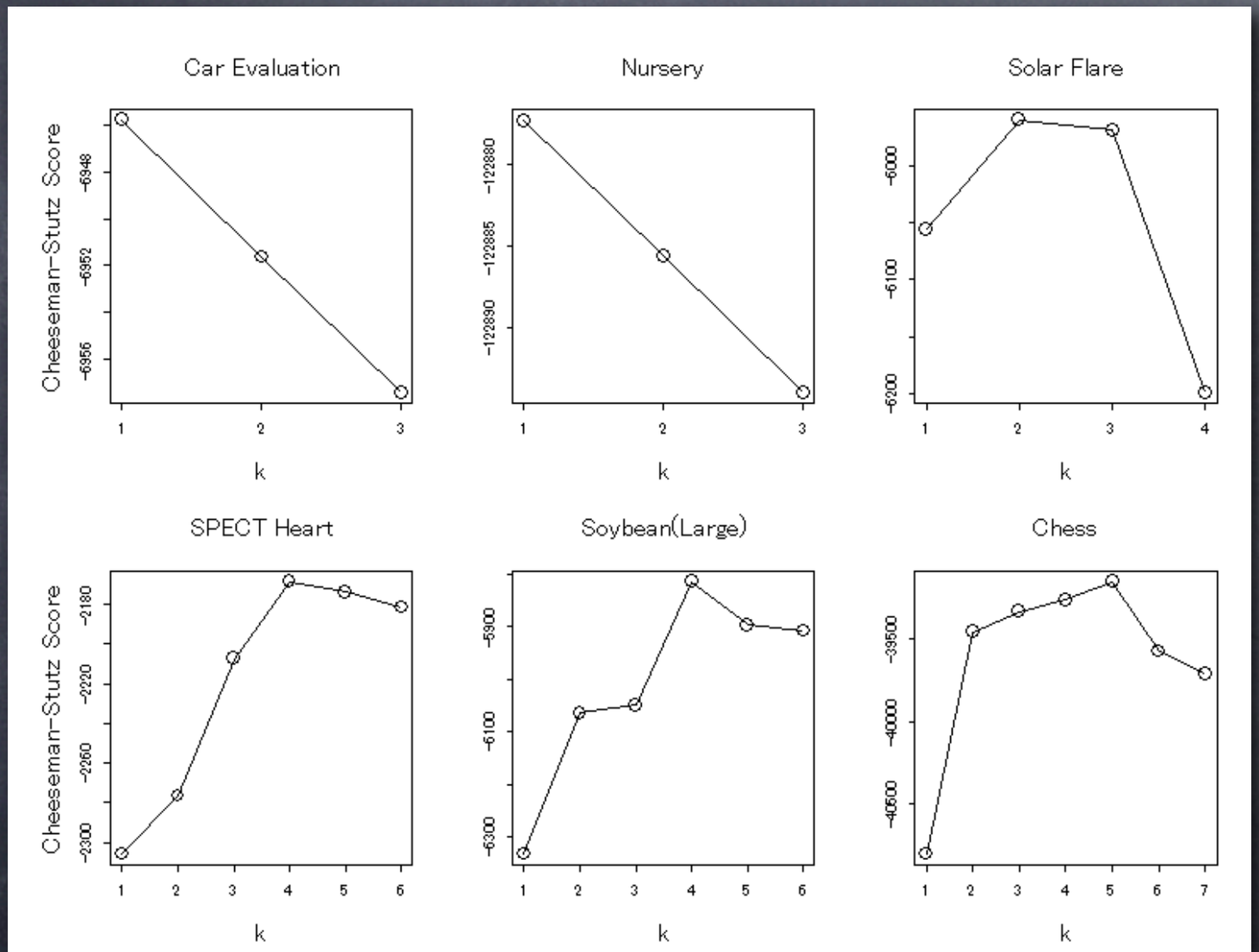
提案手法を用いた実データの分析

- 提案手法を実データに適用し次のことを行う
  - パラメタkに対しCSスコアが単峰に増減するかを検証
  - 推定結果の考察
- 実データにはUCI Machine Learning Repository*で配布されている以下のものを使用（全てカテゴリカルデータ）

データセット名	データ数	観測変数の数
Car Evaluation	1728	6
Nursery	12960	8
Solar Flare	1389	10
SPECT Heart	267	22
Soybean(large)	307	35
Chess	3196	36

* http://archive.ics.ucl.edu/ml

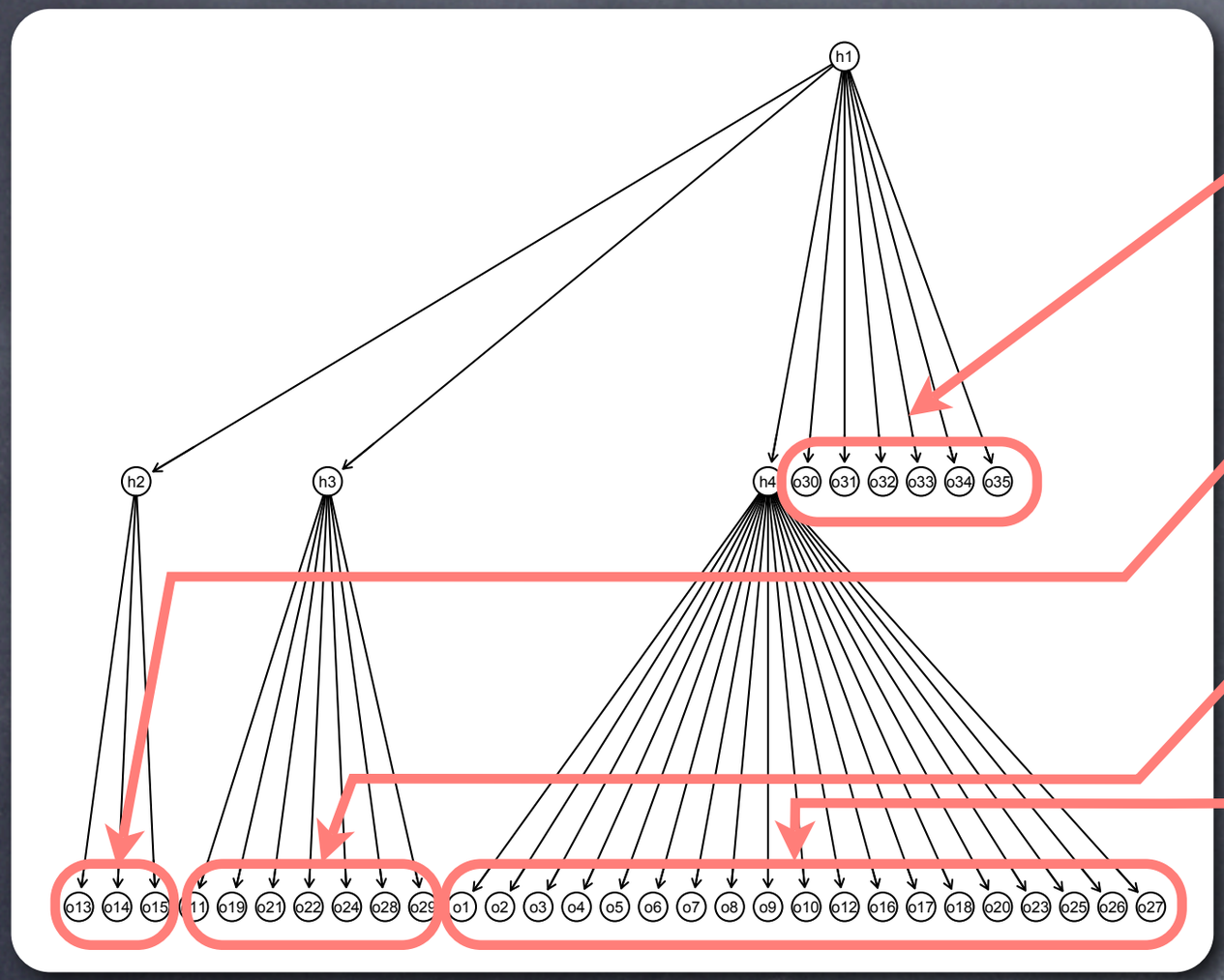
提案手法を用いた実データの分析
パラメタkに対するCSスコアの変化の検証



＊各kでの推定は３回行い，
そのうちの最大のCSスコアを推定値に採用

- すべてのデータで，CSスコアは単峰の曲線を描いている
  - トポロジー情報を利用した階層構造推定は，実データでも安定した結果を返している
- また，観測変数が多いほど潜在変数を階層化させた方がよい傾向が見て取れる

提案手法を用いた実データの分析
Soybeanに対する推定結果の考察



- ✓ 種子,根に関する項目
  カビの生息に関する項目
- ✓ 糸状菌に関する項目
- ✓ 細菌病に関する項目
  茎,枝に関する項目
- ✓ 季節,天候,土地に関する項目
  葉に関する項目

- 推定された各潜在変数は性質の似た項目をまとめている
- 種子,根に関する潜在的なクラスが決まれば，他の３つの潜在変数が局所独立になる階層構造を示している

＊上のグラフは，S-PLUSからdotスクリプトを書き出しGraphvizにより描画.

まとめ

- 本研究では，階層潜在クラスモデルの２段階推定法を提案し評価を行った
  - 数値実験により，第１段階での階層構造のトポロジー推定が有効であることが確認できた
  - 実データでの分析より，提案手法により効率的かつ頑健な推定が可能であることが示された
- また，数値計算に適したS-PLUSを用いることで短期間でプログラムの実装・テストを行う事が出来た

参考文献

- Cheeseman,P. Kelly,J. Self,M. Stutz,J. Taylor,W. Freeman,D.(1988). Autoclass: a Bayesian classification system.
  Fifth International Workshop on Machine Learning,pages 54-64.
- Elidan,G.&Friedman,N.(2003). Information Bottleneck EM Algorithm.
  Proceedings of the Nineteenth Conference on Uncertainty in Artificial.
- Elidan,G.&Friedman,N.(2005). Learning Hidden Variable Networks:The Information Bottleneck Approach.
  Journal of Machine Learning Research 6 81-127.
- Li,M. Chen,X. Li,X. Ma,B. Vitanyi,P. (2002). The similarity metric.
  E-print, arXiv.org/cs.CC/0111054.
- Pearl,J.(1988). Probabilistic Reasoning in Intelligent Systems:Networks of Plausible Inference. San Mateo,CA.: Morgan Kaufmann Publishers.
- Rose,K. (1998). Deterministic annealing for clustering,compression,classification,regression,adn realated optimization problems.
  Proc. IEEE, 86:2210-2239.
- Zhang,N.L.,Nielsen,T.D.,& Jensen,F.V.(2003). Latent variable discovery in classification models.
  Artificial Intelligence in Medicine,30:3,283-299.
- Zhang,N.L.(2004). Hierarchical latent class models for cluster analysis.
  Journal of Machine Learning Research,5:6,697-723.

Appendix:
実装したプログラムの簡単な実行例

- 提案手法により階層潜在クラスモデルの推定する

```
 # 階層構造を推定せず，クラス数のみを推定する場合
> result <- HLC.IBEM(data=data,c.flag=TRUE)
 # 階層構造(k=2)とクラス数の両方を推定する場合
> result <- HLC.IBEM(data=data,c.flag=TRUE,h.flag=TRUE,k=2)
> result
Learning Hierarhical latent class model by Information Bottleneck EM.
=== parameters of HLC.IBEM ===
*** gamma.steps: 30
*** learning cardinality: YES
*** learning hierarchy: YES
…
```

- アニーリング過程での相互情報量の変化をプロットする

```
 > plot.MI.data(result,type=1,distace="MI.XT")
 # 条件付エントロピーをプロットする場合は，type=2, distance="H.XT"を指定
```

Appendix:
実装したプログラムの簡単な実行例

- 人工データに用いたモデル”synthetic.hlc”からギブスサンプリングを行う

```
 # サンプルの先頭から3割をburn-inとしてを棄却し，500個のサンプルを得る
> samples <- sample.HLC(synthetic.hlc,n=500,burn.in=0.3)
```

- 階層潜在クラスモデルの周辺分布，事後分布の計算を行う

```
 # synthetic.hlcの各変数の周辺確率を求める場合
> bp.result <- HLC.BP(synthetic.hlc)
 # すべての観測変数が因子”1”をとったときの各潜在変数の事後確率を求める場合
> bp.result <- HLC.BP(synthetic.hlc,data=rep(1,30))
> bp.result
Result of Belief Propagation.
*** number of evidences: 1
*** calculation time: 0.009 0.001 0.011 0 0
```

- Graphviz用のdotスクリプトを書き出す

```
 # synthetic.hlcのdotスクリプトを”synthe.dot”として書き出す
> write.hlc.graph(synthetic.hlc,file=”synthetic.dot”)
```