ビッグデータに基づくモデリング

―生体・医療への適用を例として―

鷲尾　隆
大阪大学産業科学研究所

Modeling Based on Big Data

- Applications to Biogenic and Medical Domains-

Takashi Washio
The Institute of Scientific and Industrial Research,
Osaka University

要旨:

ビッグデータの解析では、データ件数の多さ以上に、変数やその組み合わせの多さに起因する様々な探索上の困難や統計的誤差の増大をどう扱うかが重要であることを述べる。そして、多チャンネル計測の脳波データから脳内各部位の影響関係を推定する問題や、多数の検査・治療項目や投薬項目を含む医療治療記録データから患者の治療効果を推定するモデルを得る問題など、生体・医療分野の解析問題を例として取り上げ、最近の我々の研究内容を交えながら、これら最新の手法の幾つかとその実適用成果について紹介する。

キーワード:ビッグデータ, 高次元データ, 統計的機械学習,
統計的因果推論, 生体・医療

ビッグデータとは？

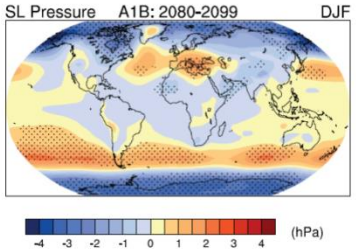
- 一般のリレーショナルデータベースなどでの扱いが困難な程に膨大なデータ
- 現状ではタスクや目的にも依るが、ペタバイトを超えるデータの扱いは難しいことが多い。

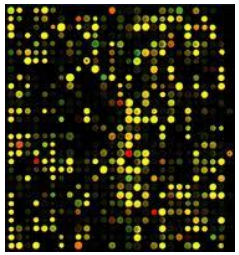
ビッグデータが生まれる背景

- データの収集、蓄積技術の発展
  - インターネット・センサーネットなど情報通信技術
  - ハードディスク・メモリなど情報保存技術

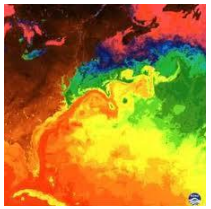
ビッグデータの例

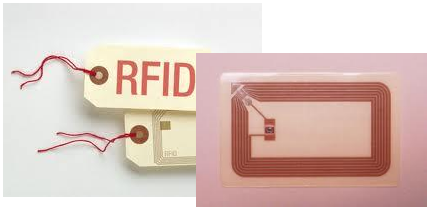
- サイエンス
  - 気象学，ゲノミクス，物理シミュレーション，環境生物学，インターネット検索，経済学，社会科学

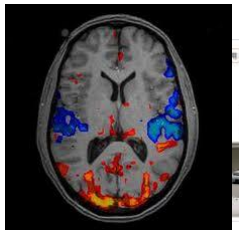




- センシング
  - モバイル機器センサー，リモートセンシング，カメラ，マイクロフォン，ＲＦＩＤ，無線センサー，ソフトウェアのログ







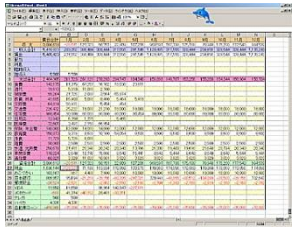
ビッグデータの量のイメージ

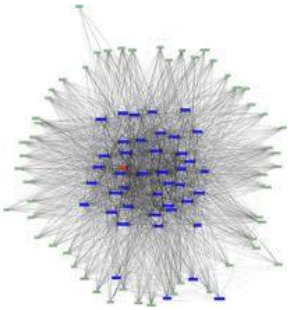
- 製造業液晶工場のログデータ／年　2PB
- 最新ゲノムシーケンサ出力データ／回　5TB

- ＮＴＴソフトウエアイノベーションセンター木村真吾氏資料（2012/4）
  - 世界のインターネットトラフィック　677EB
  - インターネット接続センサー数　10億個
  - Googleの処理量／日　20PB
  - Yahooの処理量／日　135TB
  - ツイッターの処理量／日　7TB
  - facebookの処理量／日　10TB
  - ebayのデータ発生量／日　50TB
  - ニューヨーク商取処理量／日　1TB

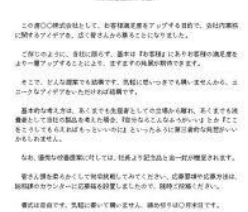
ビッグデータの内容と傾向

- **構造化データと非構造化データ**
  - **表やグラフなど**





  - **文書や画像、動画など**



何れもが急増

- **ここ数年は年率４５％成長**
  - **１０年間で４０倍に！**

ビッグデータの活用事例(1)

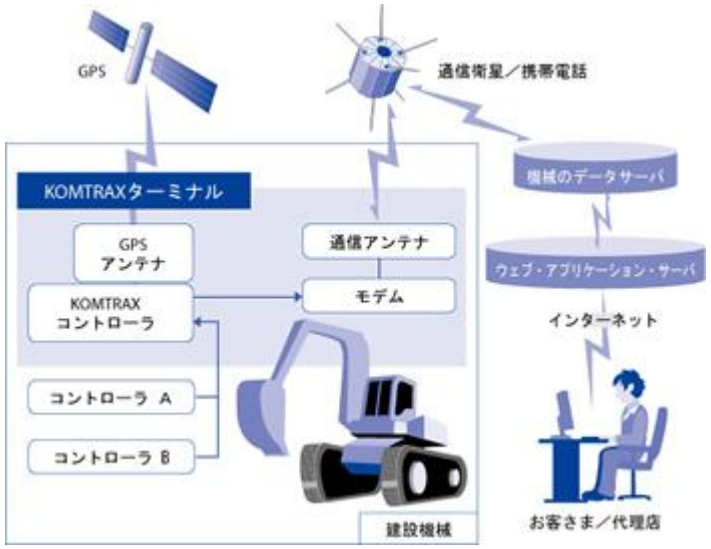
- IT**分野：**Yahoo
  - **検索インデキシング・広告レコメンデーション**
  - **過去数年分のログ数十ペタバイト**
  - Hadoop**環境，** 25000**台規模のクラスタで処理**

ビッグデータの活用事例(2)

- **センサーネットワーク分野：コマツ（建設機械）**
  - KOMTRAX
  - **API経由でのセンサーデータのアップロード・参照**
  - **全世界の自社建設機械の稼働状況をモニタ**
  - **国内62,000台(世界で数十万台)，**
  - **データ処理量／日 1TB?**

インドネシアの石炭鉱山

中国・上海のビル解体現場

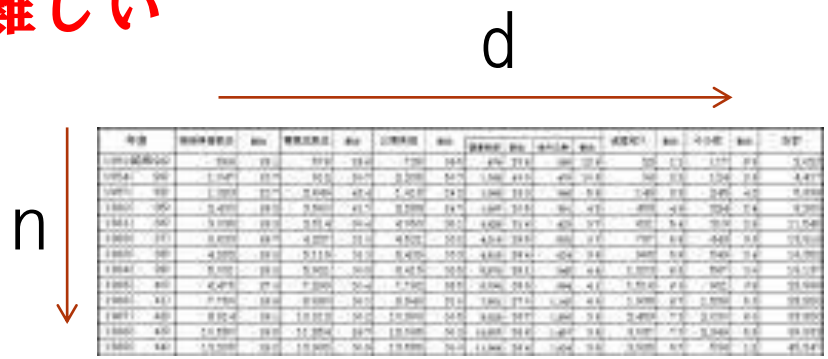


ビッグデータ処理に係る技術

- クラウドコンピューティング
  - グリッド，分散ファイルシステム，大規模記憶装置，プラットフォーム
- モバイルセンサーデータ収集
  - インターネット，API，モバイルOS
- データ処理，分析
  - 超並列データベース
  - NoSQL高速データベース管理システム
  - データマイニング

ビッグデータに関する データマイニングの課題

- **膨大なデータ件数nをどう処理するか？**
  - **マイニングのための高速アルゴリズム開発**
    - **高速事例検索，高速事例類似性計算など**
  - **並列データマイニング処理**
    - **各種分類，クラスタリング，共起分析の並列化**
- **膨大な次元ｄの属性，変数をどう解析するか？**
  - **高次元データマイニング技術**
  - **データマイニング・機械学習としてはこちらの方が本質的に難しい**



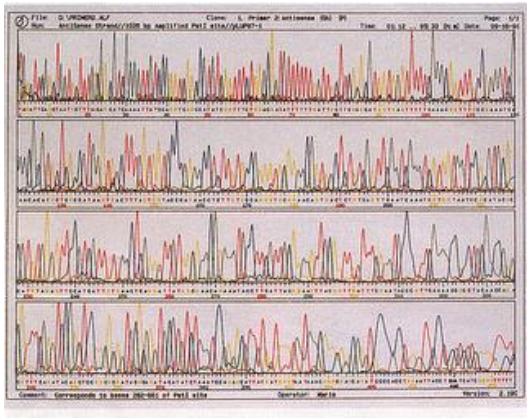
膨大な属性・変数次元を持つデータの例

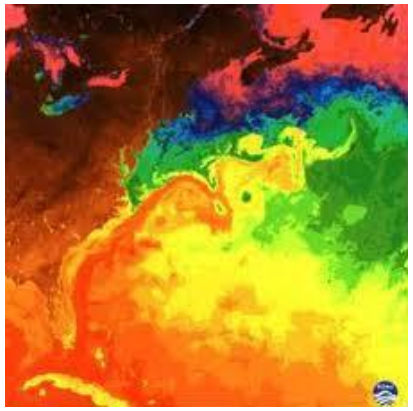
- **ユビキタスセンシング データ**
  - **膨大なセンサを含むパーソナライズドモバイル環境**

    [Dimakopoulos and Magoulas 2006]
- **科学分野のセンシング**
  - **生化学：最近のDNAシーケンサ出力**

    [Steen and Mann 2004]
  - **リモートセンシング：センサ種類と解像度の増大**

    [Lillesand et al. 2007]



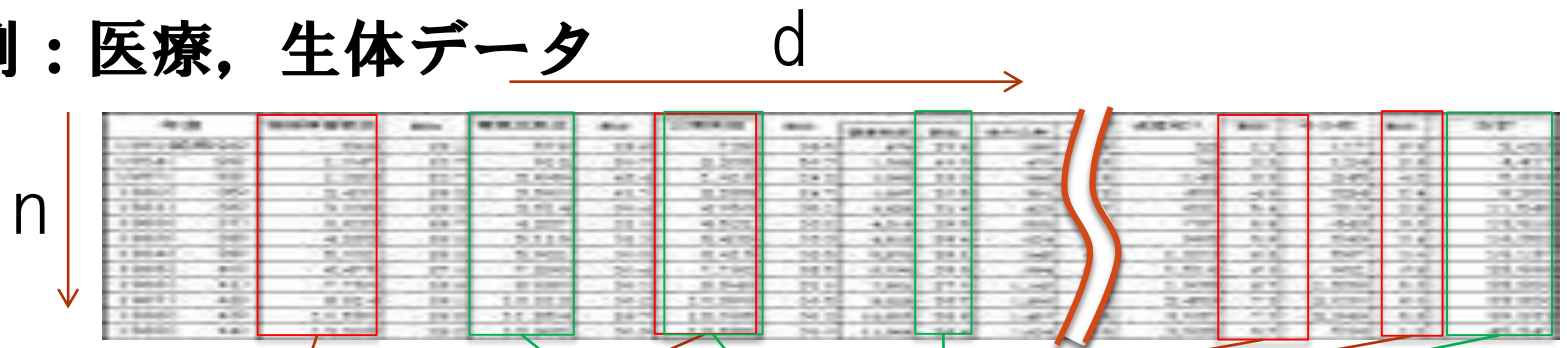


ビッグデータに関する データマイニングの課題（つづき）

- **膨大な次元dの属性，変数から、そこに埋もれる（隠された）属性や変数間の関係（モデル）を発掘する問題**

モデルマイニング

例：医療，生体データ

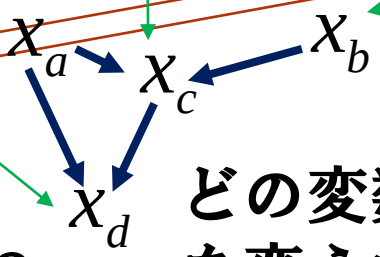


予測モデル

$y = f(x_1, x_2, x_3, x_4)$

どの検査結果変数を使えば、患者の病後ｙを予測できるか？

因果モデル



どの変数が他のどの変数を変えているか？

高次元データからのモデルマイニング

- **対象の将来を予測するモデルのマイニング**

  多数の検査・治療項目や投薬項目を含む医療治療記録データからの患者の治療効果予測モデルのマイニング

- **対象の仕組みを理解するモデルのマイニング**

  多チャンネルの脳波計測データからの脳内部位の影響因果関係モデルのマイニング

患者治療効果予測モデルのマイニング

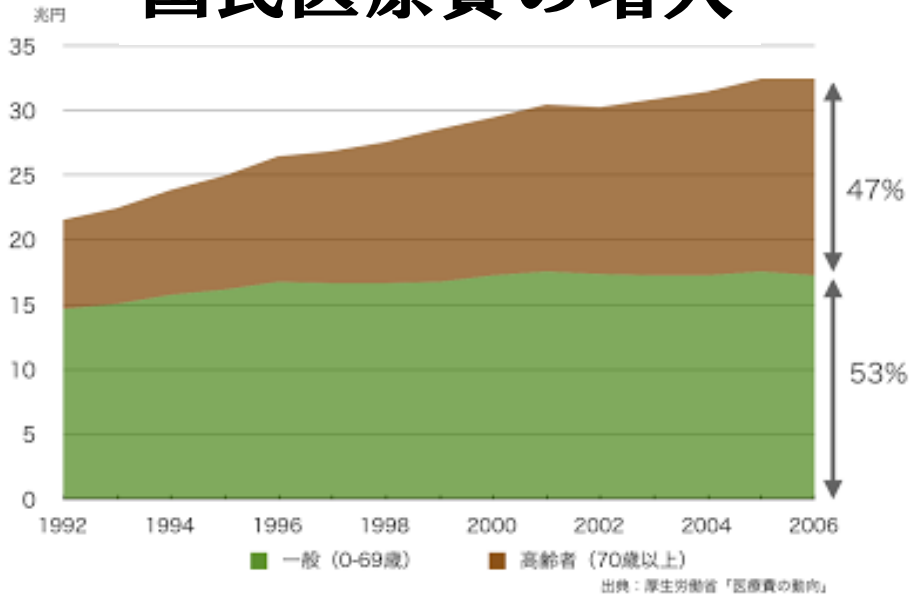
- **大阪大学産業科学研究所と国立循環器病センター臨床研究部・心臓血管内科（北風政史グループ）との共同研究成果**

  A.Yoshida et al., Derivation of a mathematical expression for predicting the time to cardiac events in patients with heart failure: a retrospective clinical study, Hypertens Res. 36(5), pp.450-456 (2013)

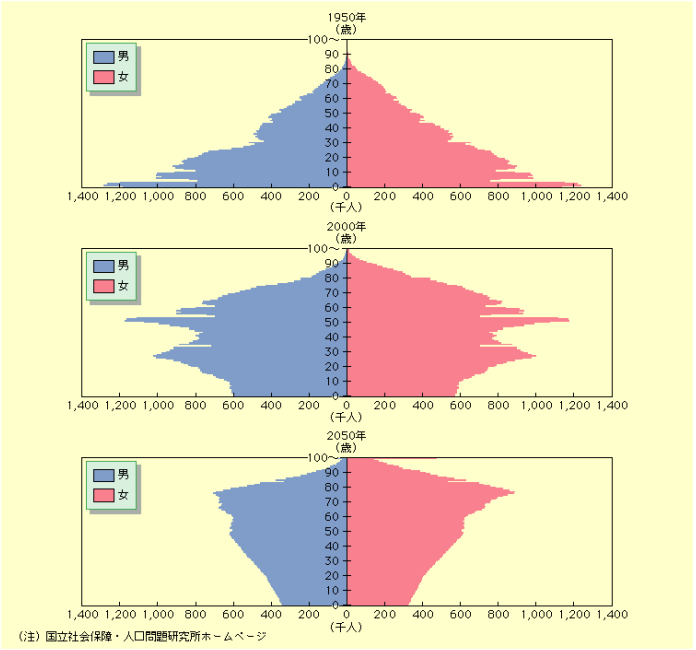
背景

- 高齢化による患者数増大
- 医療の高度化

国民医療費の増大



厚生省省HPより



国土交通省HPより

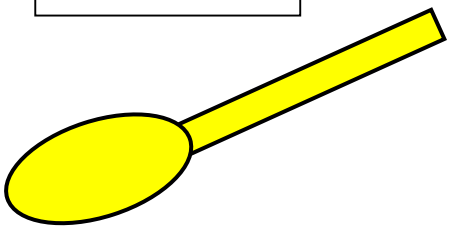
効果の高い治療により
- 国民健康寿命の延長
- 医療の効率化

の同時達成を目指す。

背景

- **治療効果の高い医療を実現するために**
  - **より客観的情報に基づく効果的治療**
  - **ベテラン，若手医師に関わらない治療効果の向上**
- **科学的根拠に基づく医療（EBM）**

経験則



医師個人の経験に基づく「さじ加減」の治療

EBM

Evidence Based Medicine

裏付けEvidenceに基づく治療

$$y = f(x_1, x_2, x_3, x_4)$$

検査・治療・投薬情報から患者の病後ｙを予測するモデルがあれば、治療成績が向上するプランを作れる。

背景

- 心疾患における患者の治療効果ｙ

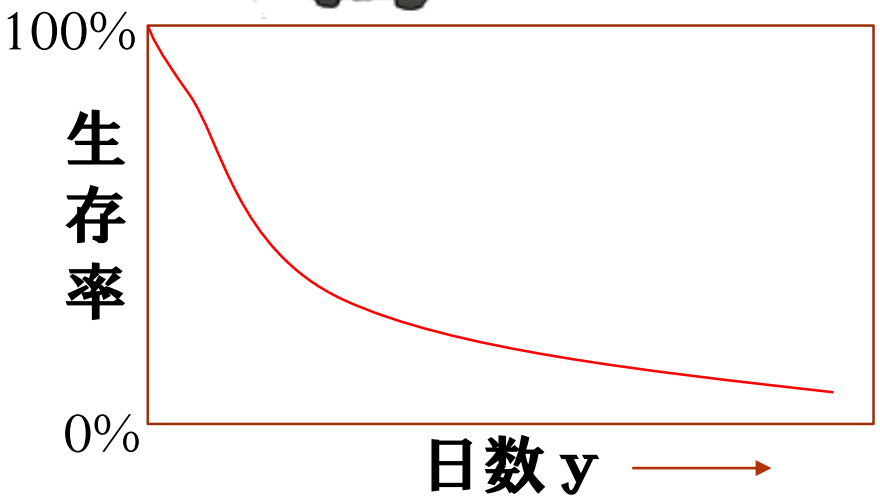
心不全
心筋梗塞
で入院

入院治療　退院

再発
再入院
or 死亡

日数

再発までの日数ｔ



再発日数ｔの期待値や
再発確率を精度よく
見積るモデルが欲しい。

$$\hat{y} = E[t] = f(x_1, x_2, x_3, x_4)$$

背景

- 医療検査の技術・機器・体制の進歩
  - ⇨検査項目の増大

- 医療治療手法・技術の進歩
  - ⇨疾患や症状に応じた治療法の多様化

- 医学・薬学の進歩
  - ⇨投薬種類の多様化

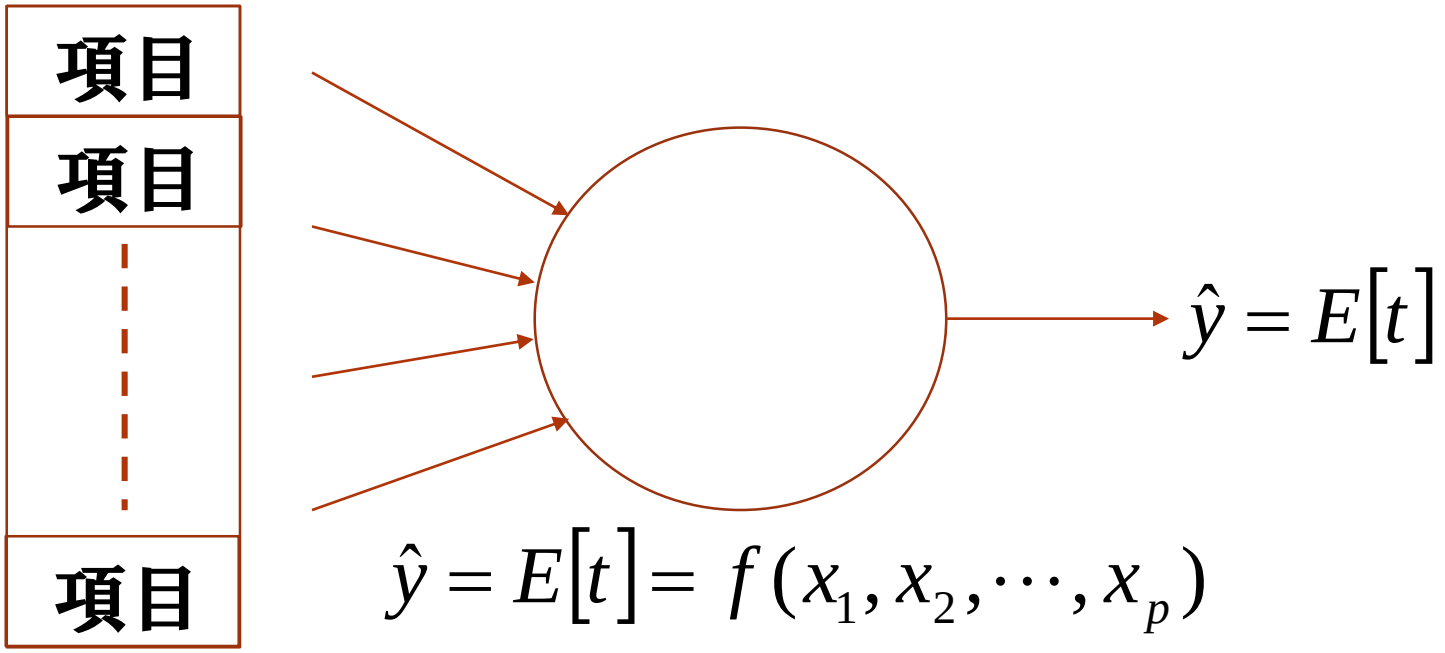
⇨各患者のデータ項目（属性,変数）の増大

国循センターでは検査治療215項目,投薬535項目

項目変数が沢山あり過ぎてモデルが作れない。

解析目的

- **多数の属性,変数項目（750項目）で構成される多数の心不全・心筋梗塞入院患者データから、適切に関係する項目を選択して、次回再発までの日数の期待値を予測するモデルを構築する。**



予測モデルの検討

- **仮定１**

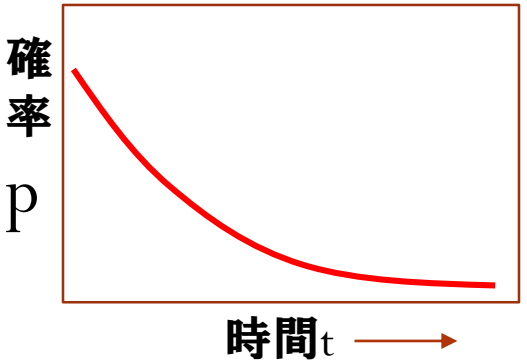
  退院後の患者 i の１日当たりの再発確率 α_i は一定

○　ポアソン過程 (Poisson process)

ランダムに生起する事象を表す基本的な確率過程で、客の到着や故障の発生、個体の出生など様々な現象のモデル化に使われる.

次回の再発時刻 T_i が t になる確率

指数分布 $$p_i(t) = \alpha_i \cdot \exp(-\alpha_i \cdot t)$$



予測モデルの検討

- **仮定２**

  患者 i の１日当たりの再発確率α_iは、近似的に前回入院中の検査，治療，投薬項目の幾つかの線形結合効果によって決まる。

$$\alpha_i = \beta^T X_i + c$$

$X_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]^T$　(x_{ij}**は患者**i**の**j**番目の項目変数，**$p \leqq d$)

$\beta = [\beta_1, \beta_2, \ldots, \beta_p]^T$**：項目変数係数パラメータのベクトル**

c **：項目に影響を受けない基礎再発確率**

予測モデルの検討

- **仮定１，２より**

$$\boxed{p_i(t) = \alpha_i \cdot \exp(-\alpha_i \cdot t)} + \boxed{\alpha_i = \beta^T X_i + c}$$

$$p(X_i, t) = \left(\beta^T X_i + c\right) \cdot \exp\left\{-\left(\beta^T X_i + c\right) \cdot t\right\}$$

患者 i の再発までの日数 y_i の期待値は以下となる。

$$\hat{y}_i = E_i[t] = \int_0^\infty t \cdot \alpha_i \cdot \exp(-\alpha_i \cdot t) \cdot dt = \frac{1}{\alpha_i} = \frac{1}{\beta^T X_i + c}$$

予測モデルのモデリング

- **患者データ**

$$D = \left\{ (Z_i, t_i) \middle| Z_i \in R^d, i = 1, \ldots, N \right\}$$

$z_i = [x_{i1}, x_{i2}, \ldots, x_{id}]^T$ (x_{ij}**は患者**i**の**j**番目の項目変数**)

t_i **：患者**i**の再発日数**

- **患者データの尤度**

$$P(D) = \prod_{i=1}^{N} p(Z_i, t_i) = \prod_{i=1}^{N} \left(\beta^T Z_i + c \right) \cdot \exp\left\{ - \left(\beta^T Z_i + c \right) \cdot t_i \right\}$$

対数尤度を最大化するようにβ**と**c**を求める。**

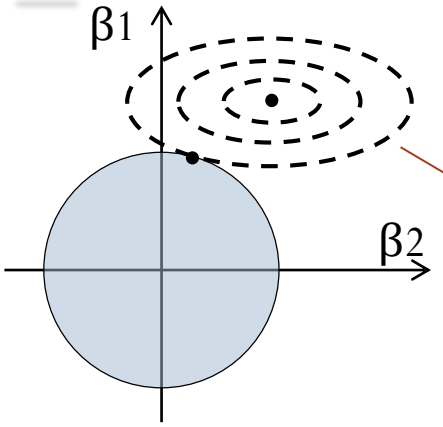
$$\ln P(D) = \sum_{i=1}^{N} \ln\left(\beta^T Z_i + c \right) - \sum_{i=1}^{N} \left(\beta^T Z_i + c \right) \cdot t_i \rightarrow \max$$

予測モデルのモデリング

- **すべての項目変数からなる$Z_i \in R^d$からt_iに関係する項目変数のみからなる$X_i \in R^p \subseteq R^d$を選ぶ必要がある。**
- **$\beta = [\beta_1, \beta_2, \ldots, \beta_p]^T$の中で$t_i$に無関係な$x_i$の係数$\beta_i$をゼロにするペナルティー項(正則化項)$L_q$を導入**

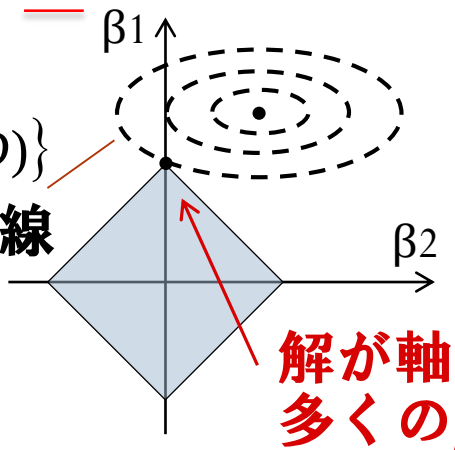
$$-\ln\{P(D)\} + L_q \rightarrow \min \qquad L_q = \left(\sum_{i=1}^{d} |b_i|^q + |c|^q \right)^{1/q}$$

L_2による正則化



$-\ln\{P(D)\}$ **の等高線**

L_1による正則化



解が軸に乗りやすく多くのβ_iがゼロに

予測モデルのモデリング

- L1-**正則化を用いる**

$$-\ln\{P(D)\}+L_1=-\sum_{i=1}^{N}\ln\left(\beta^T Z_i+c\right)+\sum_{i=1}^{N}\left(\beta^T Z_i+c\right)\cdot t_i+\lambda\cdot\left\{\sum_{k=1}^{d}\left|\beta_k\right|+\left|c\right|\right\}$$

$\rightarrow \min$

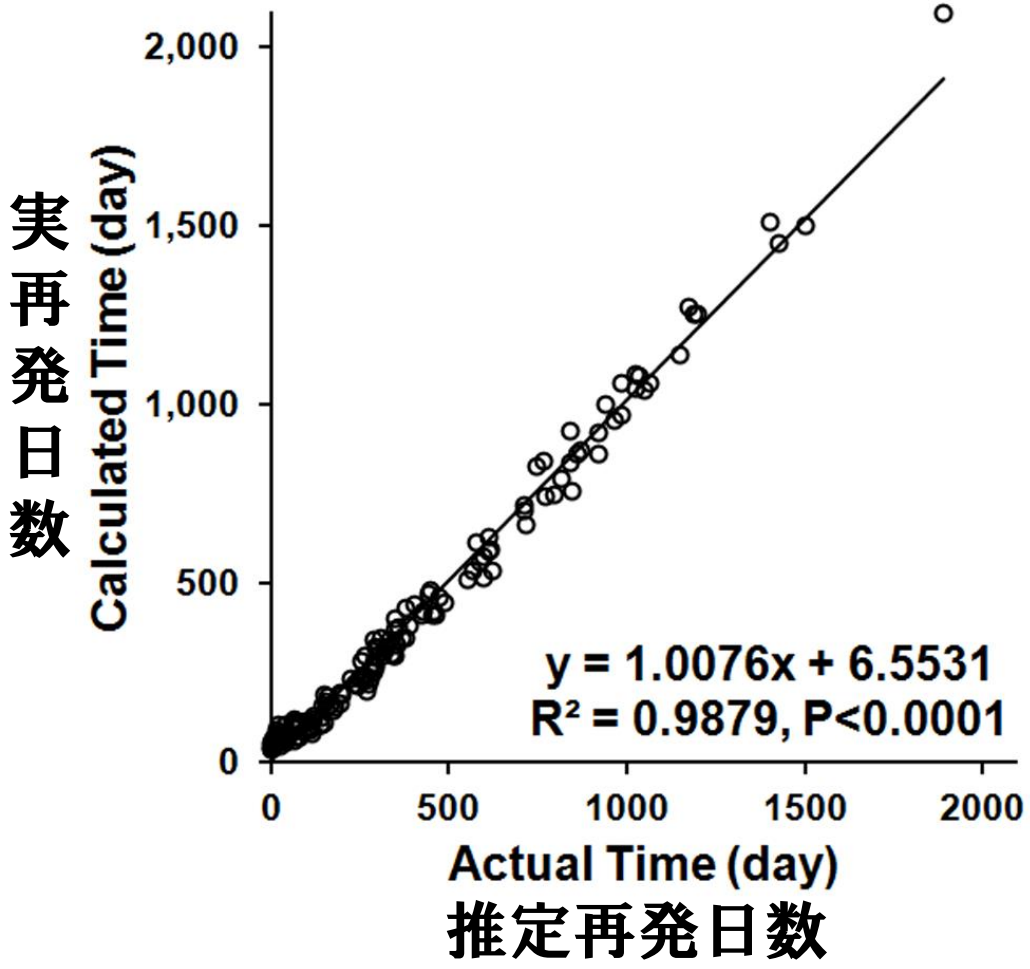
- **最適化アルゴリズム**
  - **近接勾配法** (proximal gradient method)

$$\beta(t+1)=\arg\min_{\beta}\left(\left\langle-\nabla_\beta\ln\{P(D)\},\beta-\beta(t)\right\rangle+L_1+\frac{1}{2\eta_t}\left\|\beta-\beta(t)\right\|_2^2\right)$$

  - **他により収束性のよい**
    双対拡張ラグランジュ法 (dual augmented Lagrangian)
    なども知られている。

モデルによる推定結果

- **推定再発日数期待値と実再発日数の比較**



予測モデルに選択された属性項目変数

性別

基礎疾患

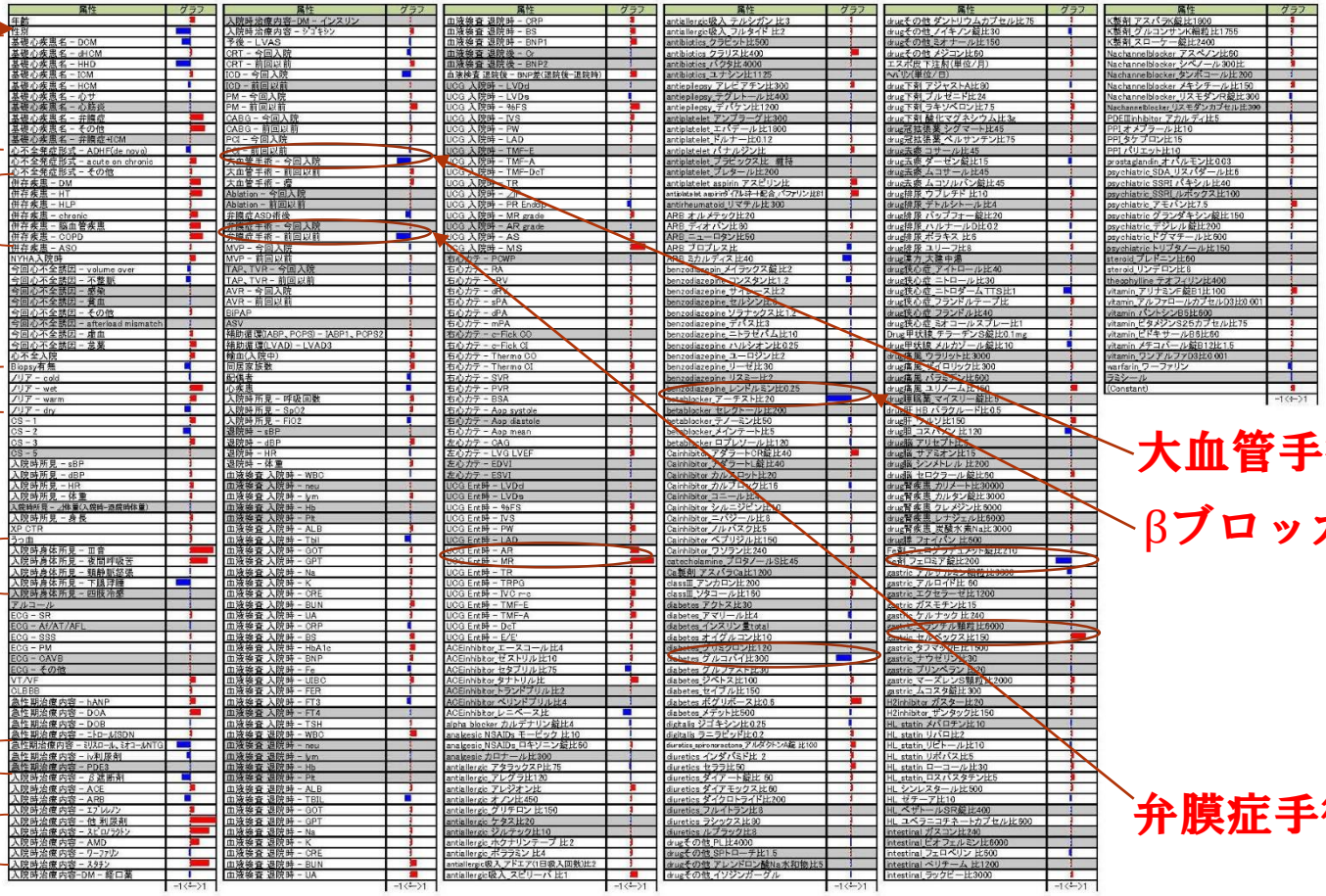
併存疾患
脳血管疾患
など

血流・鬱血

入院時
身体所見

急性期
治療内容

入院時
治療内容



大血管手術

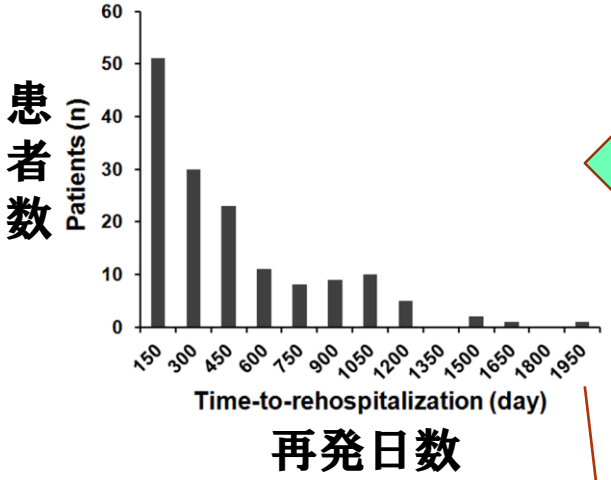
βブロッカ

弁膜症手術

薄い網掛けは選択されなかった変数
再発確率を赤は増加、青は減少させる効果

予測モデルとデータの適合度

- **再発日数データは患者個別の指数分布の混合なはず**



$$\sum_{i=1}^{N} p(X_i, t) =$$

$$\sum_{i=1}^{N} \left(\beta^T X_i + c\right) \cdot \exp\left\{-\left(\beta^T X_i + c\right) \cdot t\right\}$$

$$\text{Gスコア } G = 2\sum_{k=1}^{CL} n_k \ln \frac{n_k}{m_k}$$

階級幅(日数)	棄却レベル	Gスコア	検定結果
100	21	14.1	棄却されない
110	19.7	15.6	棄却されない
120	18.3	23	棄却される
150	15.5	9.7	棄却されない
180	14.1	10.2	棄却されない

帰無仮説：「観測度数は混合ポアソン過程である」

患者治療効果予測モデルのマイニング まとめ

- 「国民健康寿命の延長」と「医療の効率化」の同時達成のための**科学的根拠に基づく治療（ＥＢＭ）**の実現を目指す、大阪大学と国立循環器病センターの共同研究

- 心疾患について、過去の患者治療履歴データから治療成績向上プラン作成の基礎となる**患者個別の再発日数期待値や再発確率の推定モデル**を構築

- 心疾患の再発がポアソン過程であること、再発確率には多くの要因が絡む中で**幾つか大きな影響を与える治療や投薬**が明らかになった。

高次元データからのモデルマイニング

- **対象の将来を予測するモデルのマイニング**

  多数の検査・治療項目や投薬項目を含む医療治療記録データからの患者の治療効果予測モデルのマイニング

- **対象の仕組みを理解するモデルのマイニング**

  多チャンネルの脳波計測データからの脳内部位の影響因果関係モデルのマイニング

対象の仕組みを理解するモデル

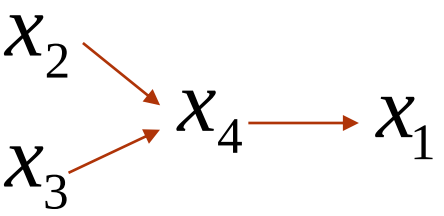
- **因果構造モデル**

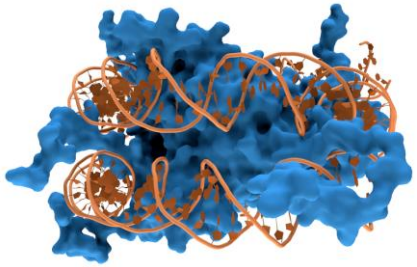
観測変数

どの変数がどの変数を変化させているか？（対象の因果構造）

x_1 x_2

x_3 x_4

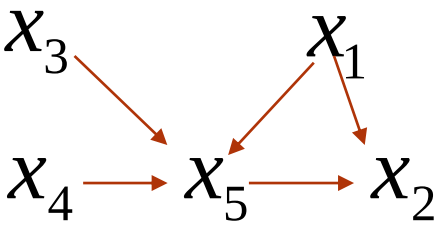




x_1 x_4

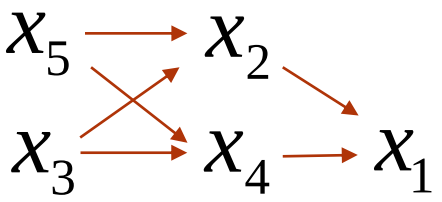
x_3 x_5

x_2



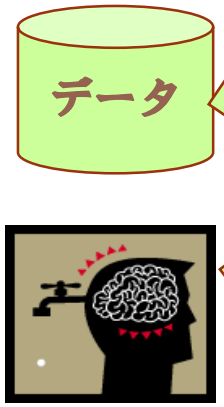
x_1 x_2

x_3 x_4 x_5

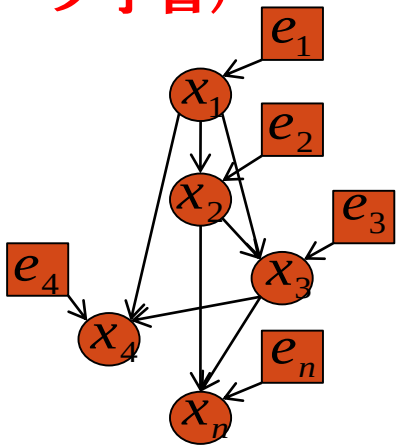


統計的因果推論（グラフィカルモデリング）の概観

因果構造の学習（構造学習）や因果強度の学習（パラメータ学習）



データと人間の背景知識から変数間の決定関係を推定



因果関係を表す有向非巡回グラフ
DAG (Directed Acyclic Graph)

共分散構造分析

$$
\begin{aligned}
x_1 &= e_1 \\
x_2 &= b_{21}x_1 + \cdots + e_2 \\
x_3 &= b_{31}x_1 + b_{32}x_2 \cdots + e_3 \\
x_4 &= b_{41}x_1 + b_{43}x_3 + \cdots + e_4 \\
&\vdots \\
x_n &= b_{n2}x_2 + b_{n3}x_3 + \cdots + e_n
\end{aligned}
$$

互いに独立な非観測外乱

変数間決定関係を線形モデルで同定

ベイジアンネットワーク

$$
\begin{aligned}
& p(x_1) \\
& p(x_2 \mid x_1) \\
& p(x_3 \mid x_1, x_2) \\
& p(x_4 \mid x_1, x_3) \\
& \vdots \\
& p(x_n \mid x_1, x_3, \cdots, x_{n-1})
\end{aligned}
$$

変数間決定関係を確率モデルで同定

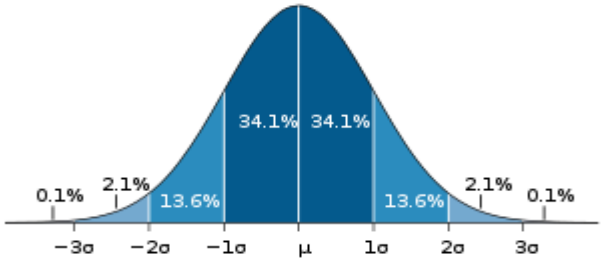
従来の統計的因果推論の問題点

- **従来の統計的因果推論は相関や偏相関など２次の統計量を用いて変数間の因果的順序を導出。**

  非観測外乱項や確率密度がガウス分布と仮定。

$$C(x_1, x_2) = \frac{\sum_{i=1}^{n} x_1 x_2}{\sqrt{\sum_{i=1}^{n} x_1^{\,2}} \sqrt{\sum_{i=1}^{n} x_2^{\,2}}} = 0$$

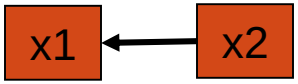
⇒ x_1とx_2は無関係



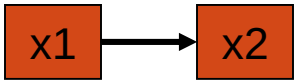
- **しかし、このガウス性の仮定（あるいは近似）により、一般に識別不可能な因果構造が存在する.**

  (Pearl, 2000; Spirtes et al. 2000):

Model 1 $x_1 := b_{12} x_2 + e_1$



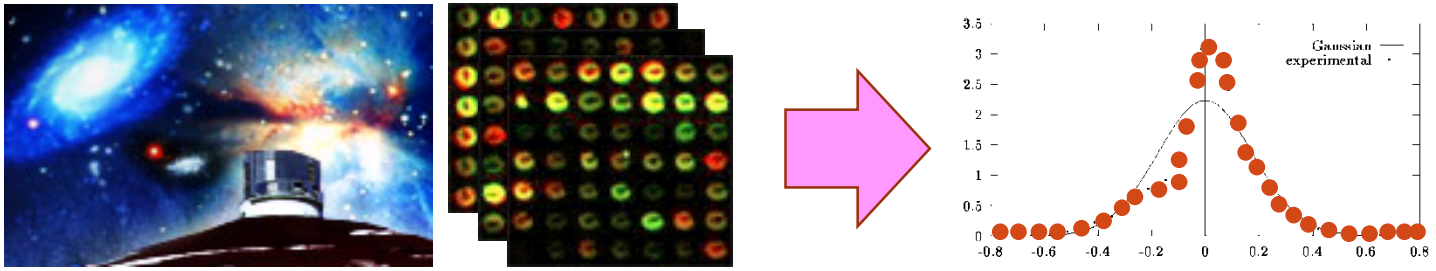
Model 2 $x_2 := b_{21} x_1 + e_2$



非観測外乱項や確率密度の非ガウス性に着目

我々研究室の清水昌平准教授中心に研究， UAI05, JMLR06, UAI09

- “Many observed data are considerably non Gaussian.” (Micceri, 1989; Hyvarinen et al. 2001)



- **非ガウス成分を考えることで、変数間の因果構造は識別可能となる。**

$$
\begin{aligned}
x_1 &= & & & e_1\\
x_2 &= b_{21}x_1 + & & \cdots + & e_2\\
x_3 &= b_{31}x_1 + b_{32}x_2 & & \cdots + & e_3\\
x_4 &= b_{41}x_1 & + b_{43}x_3 + \cdots + & & e_4\\
&\vdots\\
x_n &= & b_{n2}x_2 + b_{n3}x_3 + \cdots + & & e_n
\end{aligned}
$$

互いに独立な非ガウスの非観測外乱

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

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

LiNGAMモデル (Linear Non-Gaussian Acyclic Model)

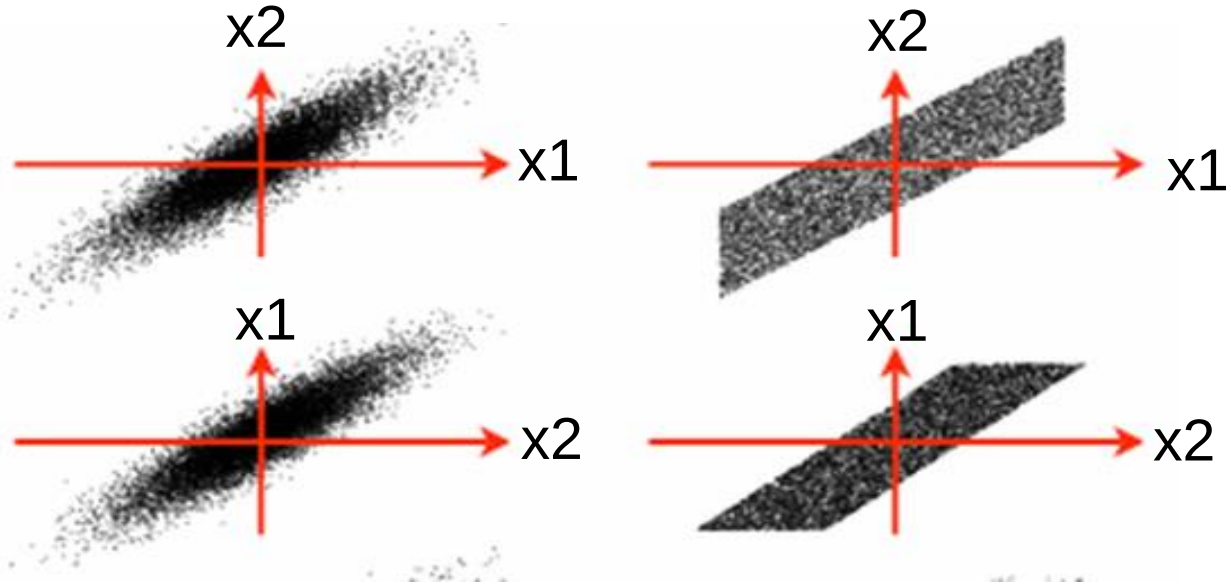
ガウス性 v.s. 非ガウス性

Model 1

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

Model 2

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

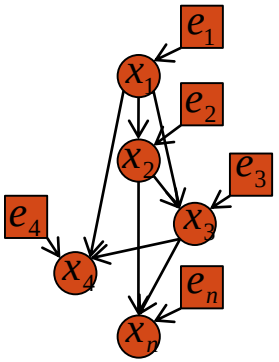


Gaussian　　Non-Gaussian

提案手法：線形非正規非巡回モデル

Linear Non-Gaussian Acyclic Model (LiNGAM model)
(Shimizu et al. UAI05)

- **構造方程式モデル**（Structural Equation Model:SEM）



DAG (Directed Acyclic Graph) Structure

$$
\begin{aligned}
x_1 &= e_1 \\
x_2 &= b_{21}x_1 + \cdots + e_2 \\
x_3 &= b_{31}x_1 + b_{32}x_2 + \cdots + e_3 \\
x_4 &= b_{41}x_1 + b_{43}x_3 + \cdots + e_4 \\
&\vdots \\
x_n &= b_{n2}x_2 + b_{n3}x_3 + \cdots + e_n
\end{aligned}
$$

- **各観測変数 x_i は連続確率変数.**
- **変数間の因果関係は線形で有向非巡回グラフ(DAG) で表される.**
- **各外乱 e_i は互いに独立で非ガウス揺らきを持つ.**

$$x_i = \sum_{k(j)<k(i)} b_{ij}x_j + e_i \quad \text{or} \quad \mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e}$$

- **行列Bは行と列の同時入れ替えによって下三角化される.**

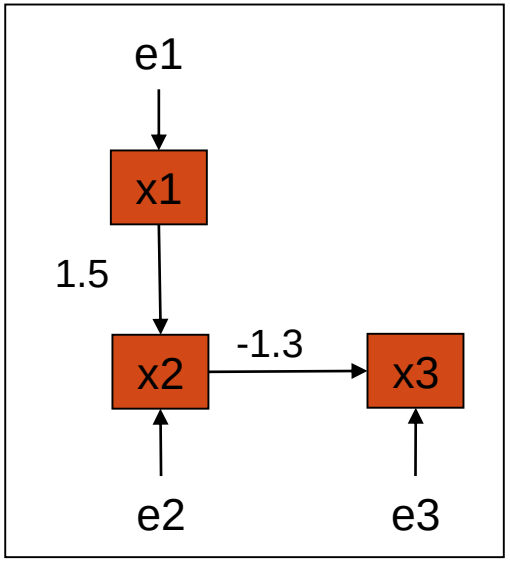
具体例

- ３変数モデル

$$x_1 = e_1$$
$$x_2 = 1.5x_1 + e_2$$
$$x_3 = -1.3x_2 + e_3$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 0 & 0 \\ 1.5 & 0 & 0 \\ 0 & -1.3 & 0 \end{bmatrix}}_{\mathbf{B}} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}$$

- 変数の因果的順序
  - $k(1)=1, k(2)=2, k(3)=3$
  - x2 はx1から影響を受けるが，x3からは影響を受けない．
- 外乱
  - 外乱x1は外乱e1に**等しい**．x1を**外生変数**という．外生変数は**最も上流の変数**である．
  - e2とe3は**付加外乱あるいは誤差** である．



提案手法(Shimizu et al., UAI09)
変数間因果の判定(２変数の例)

i) $x_1 (= e_1)$ が上流である

$$x_2 := b_{21} x_1 + e_2$$

x_1 and e_2 are independent.

x_2 and e_2 are NOT independent.

x_1とx_2の回帰式

$$r_2 = x_2 - \frac{\mathrm{cov}(x_2, x_1)}{\mathrm{var}(x_1)} x_1 = e_2$$

x_1 and r_2 are independent.

ii) x_1 が上流でない.

$$x_1 := b_{12} x_2 + e_1$$

x_2 and e_1 are independent.

x_1 and e_1 are NOT independent.

x_1とx_2の回帰式

$$r_2 = x_2 - \frac{\mathrm{cov}(x_2, x_1)}{\mathrm{var}(x_1)} x_1$$

$$= \left\{ 1 - \frac{b_{12}\,\mathrm{cov}(x_2, x_1)}{\mathrm{var}(x_1)} \right\} x_2 - \frac{\mathrm{cov}(x_2, x_1)}{\mathrm{var}(x_1)} e_1$$

x_1 and r_2 are NOT independent.

提案手法
独立性を測る尺度

- **厳密にはあらゆる非線形関数f,gの非線形相関ゼロを確認せねばならないが、現実には無理。**

独立性判定：非線形相関係数

$$C_{fg}(x_1, r_2) = \frac{\sum_{i=1}^{n} f(x_1) g(r_2)}{\sqrt{\sum_{i=1}^{n} f(x_1)^2} \sqrt{\sum_{i=1}^{n} g(r_2)^2}} = 0$$

（あらゆる非線形関数f,gについて）

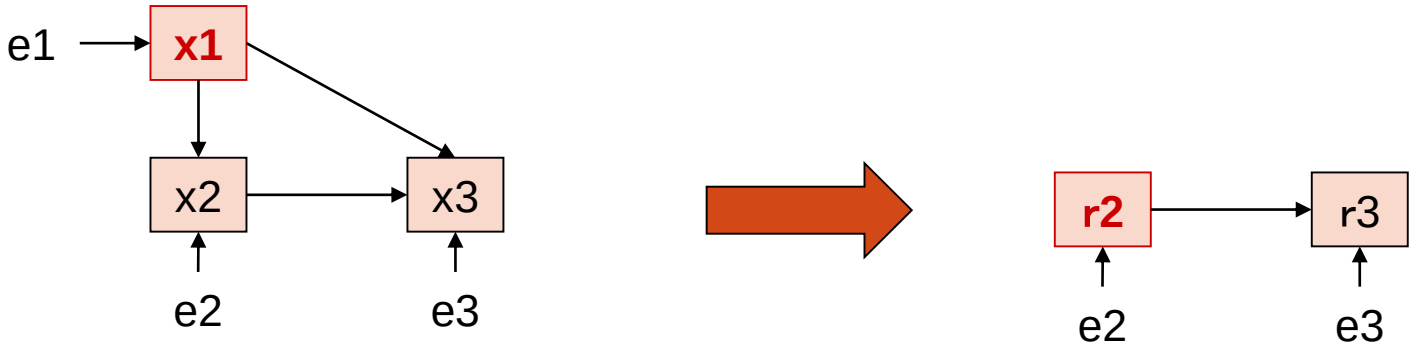
- **以下の非線形相関により近似的に変数とその残差の独立性を検定評価**

$$C_{fg}(x_1, r_2) \cong \left|\mathrm{corr}\{x_1, g(r_2)\}\right| + \left|\mathrm{corr}\{g(x_1), r_2\}\right| \quad (g = \tanh)$$

モデルの前提がデータで成立していれば，非線形相関の近似誤差を除いて，この方法は必ず正しい唯一の因果構造を同定することが証明されている．

DirectLiNGAM: 直接的因果推論法

- **基本原理:**
  1. **前掲の回帰式法より最上流の外生変数を探す.**
     - **例えば** $x_1 \rightarrow \{x_2, x_3\}$ **が判る. 回帰により**
     - **最上流の**x1**の成分を他の変数データから除く.**
  2. **“外生”の回帰残差を探す.**
     - **例えば** $r_2 \rightarrow r_3$, *i.e.*, $x_2 \rightarrow x_3$.
       残差の因果順序と元の変数の因果順序は一致する.
  3. **最終的に** $x_1 \rightarrow x_2 \rightarrow x_3$ **が導かれる.**

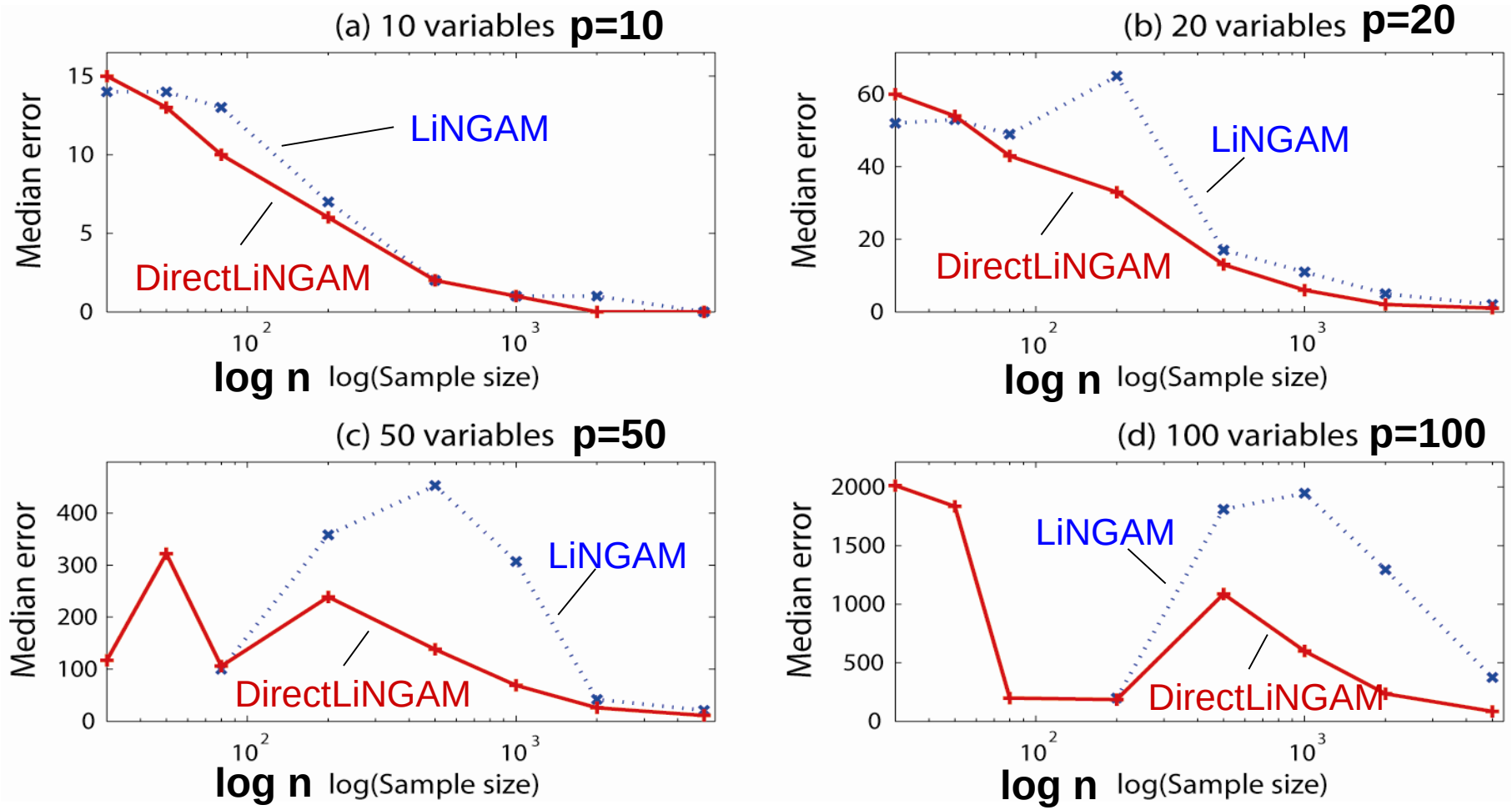


DirectLiNGAM:まとめ

- 知っていること（あるいは仮定）
  - データ Xは$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e}$によって生成されている．
- 知らないこと
  - 変数間の結合係数: $b_{ij,}$
  - 因果的順序: k(i),
  - 各外乱: e_i
- データ Xのみ観測されている．

- データXのみからBとk,eが導びかれる！

標本サイズv.s. 因果構造同定誤差



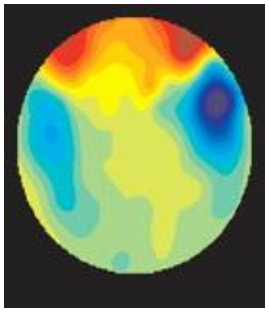
n = 30, 50, 80, 200, 500, 1000, 2000

変数が100個程度のデータでも正確に因果構造を同定可能

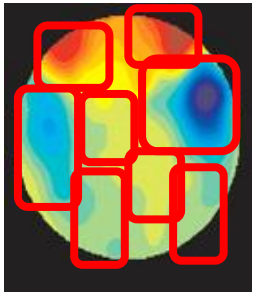
脳磁図Magneto-EncephaloGraphyデータによる脳領野関係の解析

- **脳活動で生じる磁場を超伝導量子干渉計** (SQUIDs) **によって計測イメージングする。**
- SQUIDs**は非常に微弱な磁場も検知できる感度を有する。**

http://www.nlm.nih.gov/



http://jn.physiology.org/

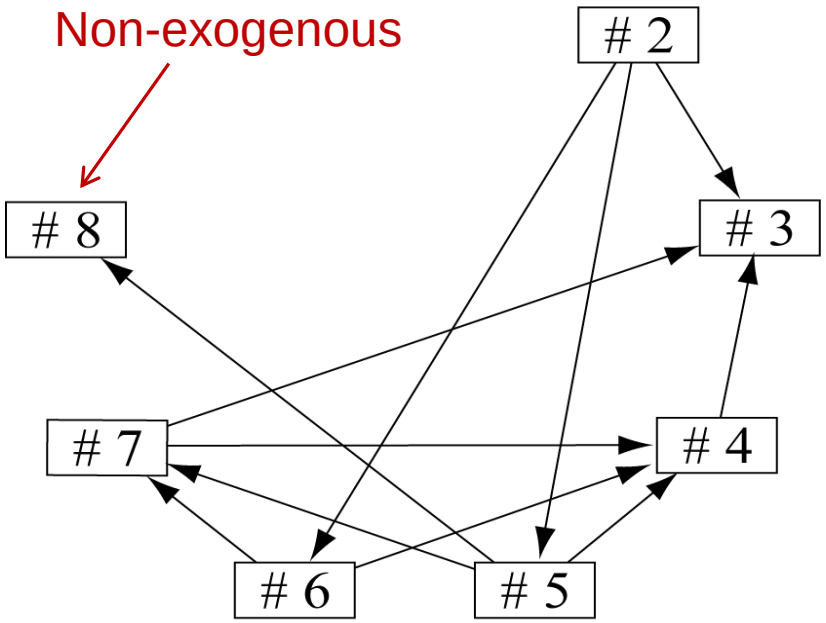


各領野毎の
平均磁場レベル
を観測変数化

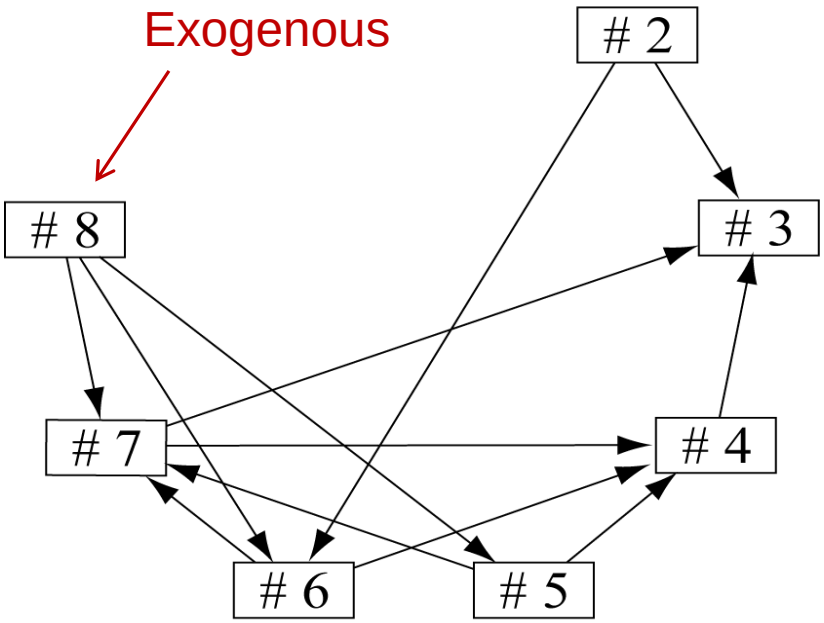
x_1 x_2 x_4 x_3 x_5 x_8 x_6 x_7

脳波Magneto-encephalography data から導いた脳磁気源の影響関係

- **各手法により推定された因果構造**:

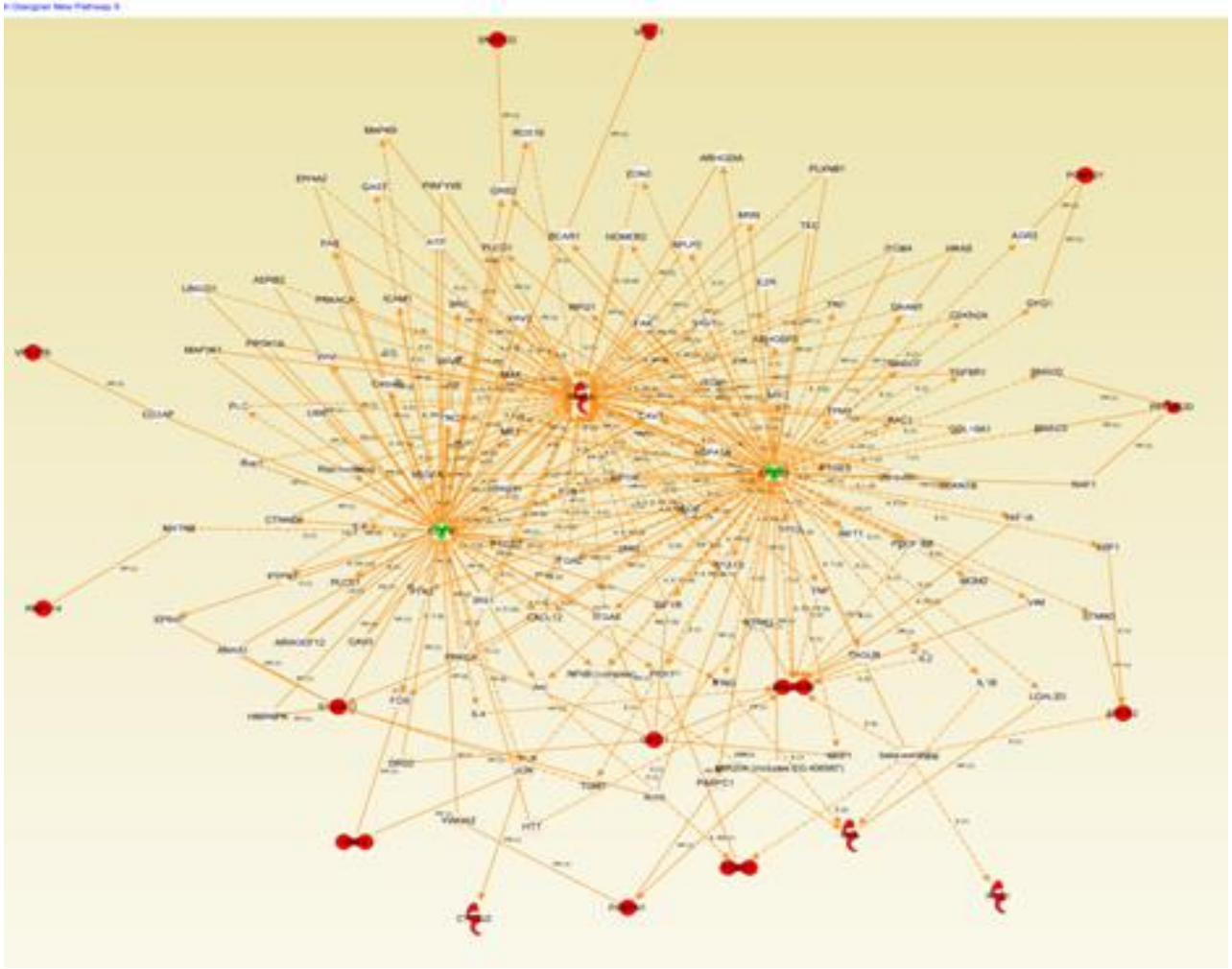


(Original) LiNGAM



DirectLiNGAM (proposed)

非ガウス性に基づく外生変数同定結果と遺伝子データベースから得た乳がん遺伝子発現因果ネットワーク



外生変数同定手法
eggFinder
(Shimizu and Sogawa et al. JSAI10)

＋

Ingenuity Pathways Database

対象の仕組みを理解するモデルのマイニング
まとめ

- 各変数の独立外乱の非ガウス分布を利用して、変数間の影響決定関係を表すモデルを推定する**統計的因果推論法である**Direct LiNGAM**法を提案**

- 脳磁図データに適用し、**脳領野間の活動の影響関係**を明らかにした。

- その他、乳がん遺伝子の発現影響関係を表す**遺伝子ネットワーク解析**への応用を示した。

最後に

- ビッグデータのマイニングでは、膨大なデータ件数の処理と平行して、多数の属性・変数を有するデータの解析が大きな課題となっている。

- 目的変数の予測に寄与する変数を正則化法により効果的に選択した予測モデルのマイニング

- 多変数間の統計的因果推論による対象の仕組みを理解するモデルのマイニング