

非ガウス構造方程式モデルにおける 因果順序の推定: 潜在交絡変数に頑健な方法

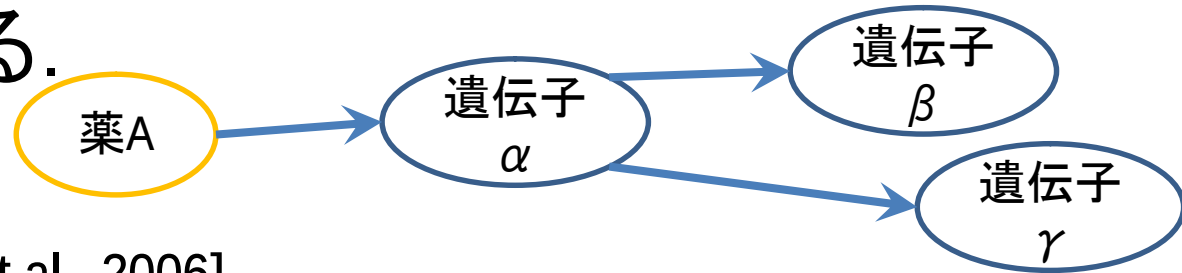
田代 竜也^{*1}, 清水 昌平^{*1},
Aapo Hyvärinen^{*2}, 鷲尾 隆^{*1}

^{*1}大阪大学 産業科学研究所

^{*2}フィンランド ヘルシンキ大学

概要

- 各事象(変数)間の因果構造を観測データのみを用いて推定する.



- LINGAM [Shimizu et al., 2006]
 - データの非ガウス性を用いることでモデルを一意に同定可能.
 - 既存の推定手法ではデータが仮定を満たしていない場合, 全く異なる結果を出力する場合がある.
- 提案手法
 - 仮定破綻の典型的原因の一つである潜在交絡変数に対してロバストに因果構造を推定.

背景

LiNGAMモデル

[Shimizu et al., 2006]

4

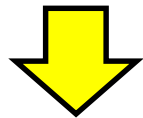
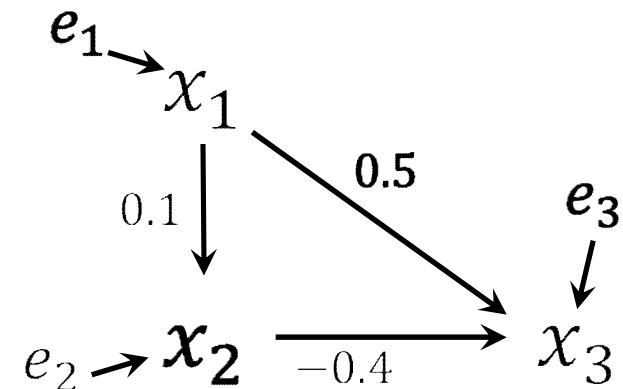
- LiNGAM (Linear Non-Gaussian Acyclic Model)
 - 観測変数 x_i が下図のような有向非巡回グラフ(DAG)から生成される統計モデル.

$$x_1 = e_1$$

$$x_2 = 0.1x_1 + e_2$$

$$x_3 = 0.5x_1 - 0.4x_2 + e_3$$

・互いに独立
・非ガウス



行列表記: $\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e}$

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 \\ 0.1 & 0 & 0 \\ 0.5 & -0.4 & 0 \end{bmatrix}$$

...

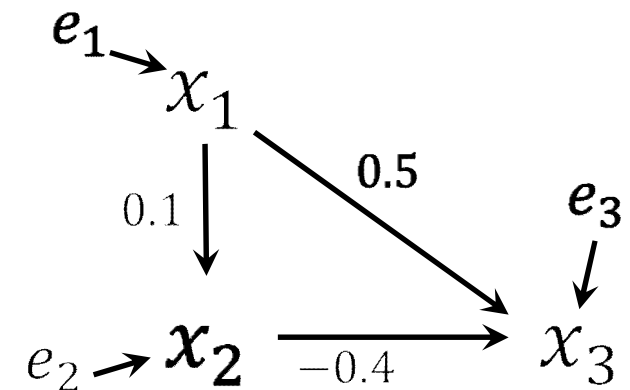
対角成分が0の下三角行列

因果的順序 $k: x_1 \rightarrow x_2 \rightarrow x_3$

- DirectLiNGAM
 - データ行列 $\mathbf{x}$ から因果的順序 k と係数行列 $\mathbf{B}$ を推定.

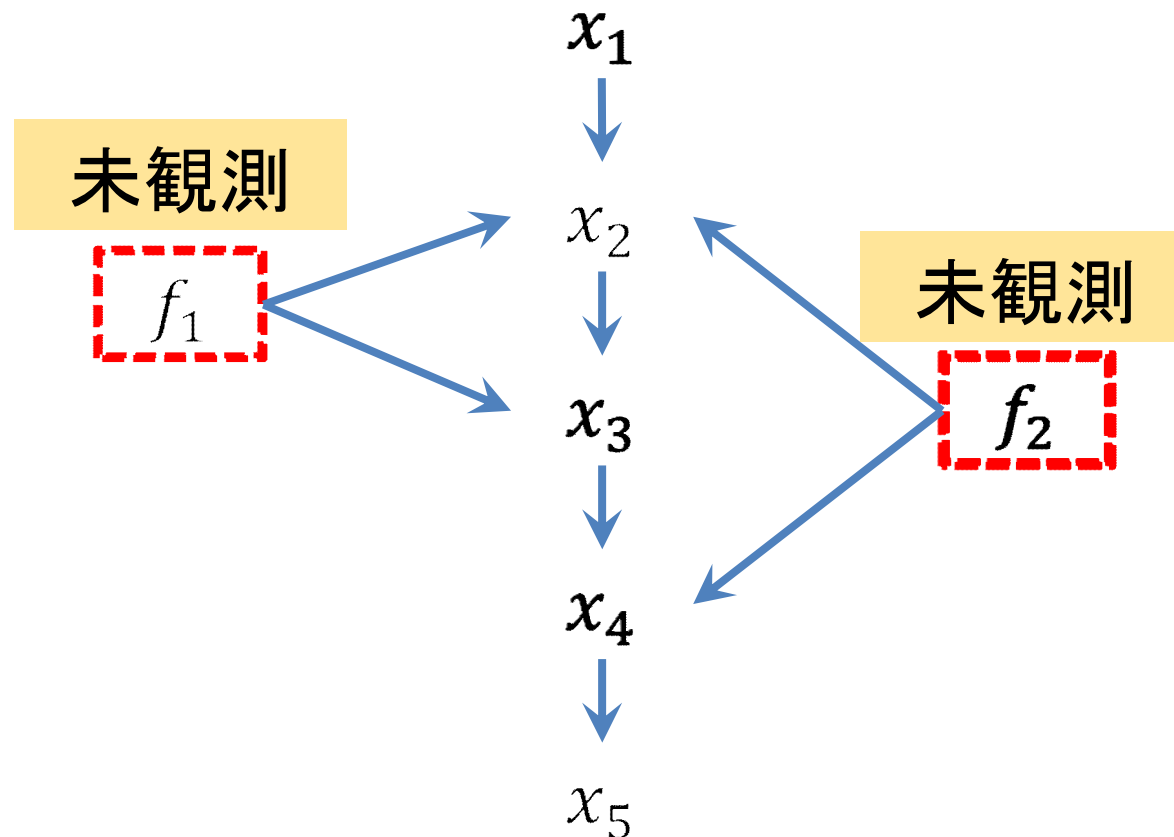
LiNGAMモデルの仮定

1. 線形
2. 非巡回性
3. 外乱 e_i は非ガウス
4. 外乱 e_i は互いに独立
5. 潜在交絡変数が存在しない



潜在交絡変数

- 潜在交絡変数 (Latent confounding variable)
 - 2つ以上の観測変数の親となるような未観測の変数.



LvLiNGAM モデル

[Hoyer et al., 2008]

- LvLiNGAM (Latent variable LiNGAM)
 - LiNGAMに潜在交絡変数 $f_i (j = 1 \dots q)$ を考慮したモデル.

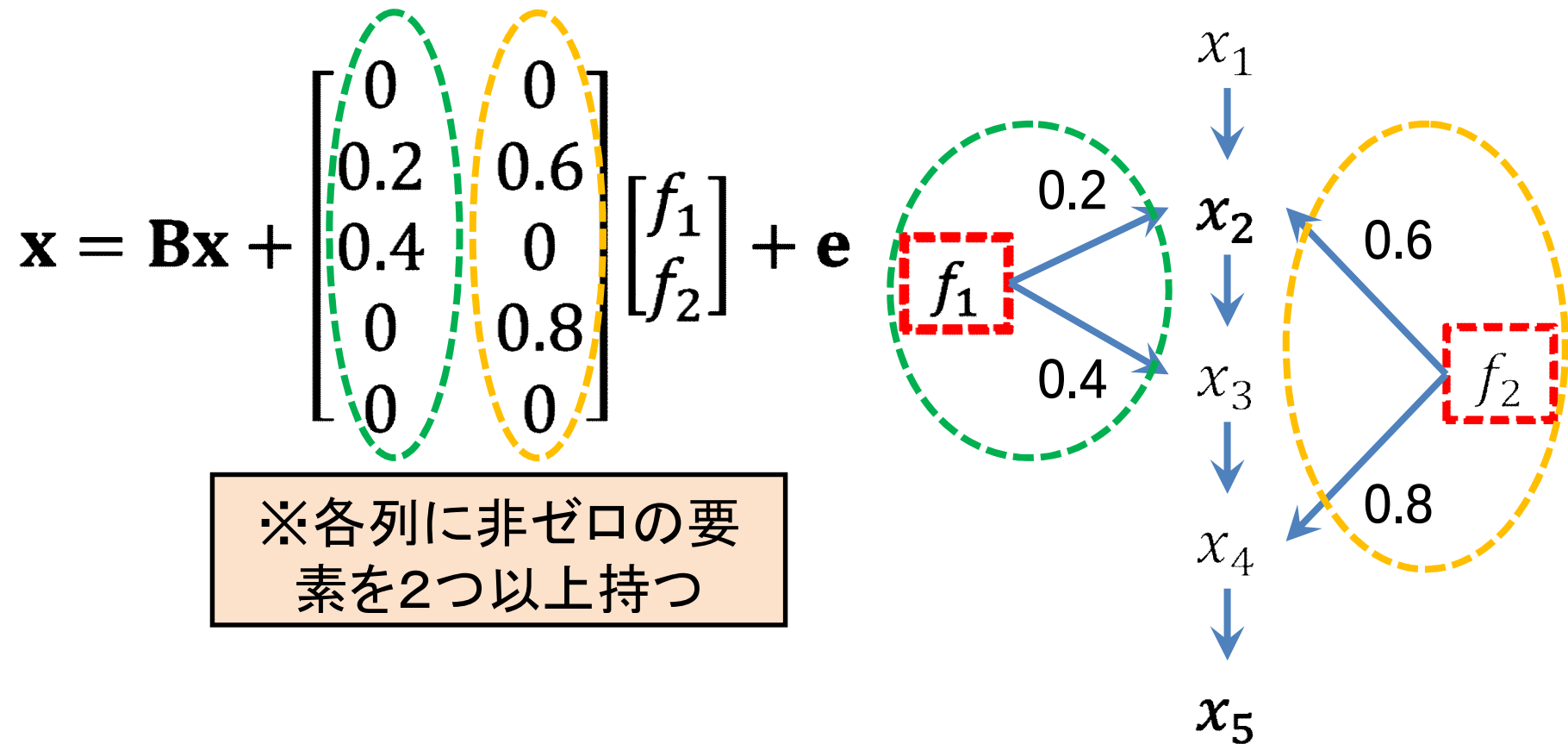
$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{\Lambda}\mathbf{f} + \mathbf{e}$$

$\mathbf{\Lambda}$: 潜在交絡変数からの
影響を表す行列

$\mathbf{f}$: 潜在交絡変数 $f_i (j = 1 \dots q)$
を集めたベクトル

- ・互いに独立
- ・非ガウス

LvLiNGAM モデル例



LvLiNGAMの既存推定手法

- LvLiNGAMアルゴリズム [Hoyer et al., 2008]
 - 過完備独立成分分析 (Overcomplete ICA) を基に係数行列 B を推定.
 - 計算が非常にハードなため4変数以上の推定は困難.
- PairwiseLvLiNGAM [Entner et al., 2011]
 - 潜在交絡変数から直接影響を受けていない2変数を探索し, その2変数間の順序を推定.
 - 3変数以上の因果的順序は推定しない.

提案

提案手法

- 潜在交絡変数の影響を受けていない部分の因果的順序を推定.
 - 潜在交絡変数に対してロバスト.
- 独立性検定 (HSIC [Gretton, et. al, 2010]) の導入.
 - DirectLiNGAMではある独立性指標のみを使用.
 - より厳密に外生変数・終点変数かどうかを判定.
- 因果的順序を上下から学習.
 - DirectLiNGAMでは上からのみ.
- LvLiNGAM ($\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{\Lambda}\mathbf{f} + \mathbf{e}$) の係数行列 $\mathbf{B}$ の要素を可能な限り求める.

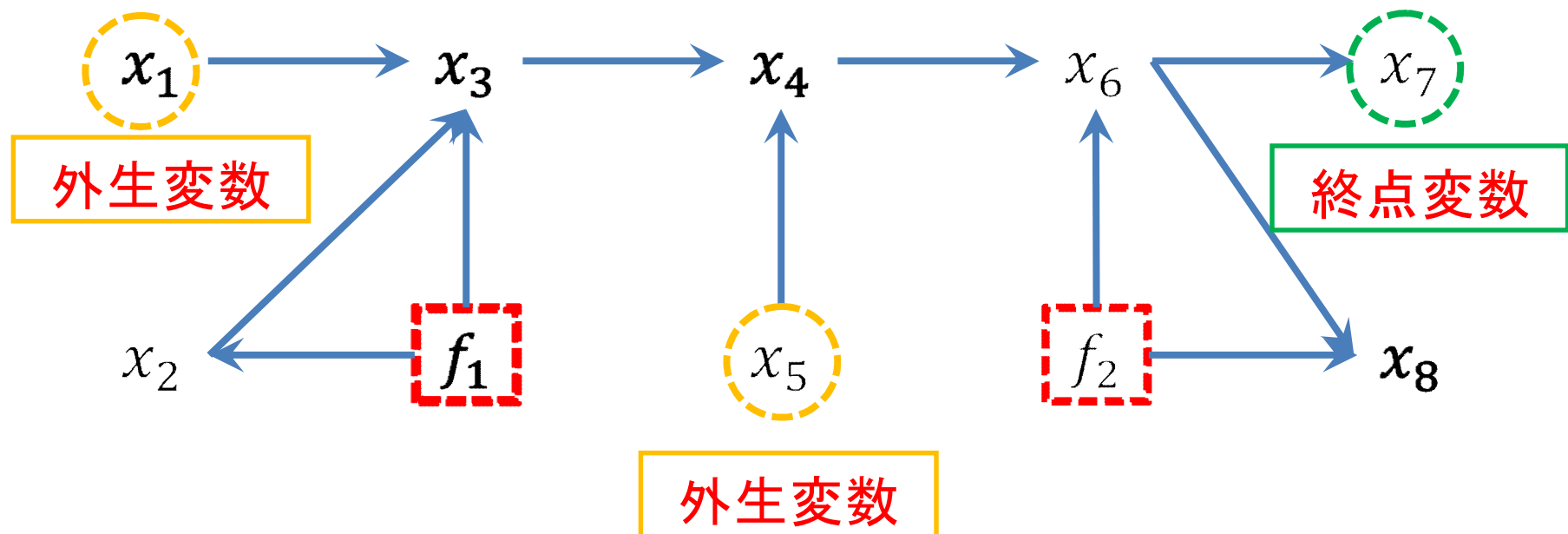
外生変数と終点変数

- 外生変数

- 親に観測変数と潜在交絡変数を持たない変数.

- 終点変数

- 親に潜在交絡変数, 子に観測変数を持たない変数.



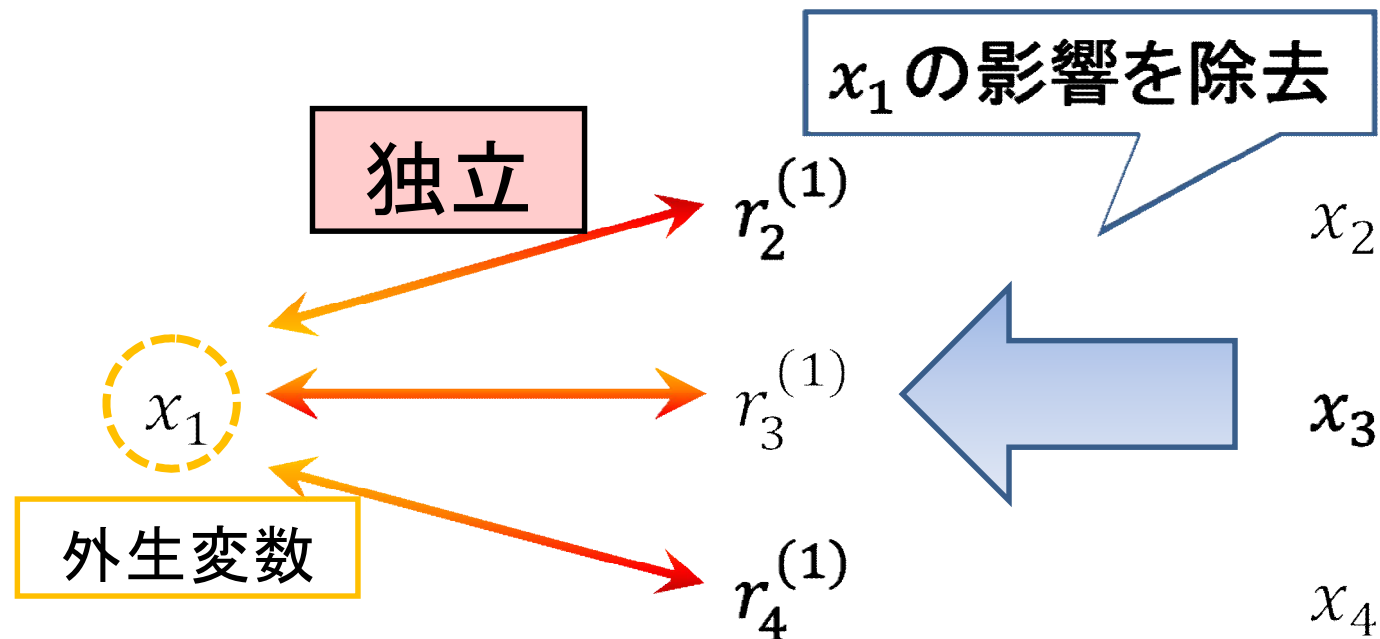
外生変数の同定

補題1.

- 入力データ $\mathbf{x}$ がLvLiNGAMに従うとき, x_i を x_j で回帰したときの残差 $r_i^{(j)}$ は次式で得られる:

$$r_i^{(j)} = x_i - \frac{\text{cov}(x_i, x_j)}{\text{var}(x_j)} x_j \quad (i \neq j)$$

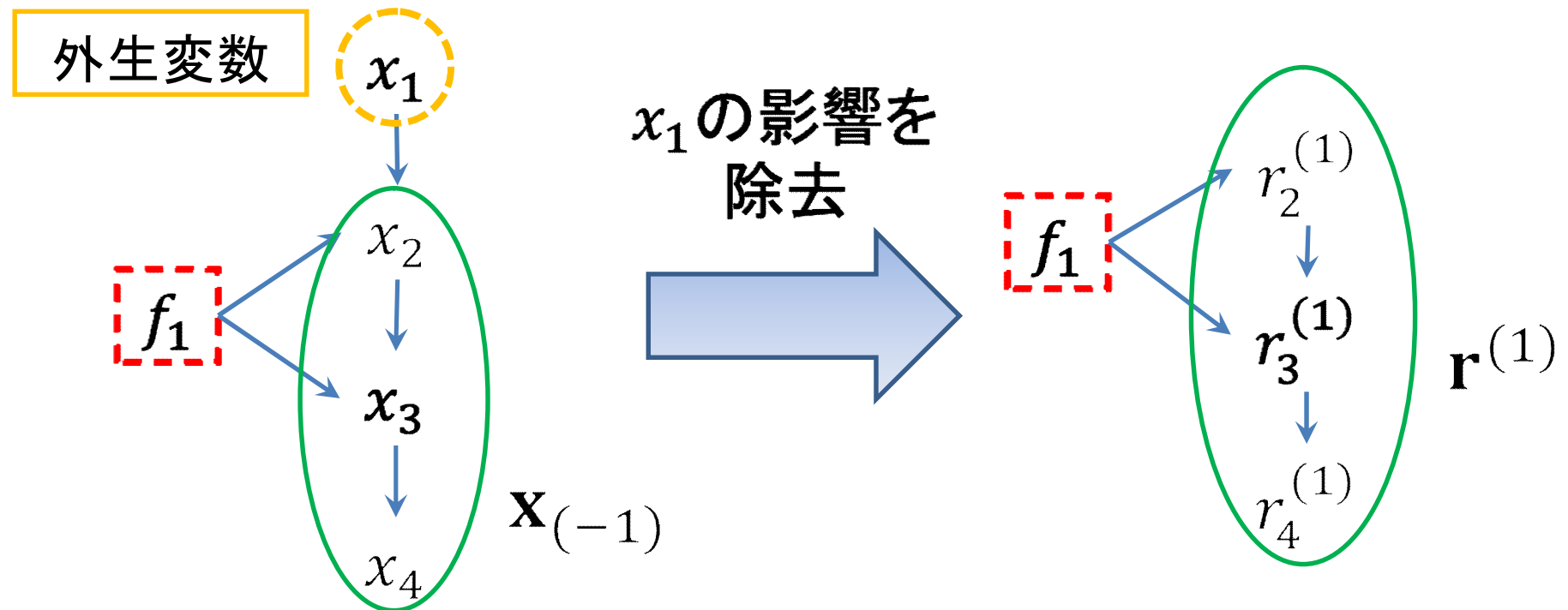
このとき, x_j と($i \neq j$)となる全ての残差 $r_i^{(j)}$ が独立ならば, x_j は外生変数であり, 親に観測変数, 潜在交絡変数を持たない.



外生変数の影響除去

• 補題2.

- 入力データ $\mathbf{x}$ がLvLINGAMに従うとき, x_j が外生変数であるとき, x_j 以外の変数を集めたベクトルを $\mathbf{x}_{(-j)}$, 残差 $r_i^{(j)} (i \neq j)$ を集めたベクトルを $\mathbf{r}^{(j)}$ としたとき, $\mathbf{x}_{(-j)}$ と $\mathbf{r}^{(j)}$ は同じ因果構造を持つ.



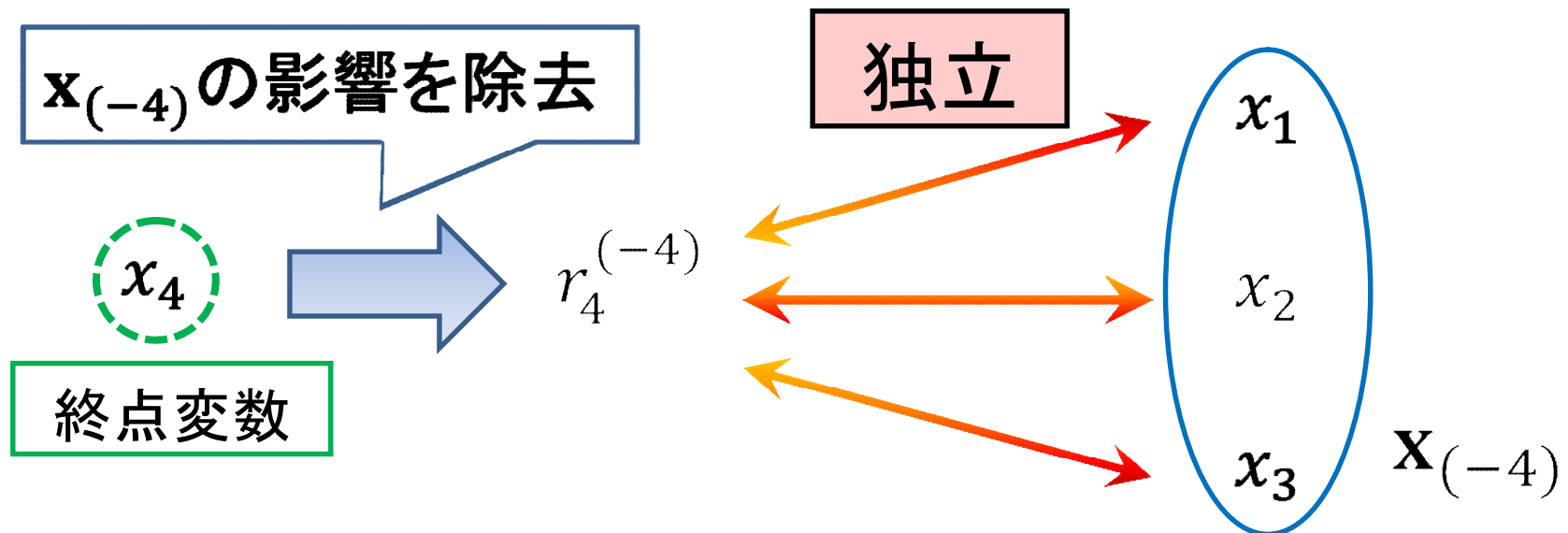
終点変数の同定

補題3.

- 入力データ $\mathbf{x}$ がLvLINGAMに従うとき, x_j 以外の変数を集めたベクトルを $\mathbf{x}_{(-j)}$ とすると, x_j を $\mathbf{x}_{(-j)}$ で回帰した時の残差 $r_j^{(-j)}$ は次式で得られる:

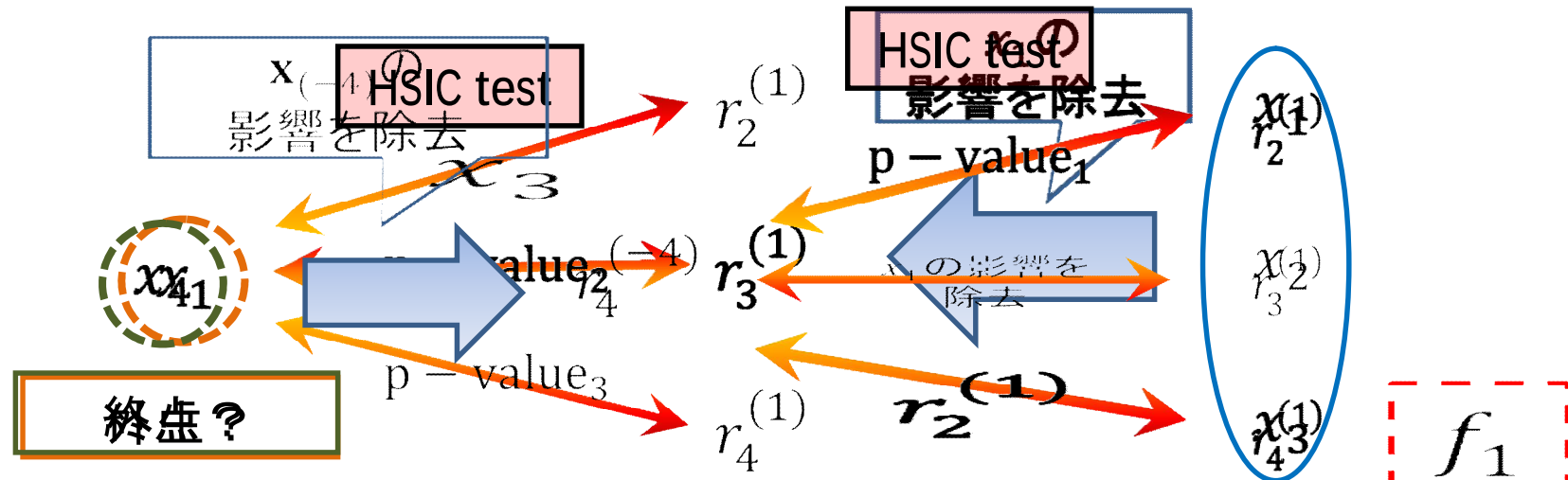
$$r_j^{(-j)} = x_j - \sigma_{(-j)j}^T \sum_{(-j)}^{-1} \mathbf{x}_{(-j)}$$

このとき, $\mathbf{x}_{(-j)}$ が残差 $r_j^{(-j)}$ と独立であるならば, x_j は終点変数であり, 親に潜在交絡変数, 子に観測変数を持たない.



独立性検定

- ノンパラメトリック独立性検定: HSIC. [Gretton, et. al, 2010]
 外生: 一つの変数と複数の残差間で独立性検定を行う.
 終点: 一つの残差と複数の変数間で独立性検定を行う.



- Fisherの手法
 - 各P値を用いて統計量 X^2 を次式で計算.

$$X^2 = -2 \sum_{i=1}^k \ln(p_k) \sim \chi_{2k}^2$$

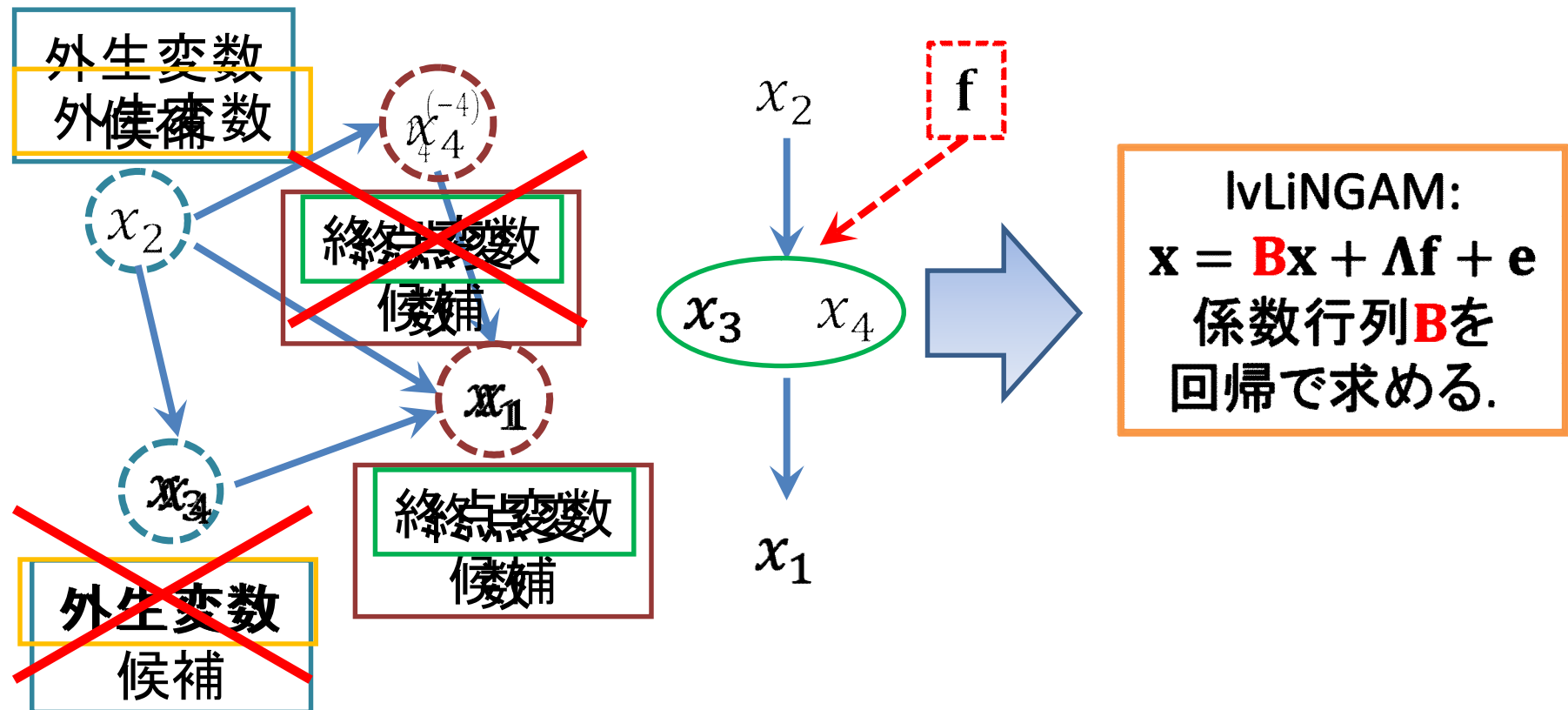
$$r_2^{(1)}$$

- X^2 はカイ二乗分布に従うため, 新たなP値が計算できる.

提案手法 - アルゴリズム

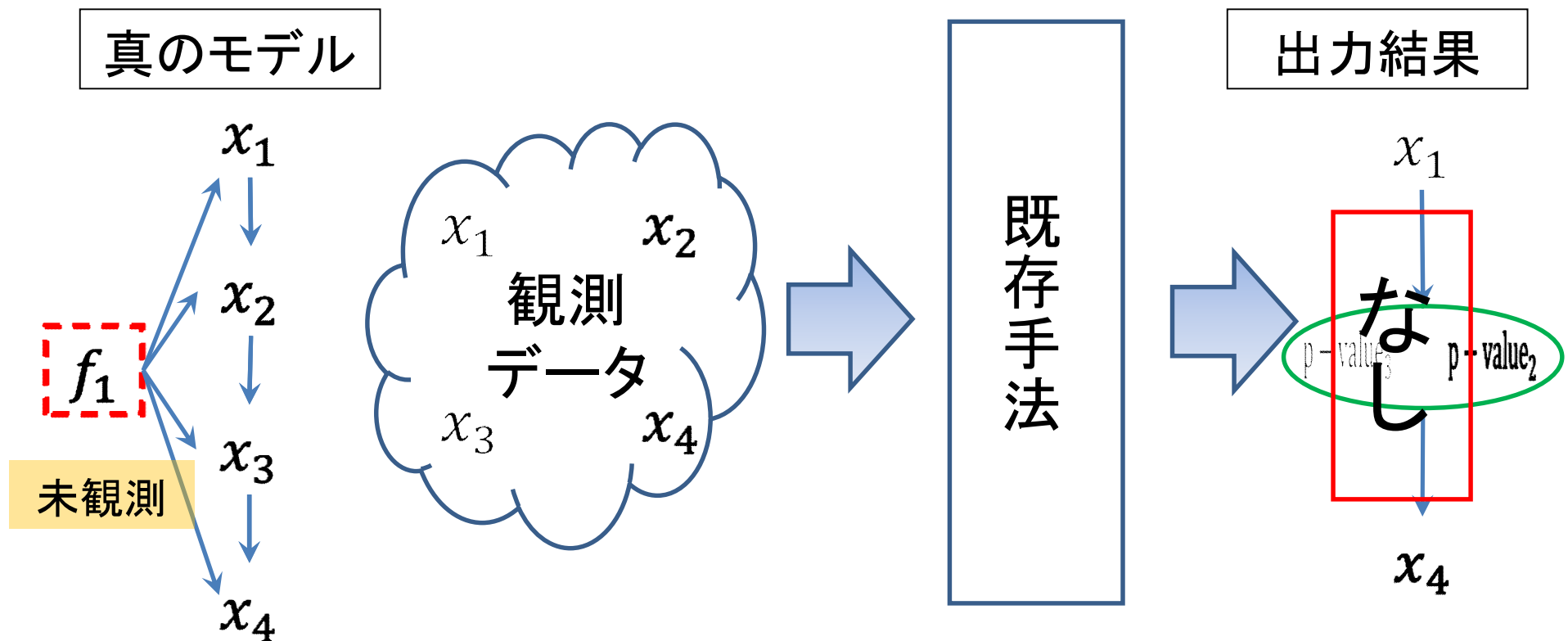
17

1. 補題1と補題2を適用して因果的順序を上から学習する.
2. 上からの探索で検定が棄却された場合, 補題3を適用して因果的順序を上から学習する.
3. 回帰を用いてLvLiNGAM($\mathbf{x} = \mathbf{B}\mathbf{x} + \Lambda\mathbf{f} + \mathbf{e}$)の係数行列 $\mathbf{B}$ を可能な限り求める.



提案手法の問題点

- 因果的順序を全く推定できない場合がある。
 - Future work : このような場合でもより多くの因果的順序を推定すること.

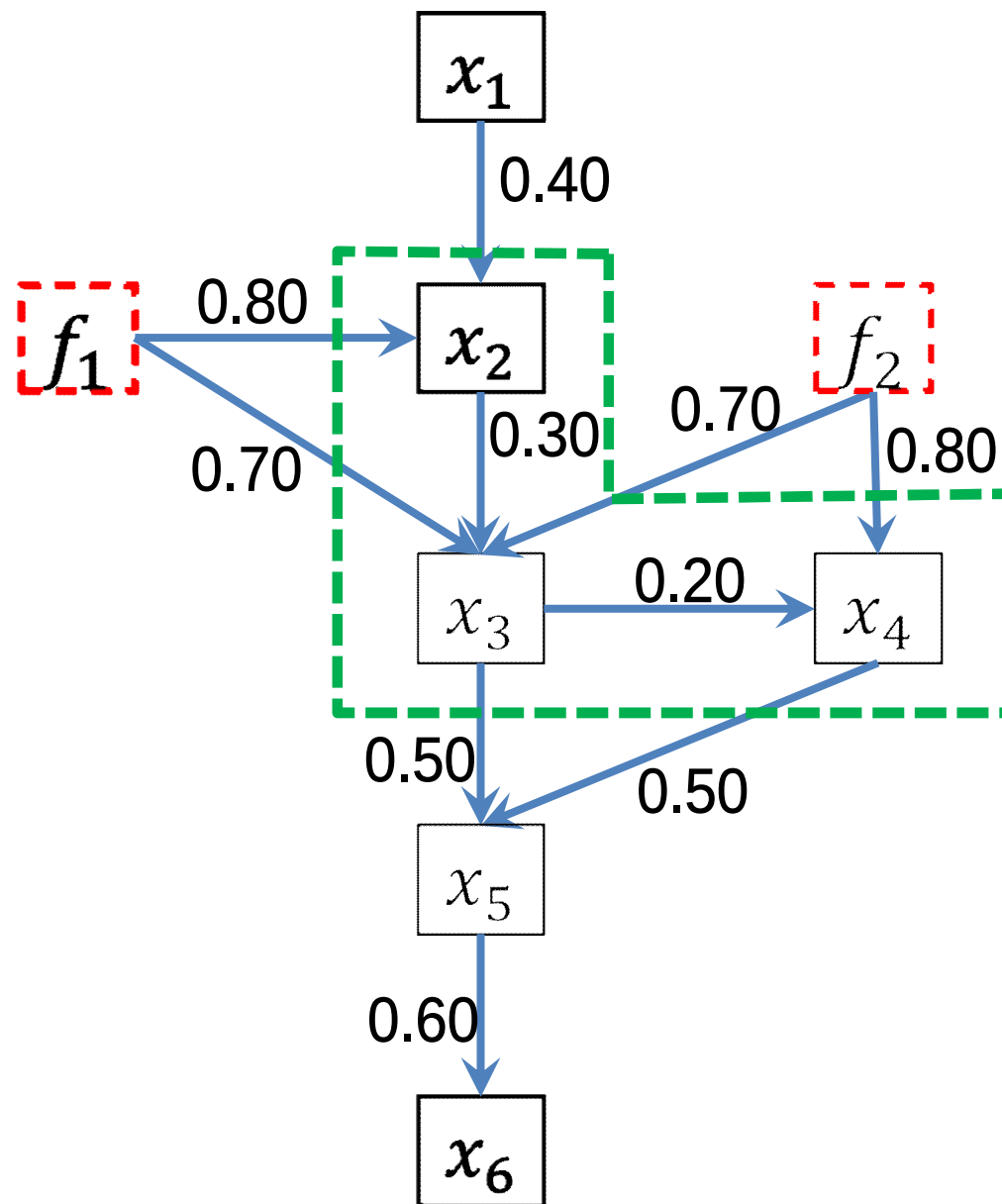


評価実験

評価実験 - 設定

- 提案手法, DirectLiNGAM, PairwiseLvLiNGAMの3つの手法の比較実験を行う.
 - 予め設定したネットワークからデータを生成.
 - 試行回数...100回
 - サンプルサイズ...500, 1000, 2000
 - 検定の有意水準 $\alpha = 0.05$
 - x_i と e_i の分散比は2:1
 - e_i, f_j は下記のいずれかの非ガウス分布で生成.
 - 2つのガウス分布の多峰型対称混合分布
 - 2つのガウス分布の多峰型非対称混合分布
 - 二重指数分布(ラプラス分布)

設定したネットワーク



評価方法

- 因果的順序の評価

- 2変数間の因果的順序を**適合率**と**再現率**で評価.

$$\text{適合率} = \frac{\text{(正しく推定できた2変数間の因果的順序の数)}}{\text{(推定した2変数間の因果的順序の数)}}$$

$$\text{再現率} = \frac{\text{(正しく推定できた2変数間の因果的順序の数)}}{\text{(潜在交絡変数を共有していない2変数間の真の因果的順序の数)}}$$

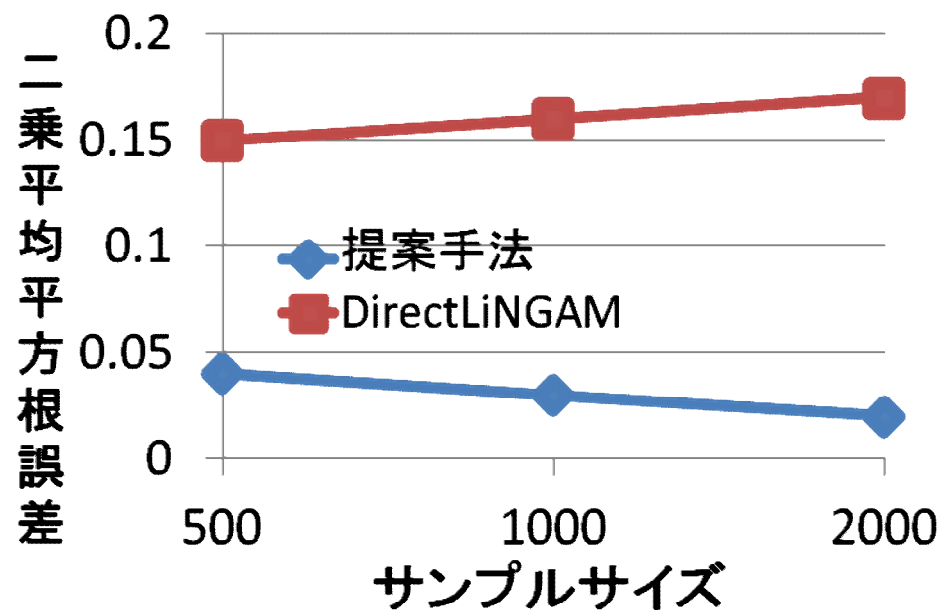
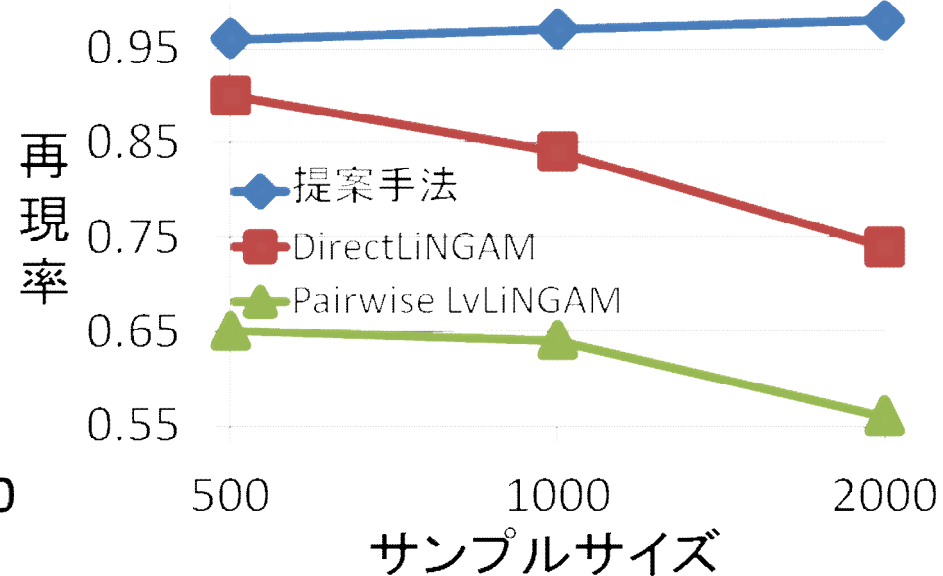
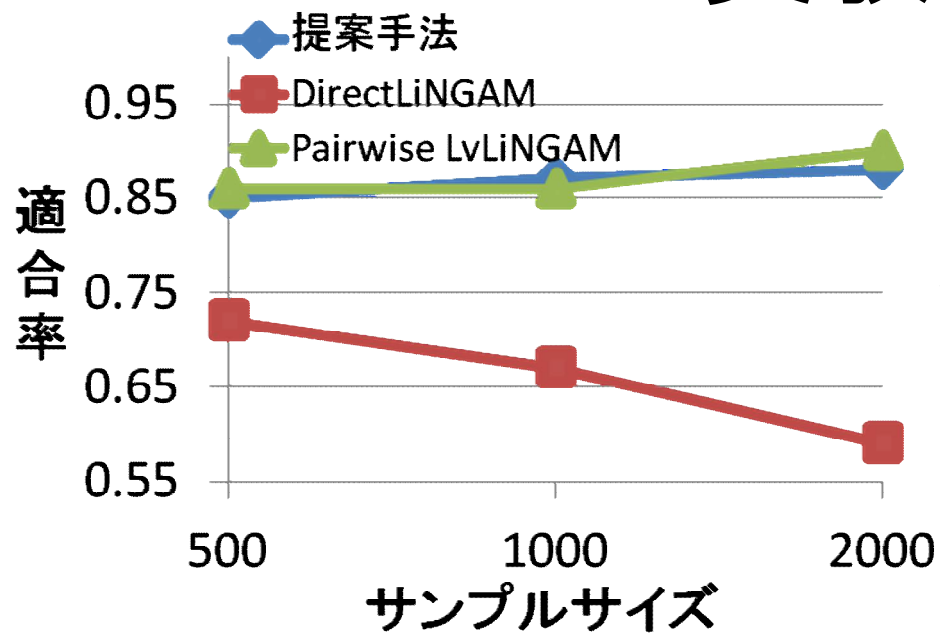
- 係数行列Bの推定精度の評価

- 推定できたBの要素の**二乗平均平方根誤差**

$$\text{RMSE}(\mathbf{B}) = \sqrt{\sum_g \sum_{i,j} \left(b_{ij(\text{true})}^{(g)} - b_{ij(\text{est})}^{(g)} \right)^2 / (\text{推定した要素の数})}$$

※ g : 試行回数

実験結果



※ PairwiseLvLiNGAM はBを推定しない.

まとめ

- LiNGAMを推定する既存手法では、仮定を満たしていないとき全く異なる結果を出力する場合があった.
- 提案手法では潜在交絡変数を考慮することで、既存手法よりも信頼性が向上.
- 実験結果では既存手法よりも良い推定結果を得ることができた.