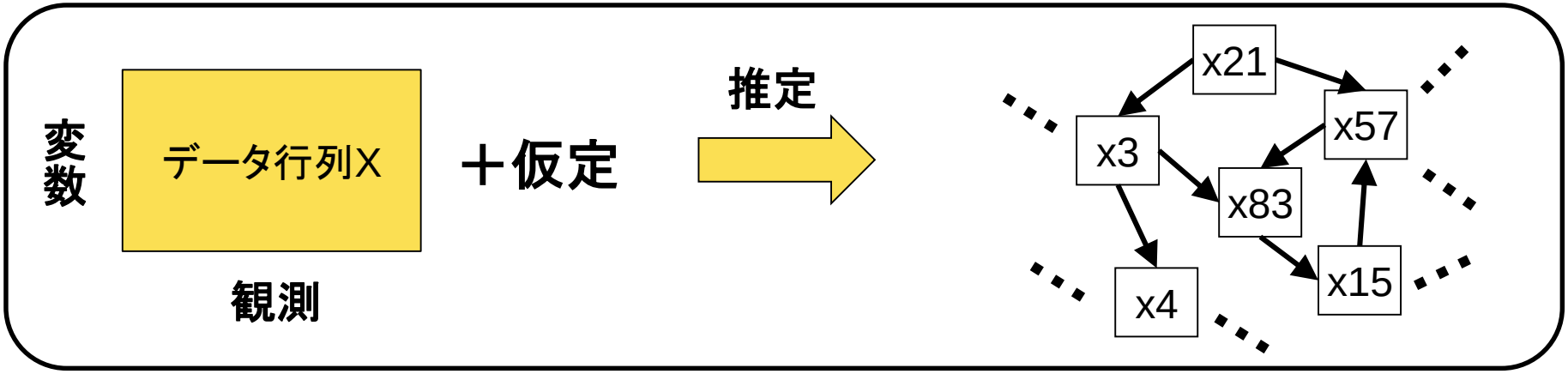
因果探索: 観察データから因果仮説を探索する

清水昌平
大阪大学 産業科学研究所

関連文献はココ: https://sites.google.com/site/sshimizu06/home/lingampapers

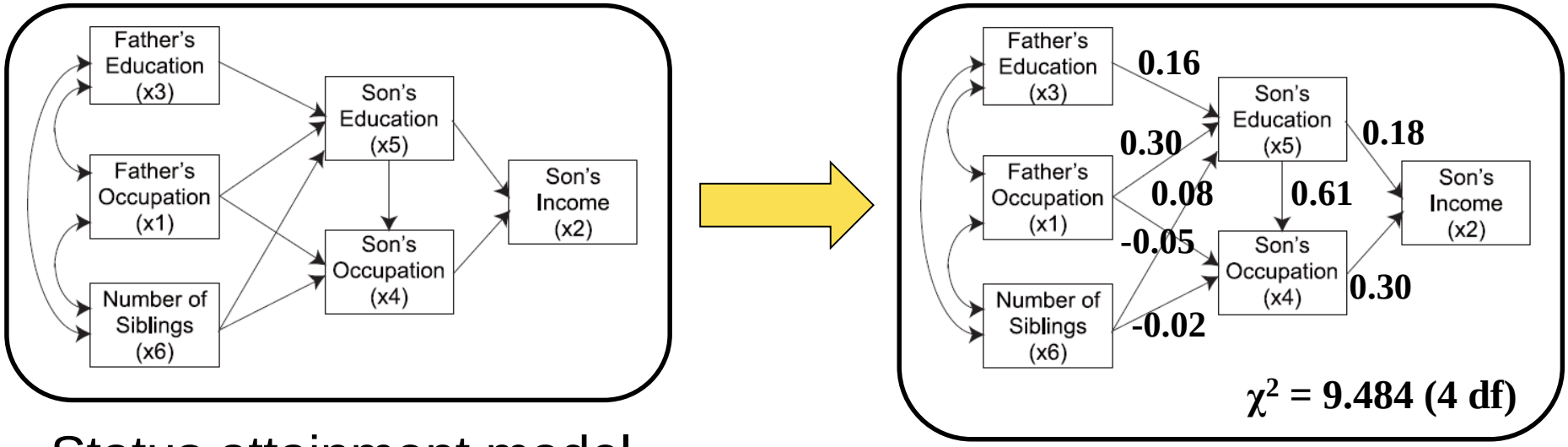
因果探索

- 観察データから因果に関する仮説を探索
- データ＋仮定 → 因果グラフ(パス図)
  - どんな仮定の下で何が導けるか?
  - 仮定の評価方法は？



構造方程式モデリング(SEM)と因果探索

- 社会学データ: General Social Survey (n=1380)

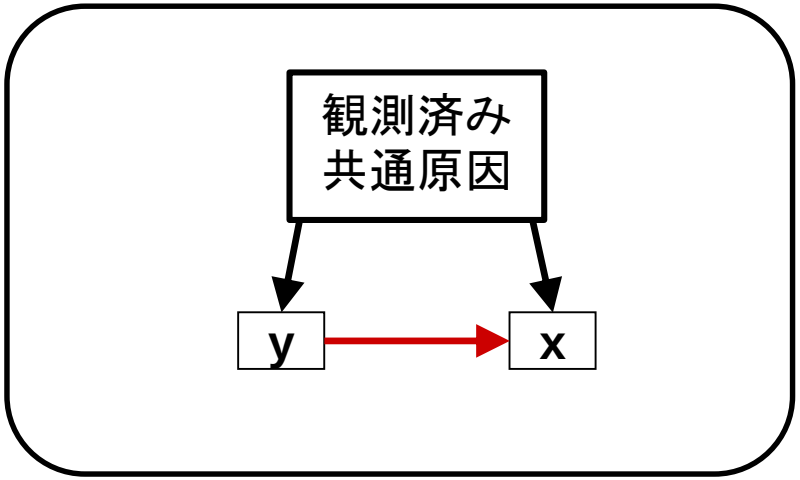


Status attainment model
(Duncan et al., 1972)

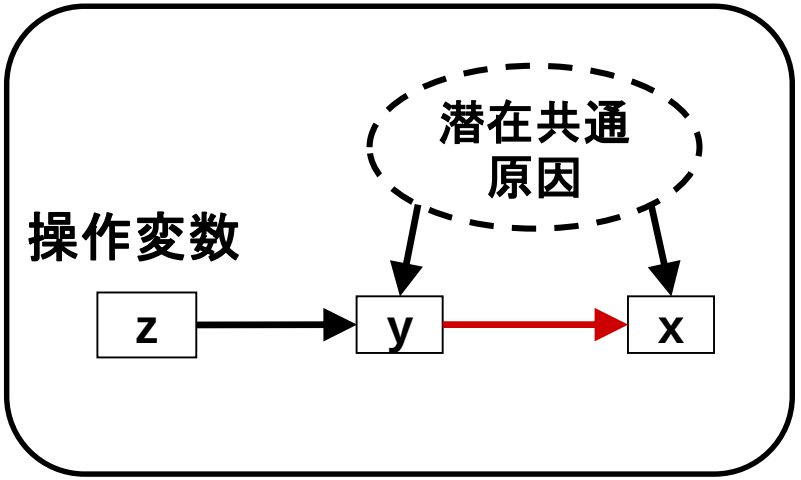
- パス図(因果グラフ)が正しいとして、因果の大きさを推定
- 因果探索法のねらい: そもそもパス図を推定

SEM: 分析者は事前に判断

- 因果方向
- 潜在共通原因がない



xはyの原因でない
「潜在」共通原因はない



zはyの原因
zからxへは直接効果なし
zに関する「潜在」共通原因なし

そんなときに、因果探索法！

- データが分析者を助けてくれる場合もあるはず
- 仮定の按排・トレードオフ
  - パス図 vs. 関数形
- そのための方法論は足りているのか？

足りてない！つくろう！

因果探索の方法論

因果探索法を
簡単にサーベイ

因果探索の基本問題

obs.1 obs.2 … obs.n

x
y

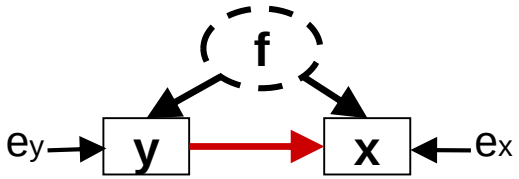
データ行列

$\sim_{i.i.d.} p(x, y)$

仮定: どれかが
データを生成

問題: どれが生成
したかを推定

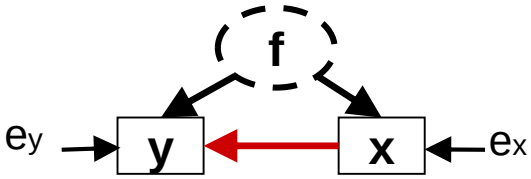
潜在共通原因 f



構造方程式モデル

$$\begin{cases} x = g_x(y, f, e_x) \\ y = g_y(f, e_y) \\ e_x, e_y, f \text{ の分布} \end{cases}$$

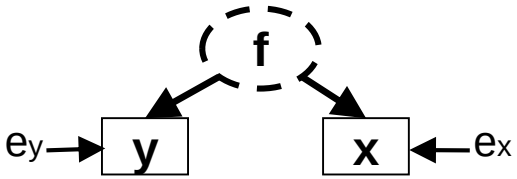
潜在共通原因f



構造方程式モデル

$$\begin{cases} x = g_x(f, e_x) \\ y = g_y(x, f, e_y) \\ e_x, e_y, f \text{ の分布} \end{cases}$$

潜在共通原因f

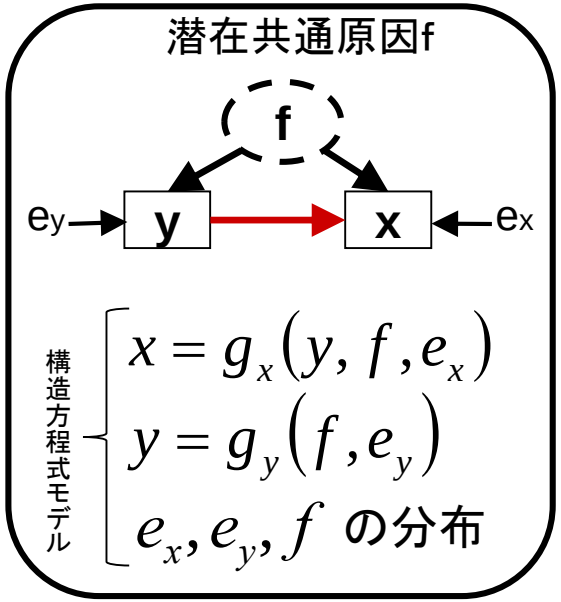


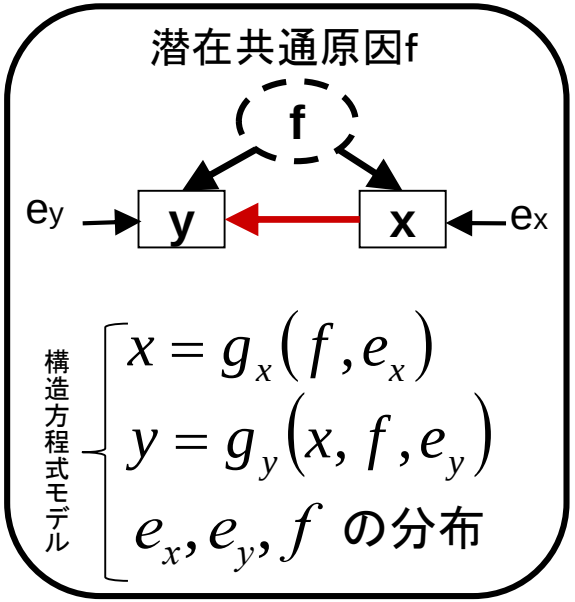
構造方程式モデル

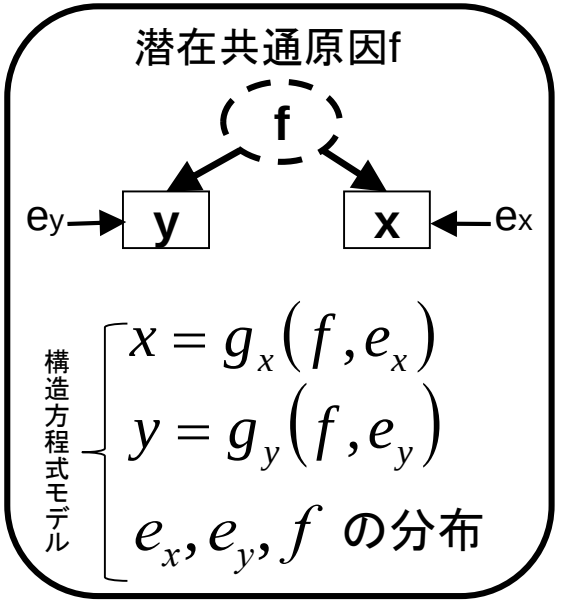
$$\begin{cases} x = g_x(f, e_x) \\ y = g_y(f, e_y) \\ e_x, e_y, f \text{ の分布} \end{cases}$$

因果探索: 3つのアプローチ

1. ノンパラ (Spirtes+93; Pearl00)
   - 関数形にも分布にも仮定おかず → どれかわからない
2. パラメトリック
   - 線形＋正規分布 → どれかわからない
3. セミパラ
   - 線形+非正規分布 → どれかわかる

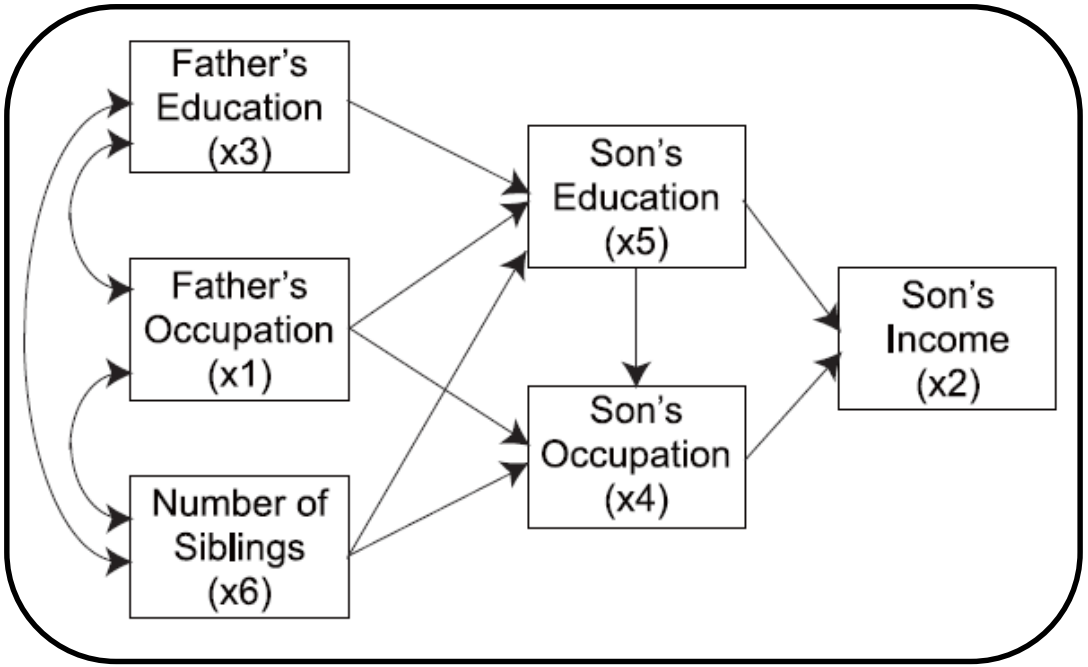


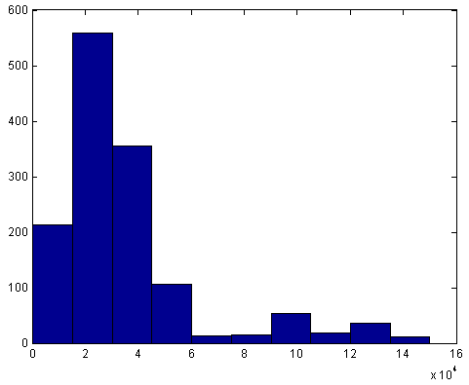




非正規分布

- General Social Survey (米国)
  - 非農業, 35-44歳, 白人、男性、就業、1972-2006
  - サンプルサイズ: 1380





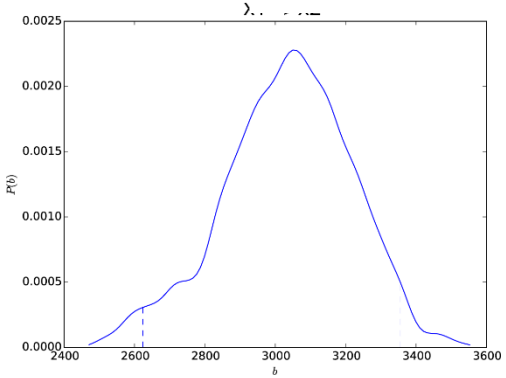
x2: Son's Income

背景知識 (Duncan et al., 1972)
Status attainment model

適用イメージ (Shimizu & Bollen, 2014)

- 社会学データ: General Social Survey (n=1380)

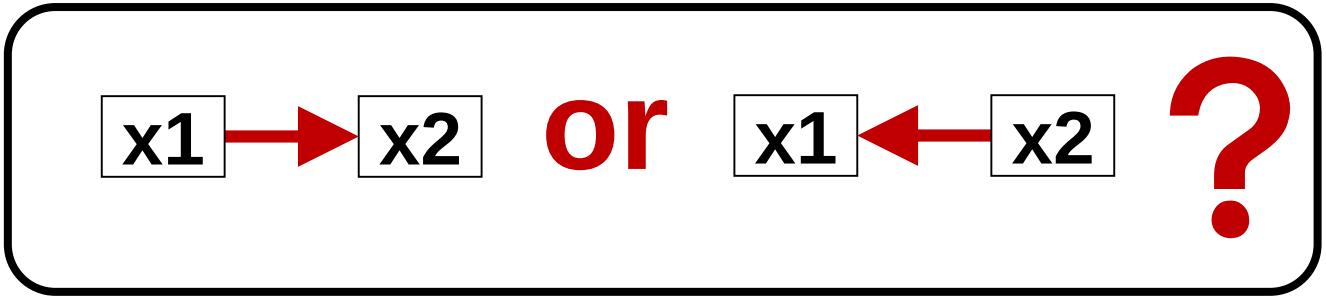
- 係数の事後分布
  - 収入←学歴



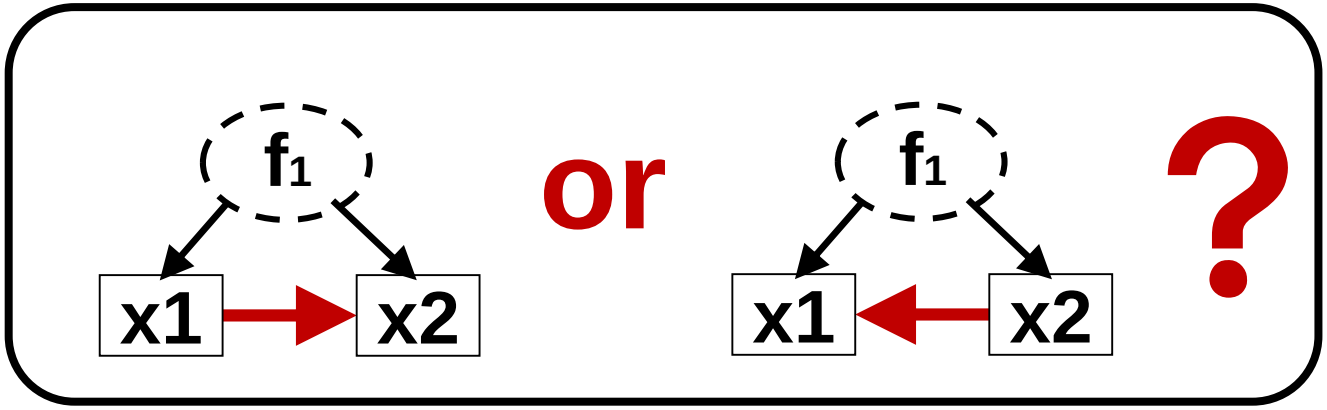
- やっとスタート地点についた！?

Major challenges

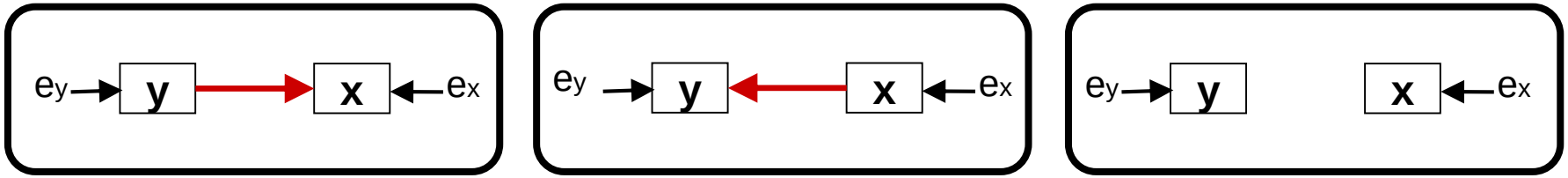
1. 時間情報がないときに因果方向を推定



2. 潜在共通原因への対処



潜在共通原因が「ない」場合

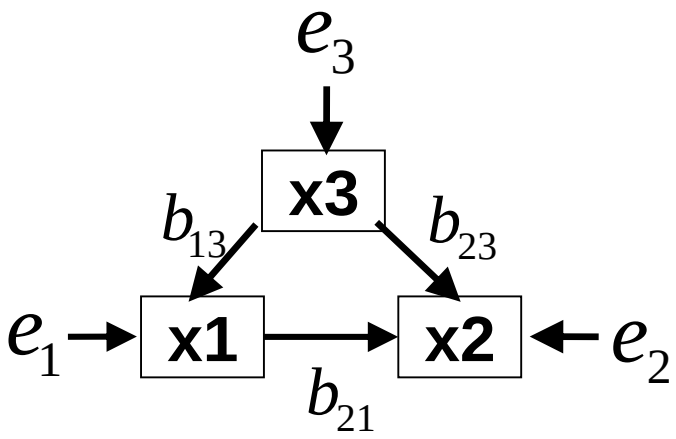


SEMにおける非正規性利用 LiNGAMモデル

(Shimizu et al., 2006, J. Machine Learning Research)

- データXから因果方向, 係数, 切片が識別可能(一意に推定可能)

$$x_i = \mu_i + \sum_{j \neq i} b_{ij} x_j + e_i$$



基礎仮定

- 線形性
- 非巡回
- 非正規誤差 e_i
- e_i は互いに独立
  (潜在共通原因なし)

識別可能: 方向が違えば分布が違う

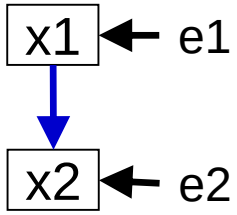
e_1, e_2がガウス

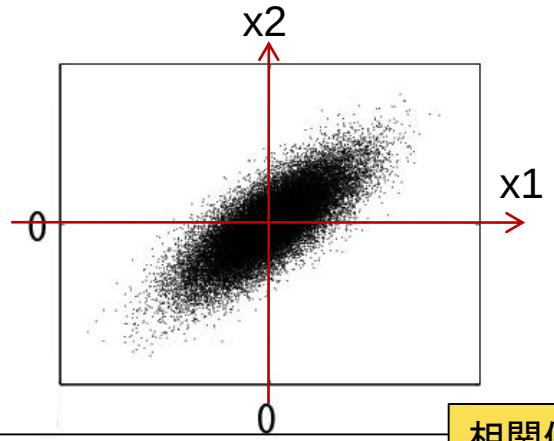
e_1, e_2が非ガウス
(一様分布)

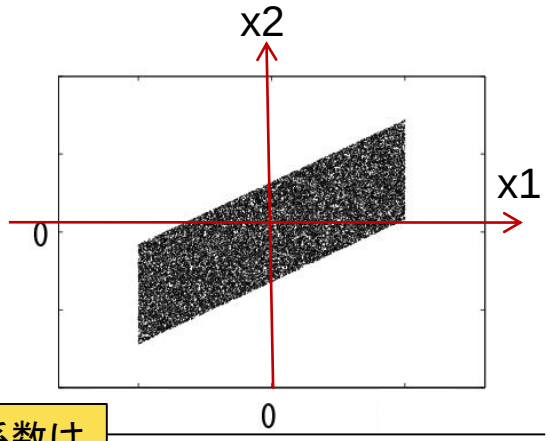
モデル1:

$x_1 = e_1$

$x_2 = 0.8x_1 + e_2$





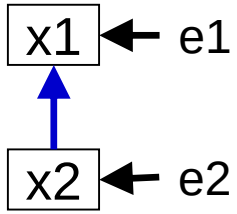


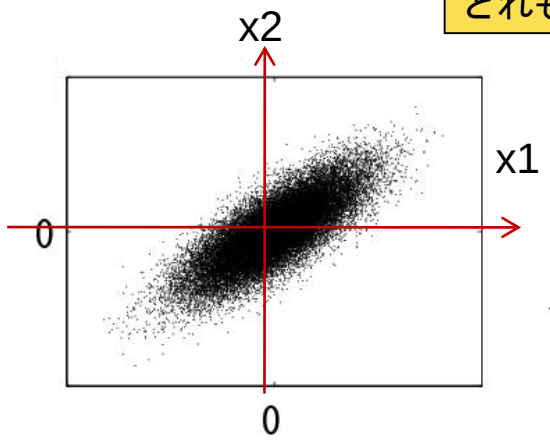
相関係数は
どれも0.8

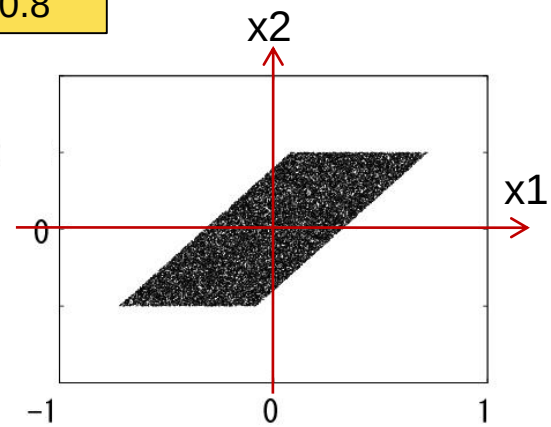
モデル2:

$x_1 = 0.8x_2 + e_1$

$x_2 = e_2$







$E(e_1) = E(e_2) = 0,$
$\mathrm{var}(x_1) = \mathrm{var}(x_2) = 1$

線形非正規以外にも

非線形＋加法の外生変数

- 「非線形＋加法の外生変数」のモデル

1. $x_i = \sum_{x_i \text{の親}} f_k(x_k) + e_i$ -- Imoto et al. (2002)
2. $x_i = f_i(x_i \text{の親}) + e_i$ -- Hoyer et al. (2008)
3. $x_i = f_{i,2}^{-1}(f_{i,1}(x_i \text{の親}) + e_i)$ -- Zhang et al. (2009)

- いくつかの非線形性と外生変数の分布を除いて識別可能
  (Zhang & Hyvarinen, 2009; Peters et al., 2014)

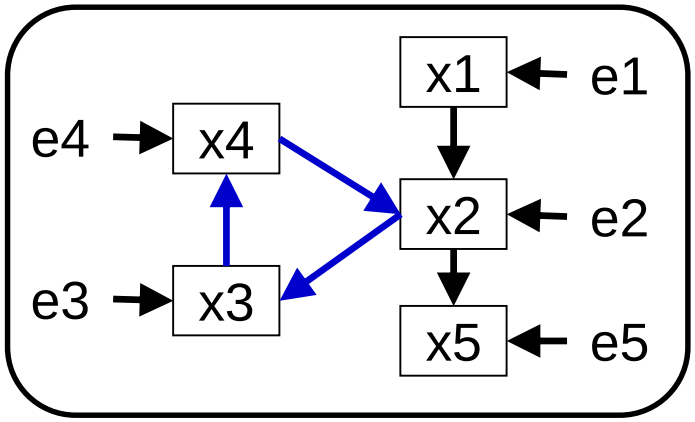
- Open problem: どこまで緩められるか?

巡回モデル

(Lacerda et al., 2008; Hyvarinen & Smith, 2013)

- モデル:

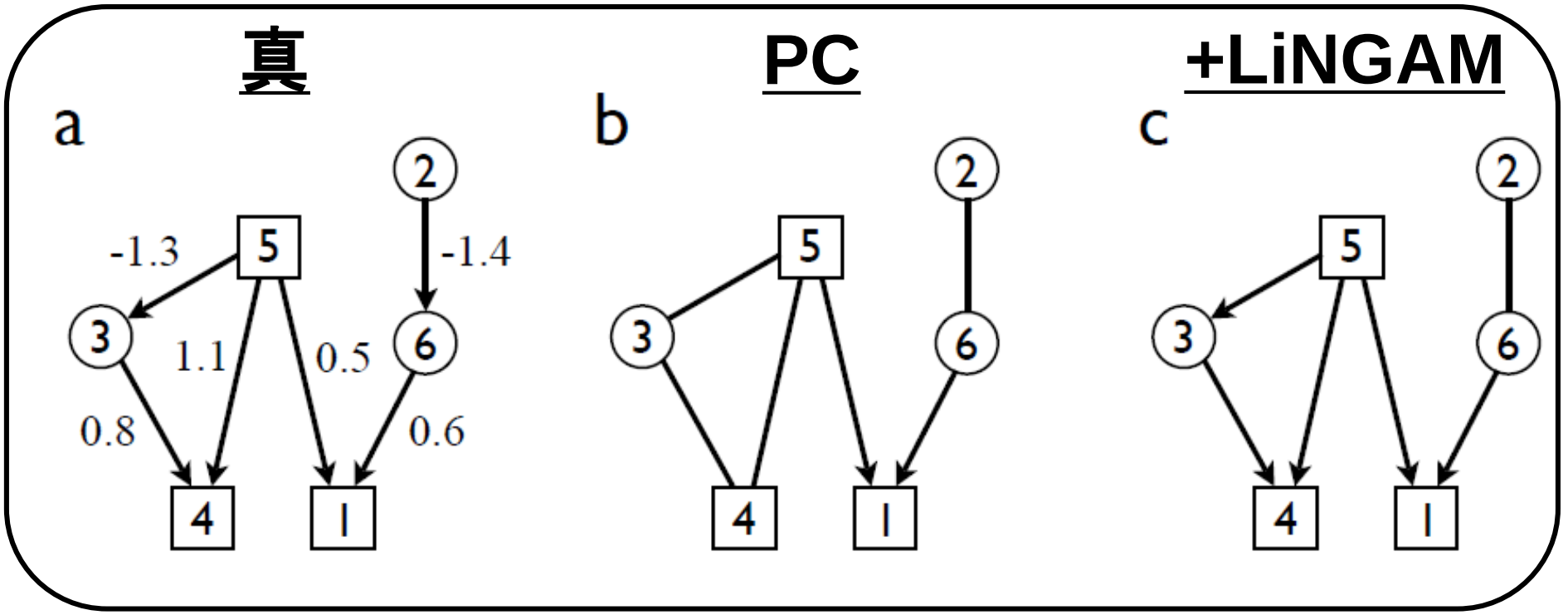
$$x_i = \mu_i + \sum_{j \neq i} b_{ij} x_j + e_i$$



- 識別性の十分条件
  - **B**の固有値の絶対値が1未満(平衡状態にある)
  - ループが交わらない
  - 自己ループなし

正規と非正規が混在

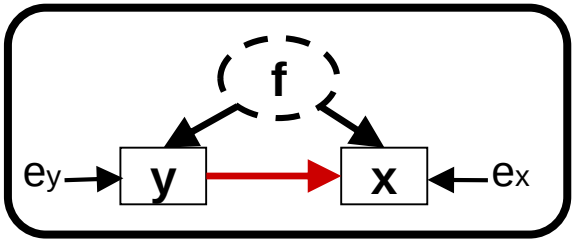
- PCアルゴリズム(or GES)+LiNGAM
  - Hoyer+08UAI; Ramsey+11NeuroImage



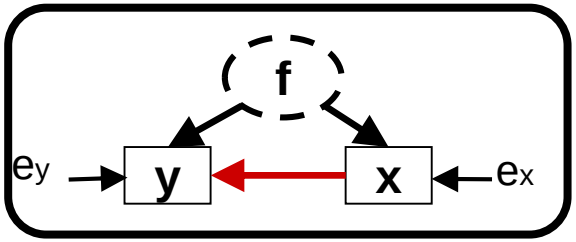
○は誤差項が正規　　□が誤差項が非正規

潜在共通原因が「ある」場合

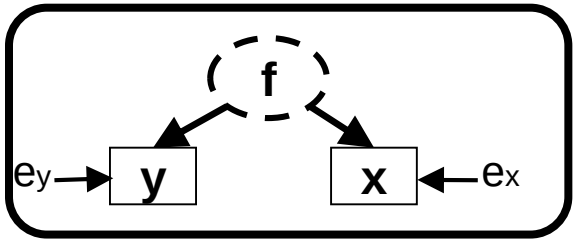
潜在共通原因 f



潜在共通原因f



潜在共通原因f

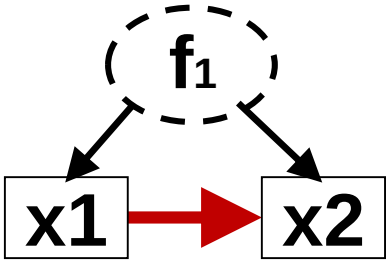


この中で、どれが一番いい？

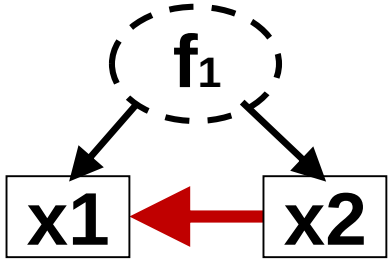
潜在共通原因がある場合

(Hoyer et al., 2008, Int. J. Approximate Reasoning
Shimizu & Bollen, 2014, J. Machine Learning Research)

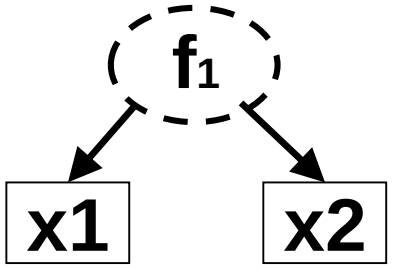
- 条件: 線形性+非巡回+非正規連続分布
  - 潜在共通原因の個数は特定不要
- 推定法に関する研究は発展途上



or



or



?

$$x_1 := \lambda_{12} f_1 + e_1$$
$$x_2 := b_{21} x_1 + \lambda_{21} f_1 + e_2$$

$$x_1 := b_{12} x_2 + \lambda_{12} f_1 + e_1$$
$$x_2 := \lambda_{21} f_1 + e_2$$

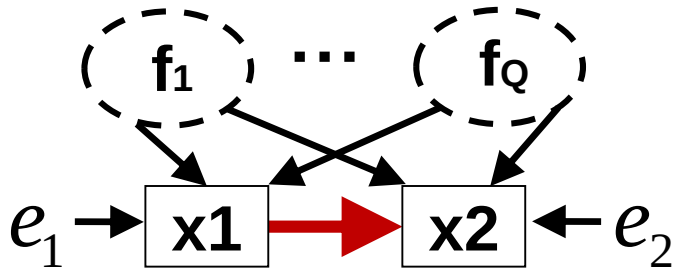
$$x_1 := \lambda_{12} f_1 + e_1$$
$$x_2 := \lambda_{21} f_2 + e_2$$

識別可能: 方向が違えば分布が違う

- 推定: モデル選択(尤度、ベイズ etc.)

$$x_1 = \sum_{q=1}^{Q} \lambda_{1q} f_q + e_1$$

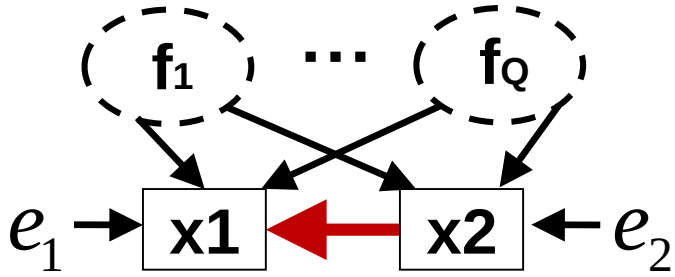
$$x_2 = \sum_{q=1}^{Q} \lambda_{2q} f_q + b_{21} x_1 + e_2$$

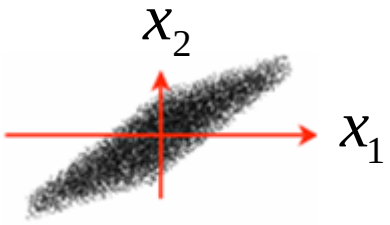


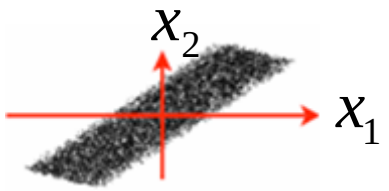
or

$$x_1 = \sum_{q=1}^{Q} \lambda_{1q} f_q + b_{12} x_2 + e_1$$

$$x_2 = \sum_{q=1}^{Q} \lambda_{2q} f_q + e_2$$



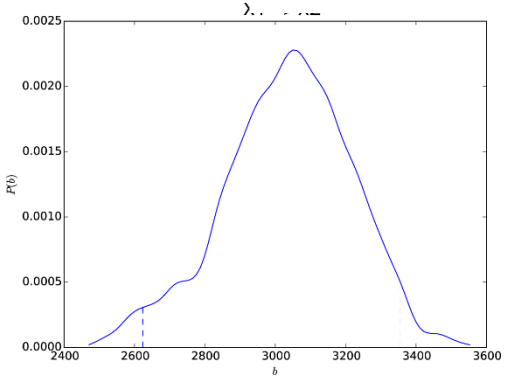




適用イメージ (Shimizu & Bollen, 2014)

- 社会学データ: General Social Survey (n=1380)

- 係数の事後分布
  - 収入←学歴



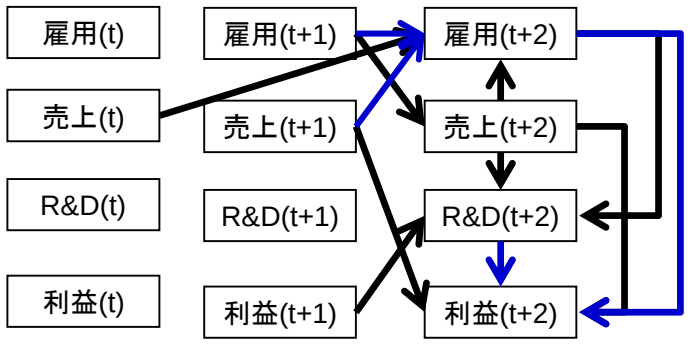
- やっとスタート地点についた！?

再

課題

適用分野からの要請

- 異質性・非定常性
  - 人により時点により因果関係が異なる
  - 潜在調整変数もココか
- 出始めた: Huang+IJCAI15

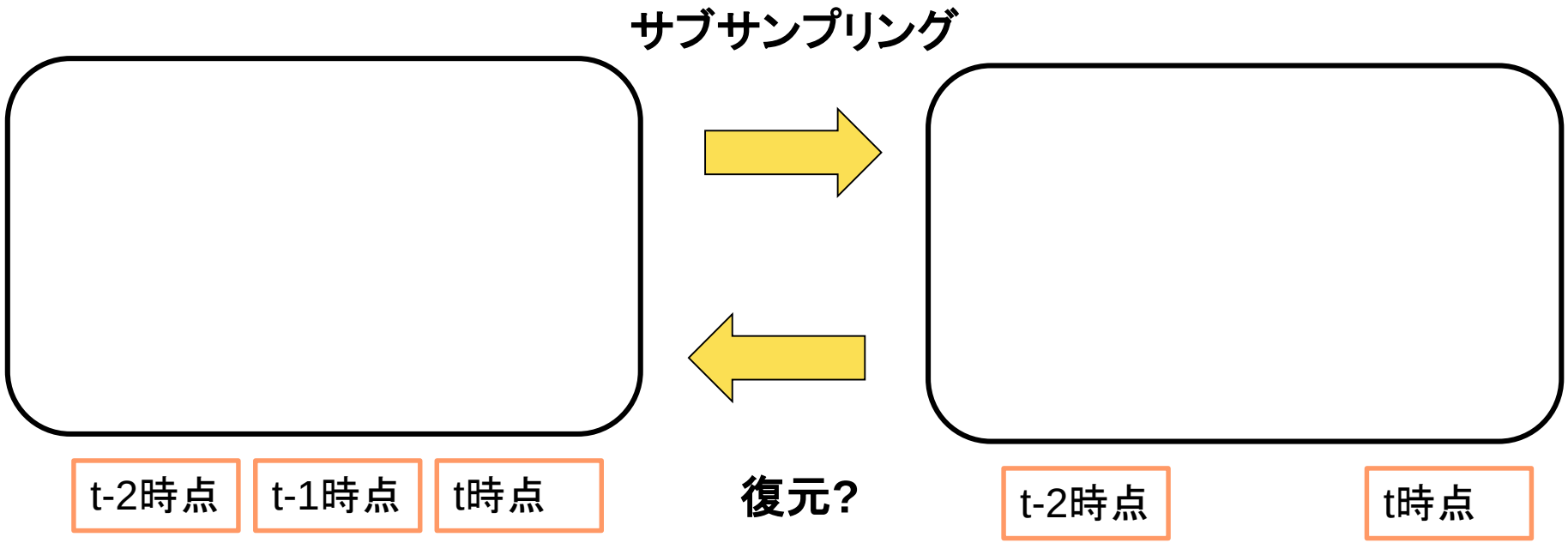


経済学: Moneta+13, Oxford B Econ. Stat.

脳: Mills-Finnerty+14, NeuroImage

潜在中間変数?

- 風が吹けば…
- 時系列データにおけるサブサンプリングの影響
  (Gong+15ICML; Hyttinen+16; cf. Hyvarinen+10)



図はHyttinen+16より

未知のこと多い

- 潜在共通原因を許しつつ
  どこまで拡張できるのか？
  - 非線形性、巡回性、異質性、非定常性
  - 局所解は避けたい
- 離散変数の場合は？混在する場合は？
  - Peters+11TPAM, Parks+15NIPS
- 選択バイアスのある場合は？(+αある?)
- 一般的話題(?)も必要
  - はずれ値？
  - 変数変換？

仮定の評価

- 正規性の検定
  - 観測変数や外生変数(誤差)の非正規性チェッ (Moneta+13)
- 外生変数(誤差)間の独立性検定
  - 従属 → 潜在共通原因あり (Entner+ 2011; 2012)
- 全体的な適合度
  - カイ二乗検定 (Shimizu & Kano, 2008)

- 参考: 統計的信頼性評価
  - ブートストラップ (Komatsu, Shimizu & Shimodaira, 2010)

おわりに

- 因果推論の数理的基盤はかなり整った
  - Rubin, Pearl, …
- 因果探索の方法論をつくろう/使ってみよう
  - どんな状況で何が何の原因で結果か
- 未知の事柄は多い
  - さぁ、これからです！