Statistical analysis

Linear Algebraic Subroutine Library
V0.1 documentation

重回帰モデル
multiple regression model

$y = b_0 + b_1 x_1 + \cdots + b_p x_p + c$

目的変数yをp個の説明変数$x_1 \cdots x_p$で予測するモデル

観測値(教師データ）を元に$b_0 \cdots b_p$を推定する

index	y	x_1	x_2	$\cdots$	x_p
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1p}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	y_i	x_{i1}	x_{i2}	$\cdots$	x_{ip}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{np}

Test

Table 3: 講義満足度と関連ある変数の観測値.

番号	y	x_1	x_2	x_3	x_4
1	4	2	3	5	4
2	4	3	3	3	4
3	4	1	2	4	4
4	4	1	3	5	3
5	5	2	2	5	5
6	4	4	1	5	4
7	4	2	4	4	4
8	3	4	3	4	3
9	3	2	1	2	3
10	3	5	1	2	4
11	4	2	2	5	5
12	5	4	3	5	4
13	4	2	4	5	4
14	4	4	3	5	5
15	3	2	2	5	3
16	5	2	1	4	5
17	4	2	2	4	4

y = 授業に対する満足度
x1 = 履修の際，シラバスは参考にしたか
x2 = 予習復習はしたか
x3 = パワポの文字や図表は見やすかったか
x4 = 教員の説明は分かりやすかったか

F値:4.211263 > F(0.05)_3,13=[3.41]=>予測に有効であると結論できる

係数・定数項の推定(信頼幅:95.00%)

係数	標準誤差	t値	p値	下限(95.0%)	上限(95.0%)
1.0877	0.8728	1.2463	0.2364	-0.8139	2.9894
-0.0984	0.1081	-0.9105	0.3805	-0.3340	0.1371
0.0375	0.1376	0.2722	0.7901	-0.2624	0.3373
0.1653	0.1402	1.1788	0.2613	-0.1402	0.4709
0.5800	0.1855	3.1274	0.0087	0.1759	0.9841

Reference【scikit-learn】

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	1.0877	0.8728	1.2463	0.2364	−0.8139	2.9894
x_1	−0.0984	0.1081	−0.9105	0.3805	−0.3340	0.1371
x_2	0.0375	0.1376	0.2722	0.7901	−0.2624	0.3373
x_3	0.1653	0.1402	1.1788	0.2613	−0.1402	0.4709
x_4	0.5800	0.1855	3.1274	0.0087	0.1759	0.9841

Test

The Boston Housing Dataset

column name	列名の意味するところ
crim	犯罪率
zn	宅地の割合
indus	非商用地の割合
chas	チャールズ川流域かどうか
nox	窒素酸化物濃度
rm	平均部屋数
age	築年数
dis	ビジネス地区への距離
rad	高速道路へのアクセス指数
tax	固定資産税
ptratio	学生と教師の割合
black	黒人の割合
lstat	低所得者の割合
mdev	住宅価格の中央値

$$\text{住宅価格中央値} = \beta_0 + \beta_1 \times \text{犯罪率} + \beta_2 \times \text{住宅の割合} + \cdots + \beta_{13} \times \text{低所得者層の割合}$$

重回帰モデルとスパースモデル

重回帰モデル

係数	
(intercept)	22.5328
	-0.9290
	1.0826
	0.1409
	0.6824
	-2.0587
	2.6769
	0.0195
	-3.1072
	2.6644
	-2.0785
	-2.0626
	0.8501
	-3.7473

Reference【scikit-learn】

```
22.5328063241
[-0.92906457  1.08263896  0.14103943  0.68241438 -2.05875361  2.67687661
  0.01948534 -3.10711605  2.6648522  -2.07883689 -2.06264585  0.85010886
 -3.74733185]
```

ボストンの住宅価格には「ビジネス地区への距離」と
「低所得者層の割合」が一番影響を与えていることがわかる・・・

自動変数選択（冗長な変数を自動的に削除）

スパースモデル

係数	
(intercept)	22.5328
	0.0000
	0.0000
	0.0000
	0.0000
	0.0000
	2.7153
	0.0000
	0.0000
	0.0000
	0.0000
	-1.3443
	0.1803
	-3.5469

Reference【scikit-learn】

```
22.5328063241
[-0.          0.         -0.          0.         -0.          2.71517992
 -0.         -0.         -0.         -0.         -1.34423287  0.18020715
 -3.54700664]
```

ボストンの住宅価格には「平均部屋数」と
「低所得者層の割合」が一番影響を与えていることがわかる

スパース回帰
Sparse regression model

Ridge回帰　（L2正則化）

$$\boldsymbol{b} = \arg\min_{b \in R^p} (\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + \lambda \|\boldsymbol{b}\|^2)$$

Lasso回帰　（L1正則化）

$$\boldsymbol{b} = \arg\min_{b \in R^p} (\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + \lambda \|\boldsymbol{b}\|)$$

ElasticNet回帰　（L1+L2正則化）

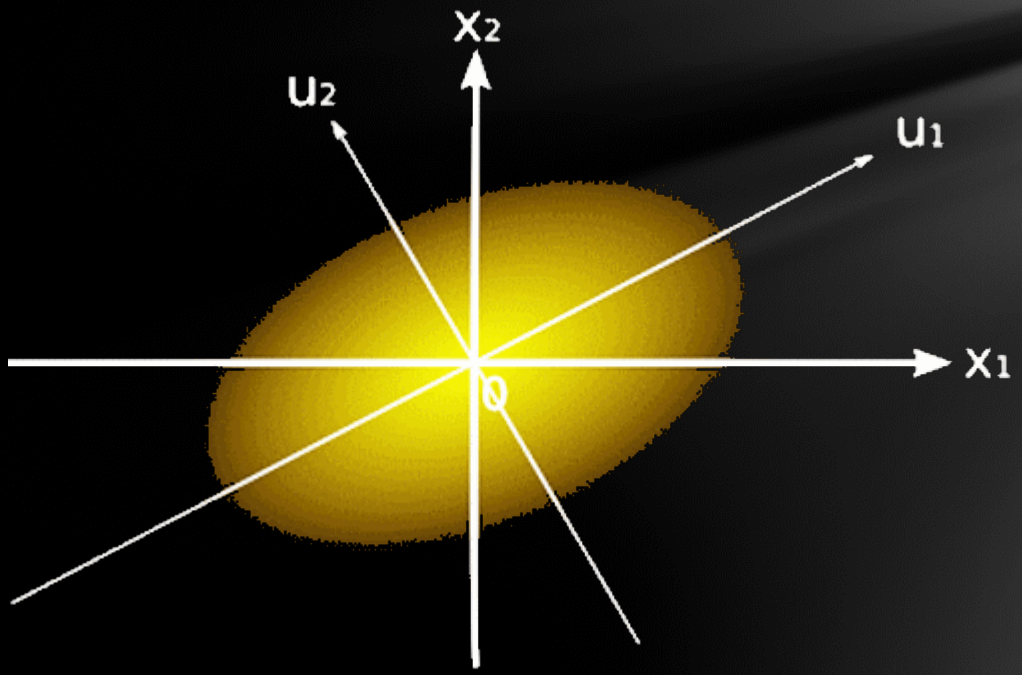
$$\boldsymbol{b} = \arg\min_{b \in R^p} (\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + \lambda_2 \|\boldsymbol{b}\|^2 + \|\boldsymbol{b}\|)$$

主成分分析
PCA: Principle Component Analysis

PCA

$$\boldsymbol{b} = \arg\min_{b \in R^p} (\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2), \boldsymbol{b}^T\boldsymbol{b} = \boldsymbol{I}$$

bは元のデータを再現しbは直交（bの各列ベクトルは互いに直交する。



独立成分分析
ICA: Independent Component Analysis

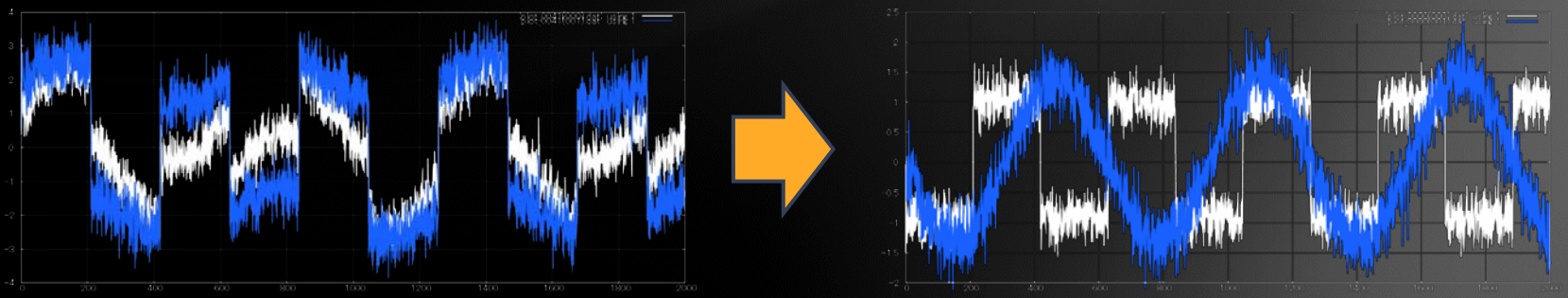
$$\boldsymbol{b} = \arg\min_{b \in R^p} (\|\boldsymbol{y} - \boldsymbol{Xb}\|^2) - \lambda \boldsymbol{H}(\boldsymbol{X})$$

bは元のデータを再現しH(X)が最大となるようなb。
H（X)はXのエントロピー。

$$\boldsymbol{y} = \boxed{\boldsymbol{Xb}}$$

X もbも未知

混合シグナルyのみ観測しているとき、元シグナルXと変換マトリクスを推定する

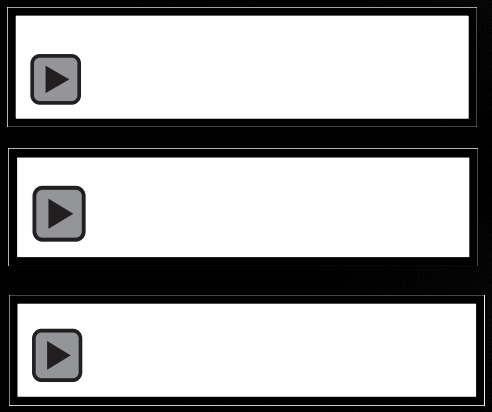


Test

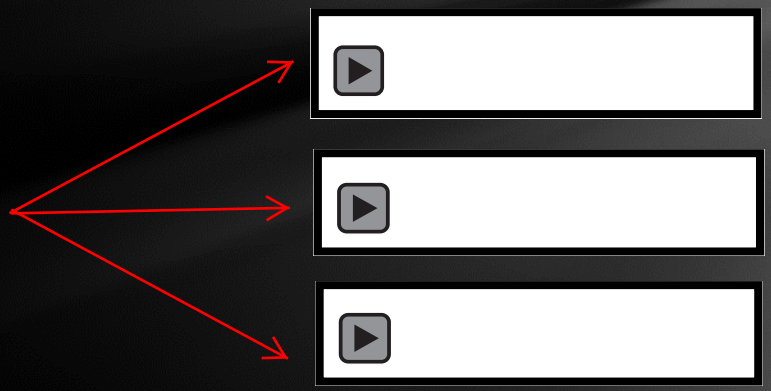
お疲れ様です。こんばんは

はじめまして、こんにちは
おはようございます。
お疲れ様です。こんばんは

混合された音声



分離された音声



統計的因果探索
LiNGAM

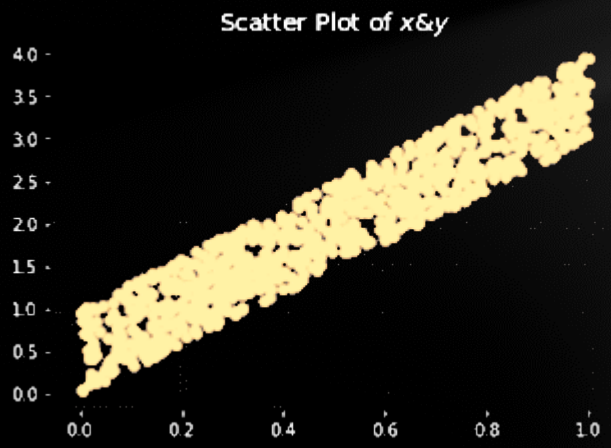
LiNGAM（linear non-Gaussian acyclic model）モデル

$$x = Bx + e$$

$$B = I - D^{-1}P^{-1}W_{ICA}$$

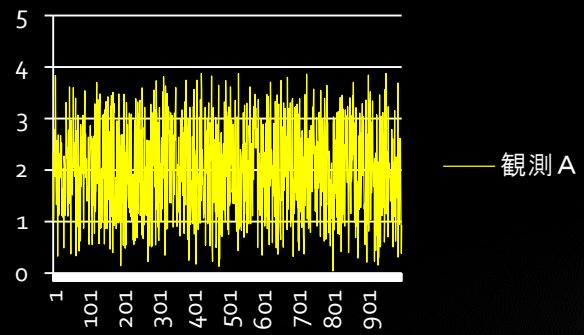
データ生成過程の構造に関する背景知識がない場合に因果関係を推定

XからYを生成したのか？
YからXを生成したのか？

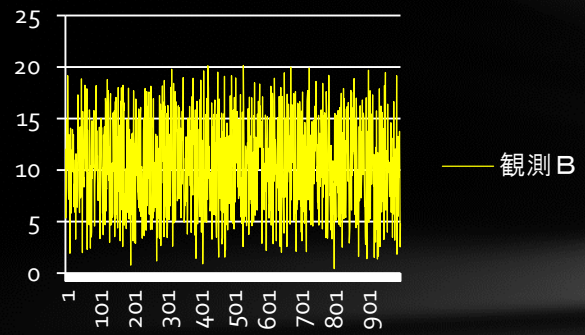


XからYを生成したはず

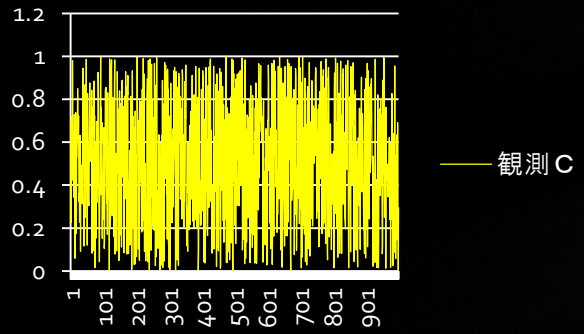
観測Ａ



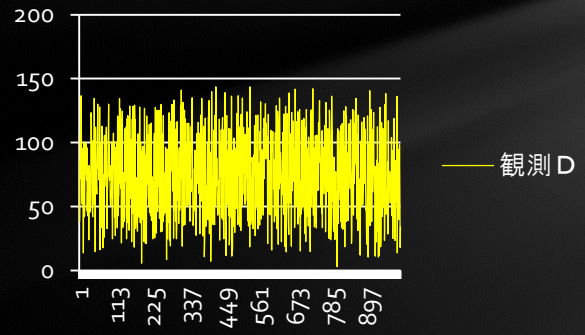
観測Ｂ

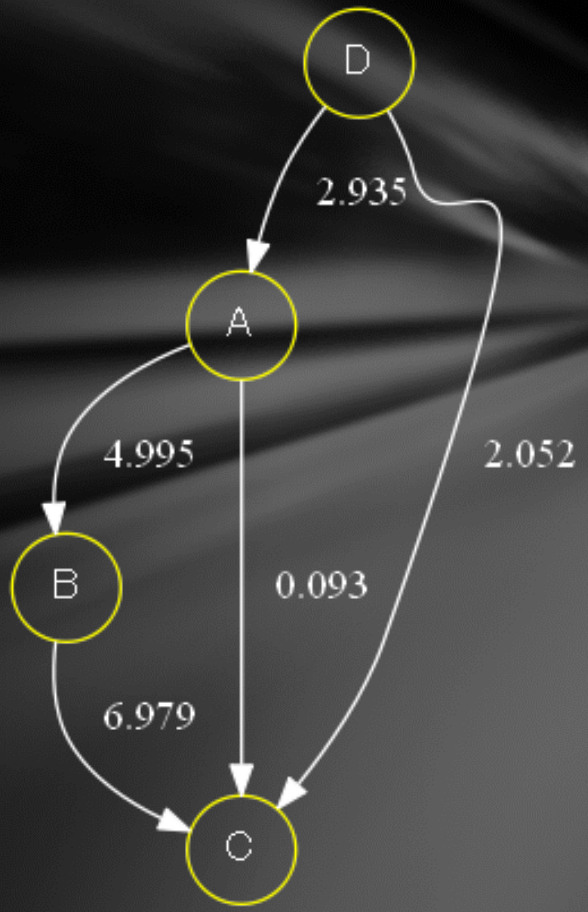


観測Ｃ



観測Ｄ





データＤ生成過程がＡ，Ｂ，Ｃのデータ生成原因になっている。

CLAPACK

CLAPACK (f2c'ed version of LAPACK)
ビルド済みのバイナリ
http://www.netlib.org/clapack/LIB_WINDOWS/prebuilt_libraries_windows.html

重回帰モデル
multiple regression model

$$y_i = b_0 + b_1 x_{i1} + \cdots + b_p x_{ip} + c_i$$

index	y	x_1	x_2	$\cdots$	x_p
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1p}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	y_i	x_{i1}	x_{i2}	$\cdots$	x_{ip}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{np}

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} + \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix}$$

$$\boldsymbol{y = Xb + c}$$

$$\sum_{i=1}^{n} \boldsymbol{c_i}^2 = \boldsymbol{c^t c} = (\boldsymbol{y - Xb})^t (\boldsymbol{y - Xb})$$

$$\|c\|^2 = \|\boldsymbol{y - Xb}\|^2$$

ゼロには出来ないが最小にする

$$\boldsymbol{b} = \arg\min_{b \in R^p} \|\boldsymbol{y - Xb}\|^2$$

重回帰モデル
multiple regression model

fit と fit2の違い

fit はマトリクスXがフルランクを仮定している。
fit2 は特異値分解（SVD）を使用してマトリックスはランク不足も考慮している。

基準化(normalization)

説明変数同士を同じ尺度で評価するためにXの各説明変数毎に基準化(平均値0分散1)

$$x_i = \frac{x_i - \mu}{\sigma}$$

最小二乗推定に失敗する場合

正則化(regularization)→スパース回帰参照