資料　２－２
科学技術・学術審議会
先端研究基盤部会
数学イノベーション委員会
（第２回）H23.9.22

大量・複雑データに埋め込まれた規則性や概念・知識の発見とその応用

平成２３年９月２２日

大阪大学産業科学研究所
鷲尾　隆

データマイニング

- 大量・複雑なデータから計算機を駆使して
  有用な規則性・概念・知識を発見
  - アルゴリズミックなパターン探索
  - クラスタリングや分類・判別・
    回帰などの機械学習・
    パターン認識・統計的解析
  - ベイズ推定を含む種々の
    統計的・確率的推定

  等々、数千種類の応用解析手法の総称

- ハンドブック
  - Handbook of Data Mining and Knowledge Discovery, by Willi Klosgen, Jan M. Zytkow, and Jan Zytkow, Morgan Kaufmann, New York (2002)
  - Handbook of Data Mining, by Sanjay Ranka, Chapman & Hall/Crc Computer& Information Science Series (2007)

我々の研究室のテーマ

- 背景
  - 観測機器・センサーの発達

  - ユビキタスセンシングの普及

同時計測された
多数項目のデータ

⇩

高次元
ベクトルデータ

複雑な関係の対象
に関するデータ

⇩

グラフ構造データ

我々の研究室のテーマ

- 基礎研究

  高次元データからの統計量やモデルの推定
  - 高次元データからの統計的推定
  - 高次元データ解析のための組合せ論的最適化

  複雑なデータからのグラフ構造の推定・発見
  - 非ガウス・非線形な統計的因果推論
  - 大量のグラフからの知識発見

- 応用研究
  - 遺伝子発現・タンパク質生成ネットワークの推定
  - データからの希少シナリオの確率分布推定（化学反応）
  - 患者治療履歴データからの新薬候補の発見手法の開発

  等々

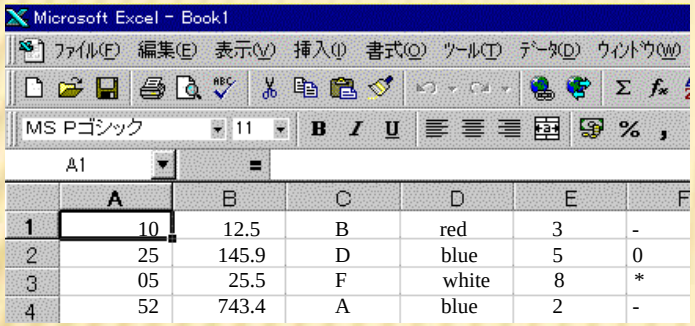
いずれも数学・数理科学的アプローチの研究

例：高次元データ解析のための組合せ論的最適化(1)

- データからの分類・判別・回帰式等のモデリング
  （基本的な機械学習・統計的モデリング問題）

説明変数: $x_1, x_2, \ldots$, 目的変数:y

事例



	A	B	C	D	E	F
1	10	12.5	B	red	3	-
2	25	145.9	D	blue	5	0
3	05	25.5	F	white	8	*
4	52	743.4	A	blue	2	-

推定モデル

$$\hat{y} = f(X_s)$$

$$X_s \subseteq X = \{x_1, x_2, \cdots, x_m\}$$

目的関数(対数尤度やAIC等)

$$\log p(y|\hat{y}) \propto -\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

- 問題

  尤度など目的関数最大となる X_s と f を求めたい.

  + X_sを決める問題：説明変数選択・モデル選択問題
  + fの係数を決める問題：係数フィッティング・探索問題

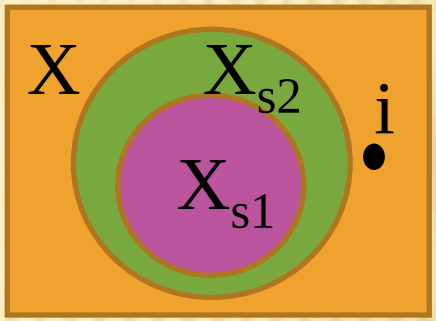
例：高次元データ解析のための組合せ論的最適化(2)

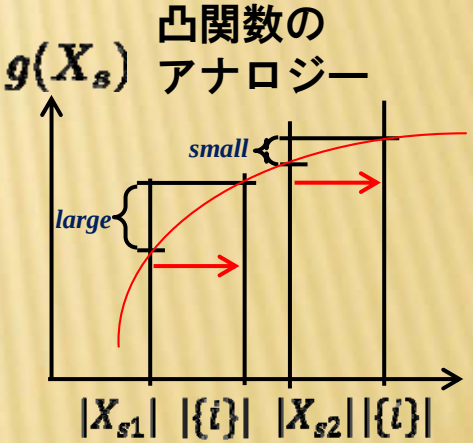
- 劣モジュラ関数
  - 集合関数：集合を引数とするスカラー関数

$$g(X_s) \quad s.t. \quad X_s \subseteq X$$

  - 劣モジュラ性

$$g(X_{s1} \cup \{i\}) - g(X_{s1}) \geq g(X_{s2} \cup \{i\}) - g(X_{s2})$$





- 具体例
  - 経済における「限界効用逓減効果」や「規模の経済性」
  - モデリング目的関数(対数尤度)

$$g(X_s) = \log p(y|\hat{y}) \propto -\frac{1}{N}\sum_{i=1}^{N}(y_i - f(X_{si}))^2 \quad s.t. \quad X_s \subseteq X$$

例：高次元データ解析のための組合せ論的最適化(３)

- 説明変数が1000個の場合，XからX_sの選び方は2^{1000}通り！
  （宇宙にある原子の個数は2^{300}個くらい）
- 組み合わせ爆発問題(NP-困難)を回避する従来の近似法
  正則化：説明変数が増えないようにペナルティー項を導入
  $$\tilde{g}(X_s) = \log p(y|\hat{y}) - ||X_s|| \Rightarrow$$ 所詮近似

- **本来はすべての組み合わせの中で最良のX_sを選びたい。**
  我々の研究：劣モジュラ関数最大化問題 (NIPS2009,2011)
  限られた説明変数個数$|X_s| \leq k$の下で，劣モジュラ関数
  $g(X_s) = \log p(y|\hat{y})$を最大にする$X_s$を効率的に**完全探索**。
  - 劣モジュラ関数の区間連続関数近似(*Lovász extension*)
  - カッティング・プレーンと上下界計算により枝刈り(Brach and Bound)

例：高次元データ解析のための組合せ論的最適化(4)

- 文書分類への適用

多数の文書ファイルをその出現単語を
説明変数として文書を表す五つに分類

従来のGreedy法と
本手法の精度を比較

（カテゴリ）		（選ばれた単語）	（利得）	（精度）
‘earn’	Greedy	vs, dividend, qtr, split, **writedown**, **payout**, rev, **frederick**, **nonperform**, **startup**	0.615	0.834
	SubmoCut	**profit**, split, vs, qtr, **earn**, dividend, **payabl**, **auditor**, **restat**, rev	0.653	0.816
‘acq’	Greedy	cyclop, twa, unsolicit, purol, chemlawn, tfb, ferruzzi, bor, spc, chalmer	0.187	0.788
	SubmoCut	cyclop, twa, unsolicit, purol, chemlawn, tfb, ferruzzi, bor, spc, chalmer	0.187	0.788
‘money-fx’	Greedy	bank, dollar, sai, pct, currenc, blah, **treasuri**, **new**, **two**, **prior**	0.331	0.902
	SubmoCut	**dlr**, **year**, currenc, pct, bank, **rate**, **record**, sai, blah, dollar	0.335	0.924
‘grain’	Greedy	wheat, grain, corn, **maiz**, rice, **barlei**, year, **april**, **three**, **period**	0.279	0.951
	SubmoCut	**dlr**, **tonn**, year, wheat, corn, **washington**, grain, **agricultur**, **ct**, **offici**	0.290	0.937
‘crude’	Greedy	oil, **dlr**, **sai**, **barrel**, **total**, **unit**, **two**, todai, **y**, **year**	0.269	0.952
	SubmoCut	**mln**, **april**, oil, **pct**, **compani**, **report**, todai, **state**, blah, **crude**	0.271	0.953

全体の精度が向上

応用数学における劣モジュラ関数最適化の現状

- 劣モジュラ関数最小化問題(多項式時間)の研究は進んでいる。
  - 最初のアルゴリズム Grotschel, Lovasz and Schrijver (1981)
  - その後の組合せ的アルゴリズム

    Schrijver (2000)

    Iwata, Fleischer and Fujishige (2001) (京大数理解析研究所)

    Iwata (2002), Fleischer and Iwata (2003), Iwata (2009)
- しかし、劣モジュラ最大化問題(NP-困難)はニーズはあるがほぼ近似アルゴリズム研究のみである。

  Nemhauser, Wolsey and Fisher (1978)

  Feige, Mirrokni Vondrak (2007), Vondrak (2008)

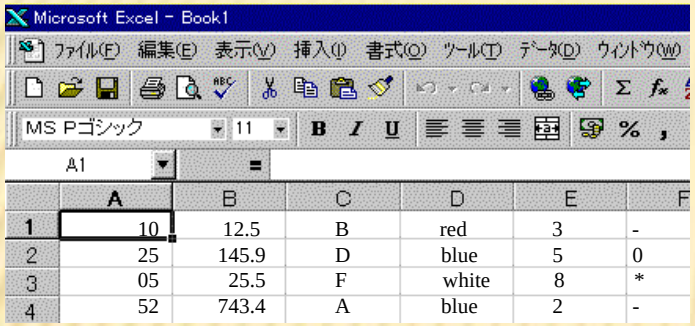
⇒ 最も離散数学者との連携・協力が期待される分野の１つ

例：非ガウス・非線形性な統計的因果推論(1)

- 統計的因果推論

変数: $x_1, x_2, \ldots, x_n$

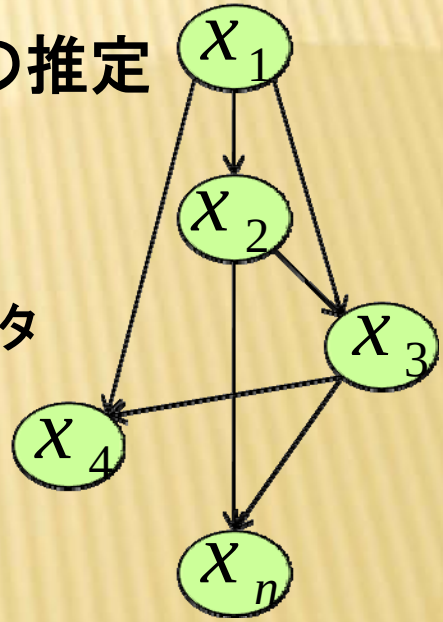
事例



	A	B	C	D	E	F
1	10	12.5	B	red	3	-
2	25	145.9	D	blue	5	0
3	05	25.5	F	white	8	*
4	52	743.4	A	blue	2	-

データ生成過程の推定

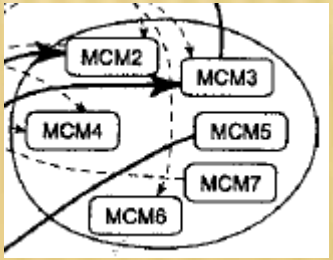
どのような順序で
いずれの変数から
いずれの変数のデータ
が生成されたか？



- 応用
  - 経済・社会学 [H.A.Simon (1977)]

    調査データ ⇨ 都心と住居の距離 D ⇄? 収入M
  - バイオオインフォマティクス

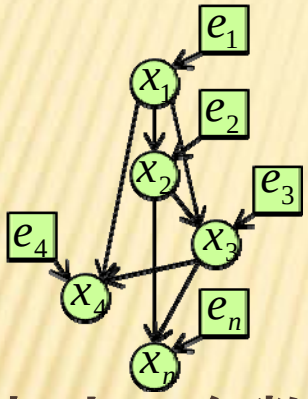
    実験データ ⇨ 遺伝子発現間の決定関係



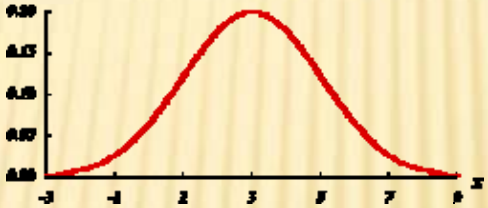
例：非ガウス・非線形性な統計的因果推論(2)

従来法

- 観測変数 $x_1, \cdots, x_n$がガウス分布すると仮定



各観測変数 x_iは非観測外乱 e_iにより揺らぐ。
揺らぎはすべてガウス分布と仮定。



- これまで多数の推定手法が提案されて来た。
  - + PC algorithm (Spirtes et al., 2000)などの制約ベース
  - +GES algorithm (Chickering, 2002)などのスコアベース
- 過去４０年の問題点：一部モデルを同定できない。

? Model 1 $x_1 := b_{12}\, x_2 + e_1$ x1 ← x2

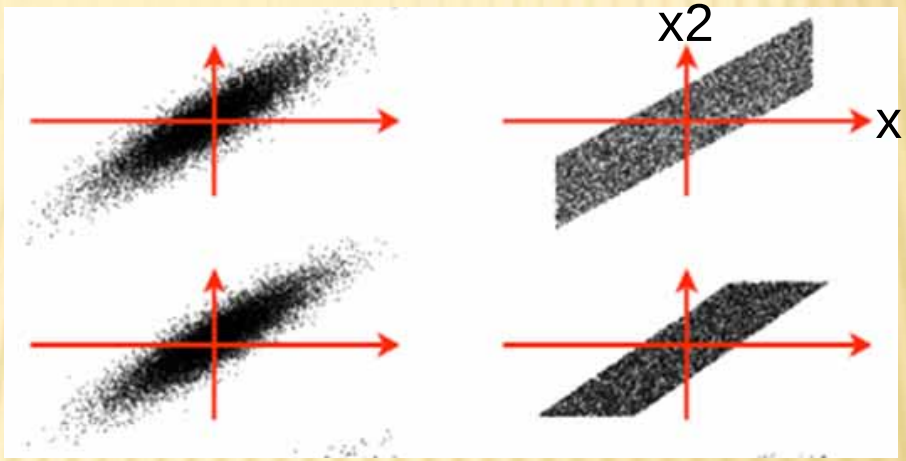
Model 2 $x_2 := b_{21} x_1 + e_2$ x1 → x2

例：非ガウス・非線形性な統計的因果推論(3)

- 我々の研究(UAI2005,09,11, JMLR2011)
  - 外乱が非ガウスだと一意に同定可能であることを示した。

Model 1
$x_2 := \beta x_1 + e_2$

Model 2
$x_1 := \beta x_2 + e_1$



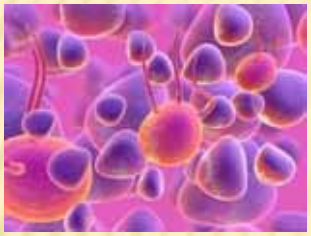
Gaussian　　Non-Gaussian

- その後Hoyer et al (2008), Zhang and Hyvarinen (2009)が、外乱がガウスでも一部の非線形モデルでは同定可能であることを示した。

$$x_2 = f(x_1) + e_2 \quad \neq \quad x_1 = g(x_2) + e_1$$

例：非ガウス・非線形性な統計的因果推論(4)

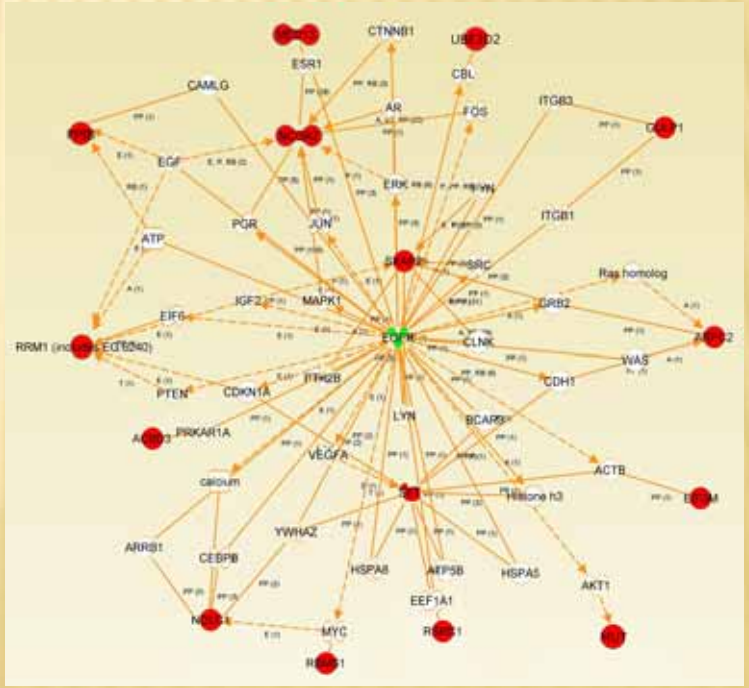
- 遺伝子発現ネットワーク推定への適用
  - 乳がん細胞にホルモンを投与時の遺伝子発現データ



ホルモンが細胞遺伝子を刺激

1000個の遺伝子発現レベルを測定

我々の開発手法
Direct-LiNGAM
を適用



これまで困難だった
遺伝子発現ネットの
一意同定に成功

これまでの統計的因果推論研究の問題点

- 統計的因果推論はPearson and Moul (1927), Simon (1953), Blalock (1961)等の時代からガウス分布(２次統計量)の世界で研究が行われて来た。最近ではPearl, Glymour (1980-)等。我が国ではKano (1983-86), Miyakawa (1997-), Kuroki(2000-)等。

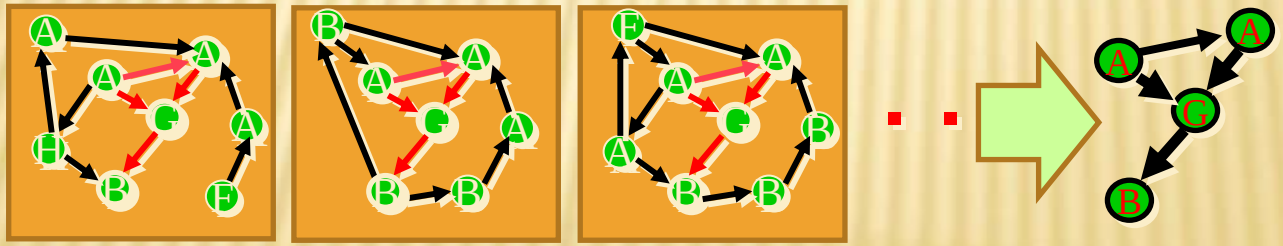
- これに対してFrechet, Fisher and Tippett, von Mises, Gnedenko (1920-1940)等，我が国ではKuriki (2000-)等の極値統計(非ガウス統計)や独立成分分析Jutten and Herault (1991)等，我が国ではAmari (1995-)の統計理論と長い間交わることがなかった。

⇨ 統計学者との更なる連携・協力が期待される分野の１つ

その他データマイニングと数学・数理科学の接点(1)

- 高次元データからの統計的推定（次元の呪いの克服）
  - 例　高次元データベクトル分布は超球上に縮退する。

- グラフ構造マイニング

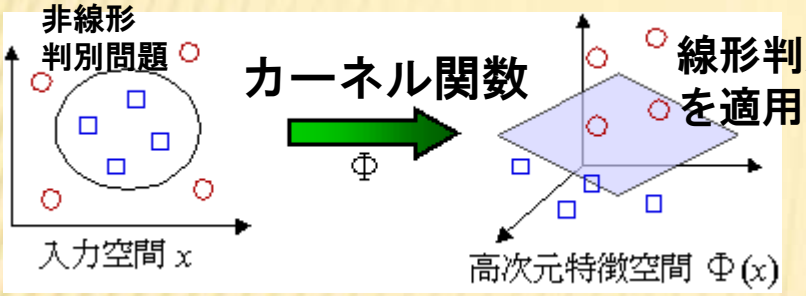


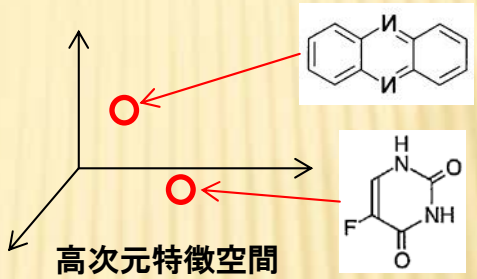
数学・数理科学では部分グラフ同形問題は研究されている

パターン等の列挙や探索と離散数学の連携・協力研究は不足

その他データマイニングと数学・数理科学の接点(2)

- カーネル関数やデータ空間の埋め込み





どの空間に
どの空間を
写像可能か？
効率的な写像は？

空間の性質や写像に関する数学・数理科学が重要

- ベイズ推定をはじめとする確率・統計的推定・推論

分類・判別・回帰の確率モデル $\hat{y} = f(X_s)$ の拡張として、
特定の条件 C での $p(y|X_s, C)$ を推定する研究が盛んである。
しかし、例えば C を容易に導入できない（例：ある病気 C とその時の遺伝子発現パターン X_s の関係が単純に決まらない）場合どうするか？

⇒ **種々のモデルシミュレーション技術と確率・統計的推定・推論の融合が必要**

データマイニングの立場からの期待

- 機械楽学習・統計的データ解析を含むデータマイニングと数学・数理科学の接点
  - 確率・統計分野の基礎的発展とその応用
  - 離散数学の基礎的発展とその応用
  - 位相空間や写像・関数論の基礎的発展とその応用
- 何をおいても基礎研究が重要である。
- Plus その応用の視点もあればギャップが埋まる。

- 数学者や数理科学者と応用視点を共有するには、出会いと連携・協力が必要。
- これまでの研究成果もこのような出会いに基づく。

これまでの連携・協力の機会

- 我々の分野で組織的に行われてきたこと
  - プロジェクト
    - 特定領域研究(A)「発見科学」H9年度～H12年度
    - 特定領域研究(B)「情報洪水時代におけるアクティブマイニングの実現」H13年度～H16年度
  - 学会活動
    - 人工知能学会
      データマイニングと統計数理研究会 H18年度～H21年度
    - 電子情報通信学会
      情報論的学習理論と機械学習研究会 H10年度～

- 異分野研究者の出会いの場を提供して来た。
- しかし、どうしても情報系・工学系研究者に偏っており、数学者・数理科学者と議論する機会が少ない。