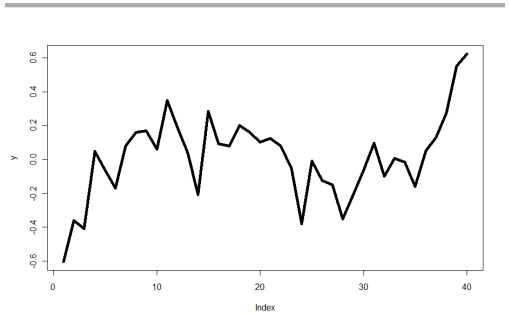
不完全データの一部に興味がある場合の情報量規準

東工大 情報理工学研究科
原　照雅　　下平　英寿

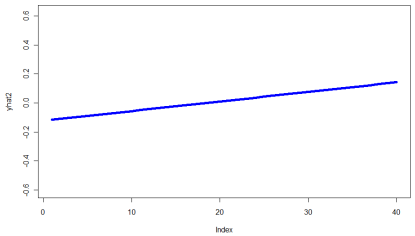
発表の概要

- 情報量規準とは
- 潜在データを考慮した情報量規準とは
  - 完全データを重視できる情報量規準PDIO
  - 完全データの一部を重視できる情報量規準PDIOp
- 数値実験

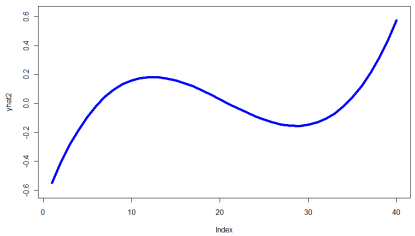
予測のためにモデル選択を行う



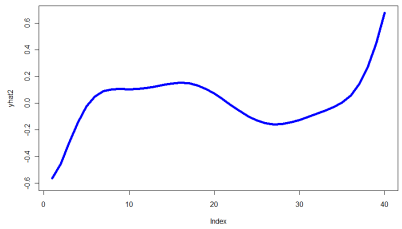
モデル1



モデル2

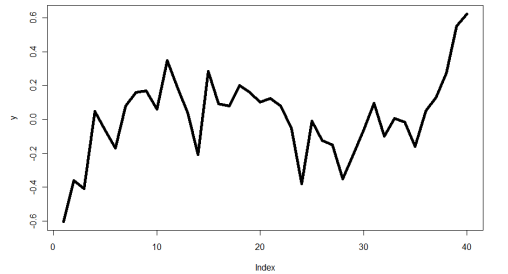


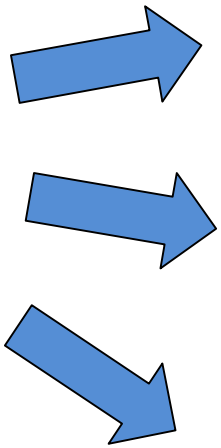
モデル3



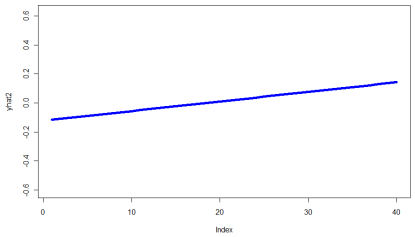
- データの予測のためには
  適切なモデルを選ぶ必要がある
  - データは例えば株価など

情報量規準でモデルの良さを測ることができる

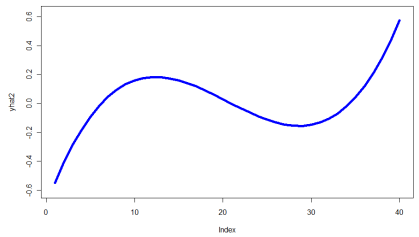




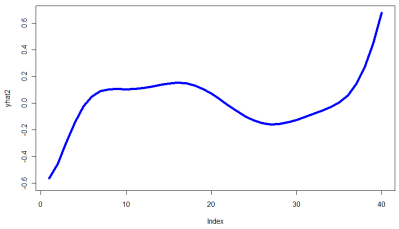
モデル1
AIC=97.7



モデル2
AIC=37.1



モデル3
AIC=46.2



- 赤池情報量規準(AIC)が特に有名
  - しかし、AICでは対応できない場合がある
  - まずは、このAICを拡張する

$$\mathrm{AIC} = -2l(\hat{\theta}(y)) + 2m$$

不完全データとは

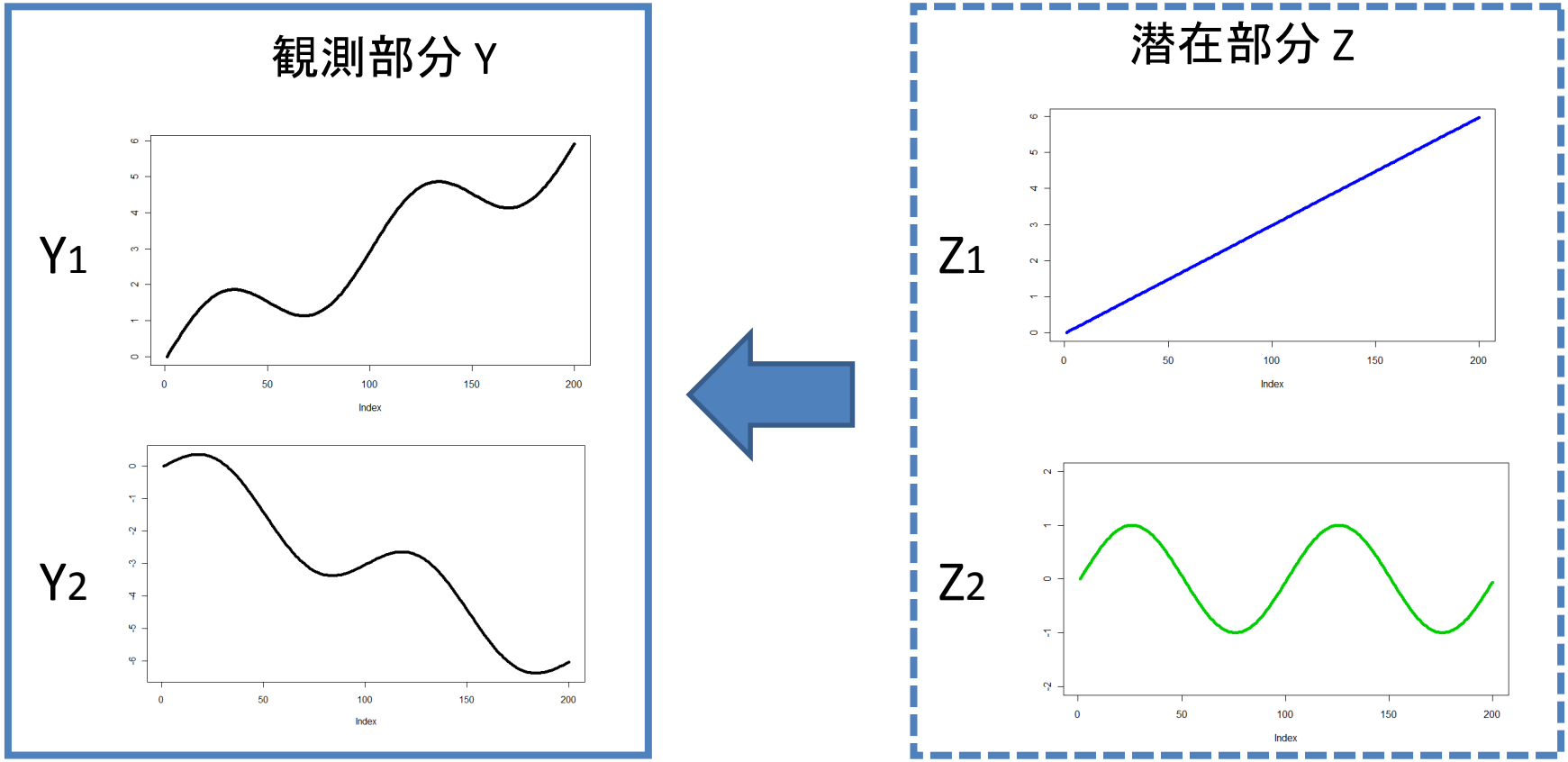
$$X = (Y, Z)$$

完全データ　　観測データ（不完全データ）　　潜在データ（損失データ）

- 不完全データYのみ観測でき、Xは直接観測できない
  - 本当に知りたいのはXの性質
  - これを、$(Y,Z)=((Y_1,Y_2),(Z_1,Z_2))$と分割できる場合を考える

データが分割できる場合

- 観測データ Y_1 , Y_2 はZ_1 , Z_2 から生成されるとする。
  - 例えば、観測データは株価、潜在成分はトレンド、周期成分



観測できる

観測できない！

- 通常はY_1 ,Y_2 , Z_1 , Z_2 について予測を行いたい

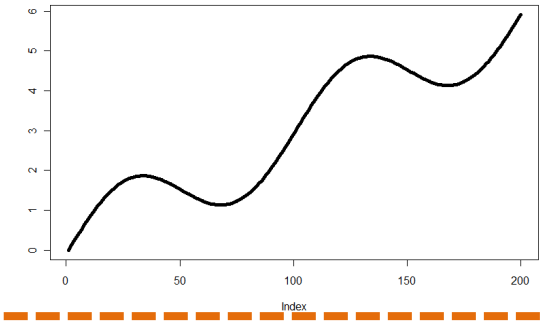
興味がある部分だけ予測したい

- 観測データ Y_1 , 潜在データ Z_1 に興味があるとする

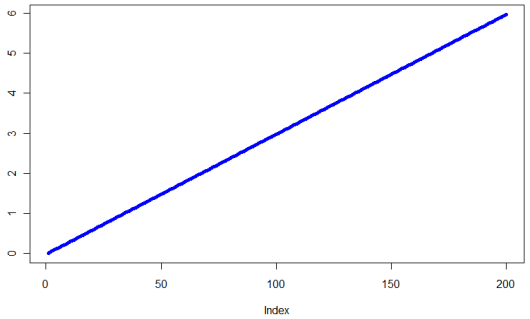
興味がある部分X_1

観測部分 Y　　潜在部分 Z

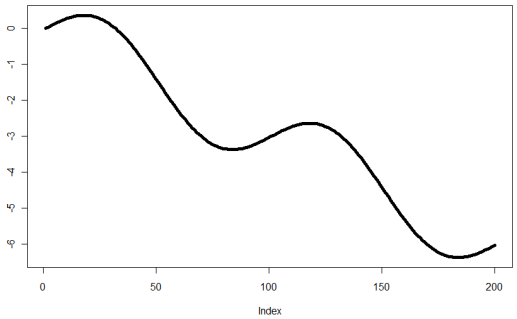
Y_1



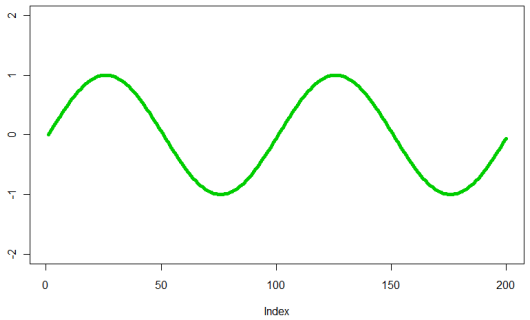
Z_1



Y_2



Z_2

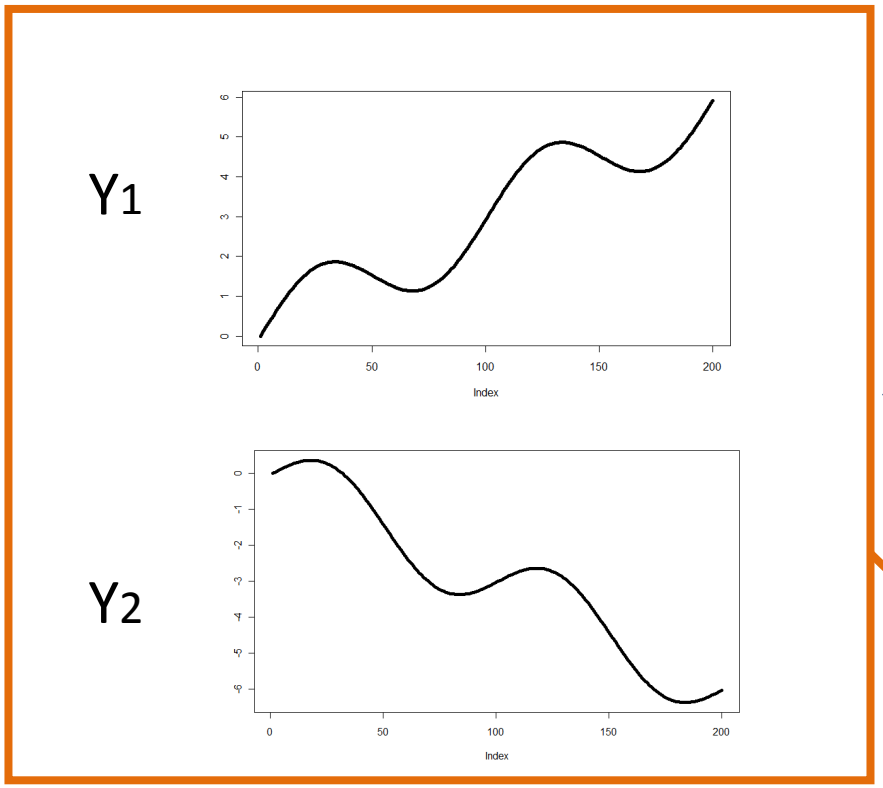


- 本研究では$X_1=(Y_1 , Z_1)$ について予測を行いたい

AICでは観測部分しか評価できない

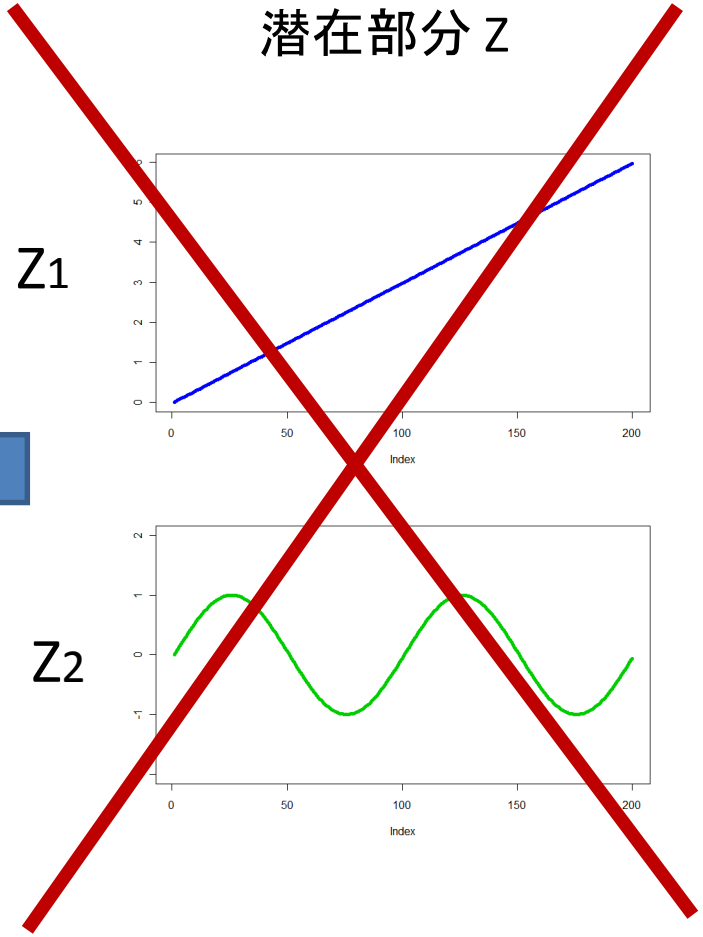
- AIC が重視する部分

観測部分 Y



AIC

潜在部分 Z

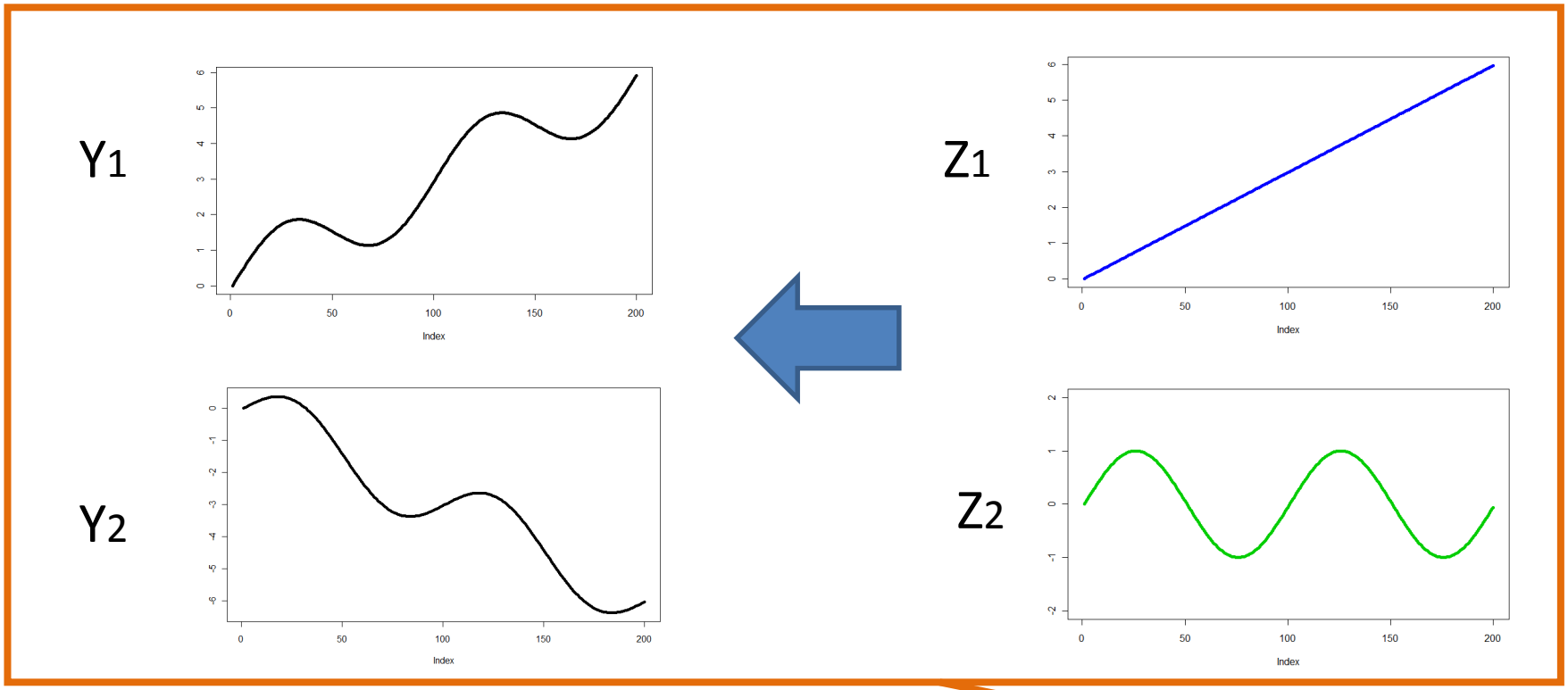


観測部分しか評価できない

PDIOでは潜在部分も含めて評価できる

- PDIO(Predictive Divergence for Indirect Observation)(Shimodaira, 1994)
- PDIOはAICの拡張で、完全データへの当てはまりを考慮する

観測部分 Y　　潜在部分 Z

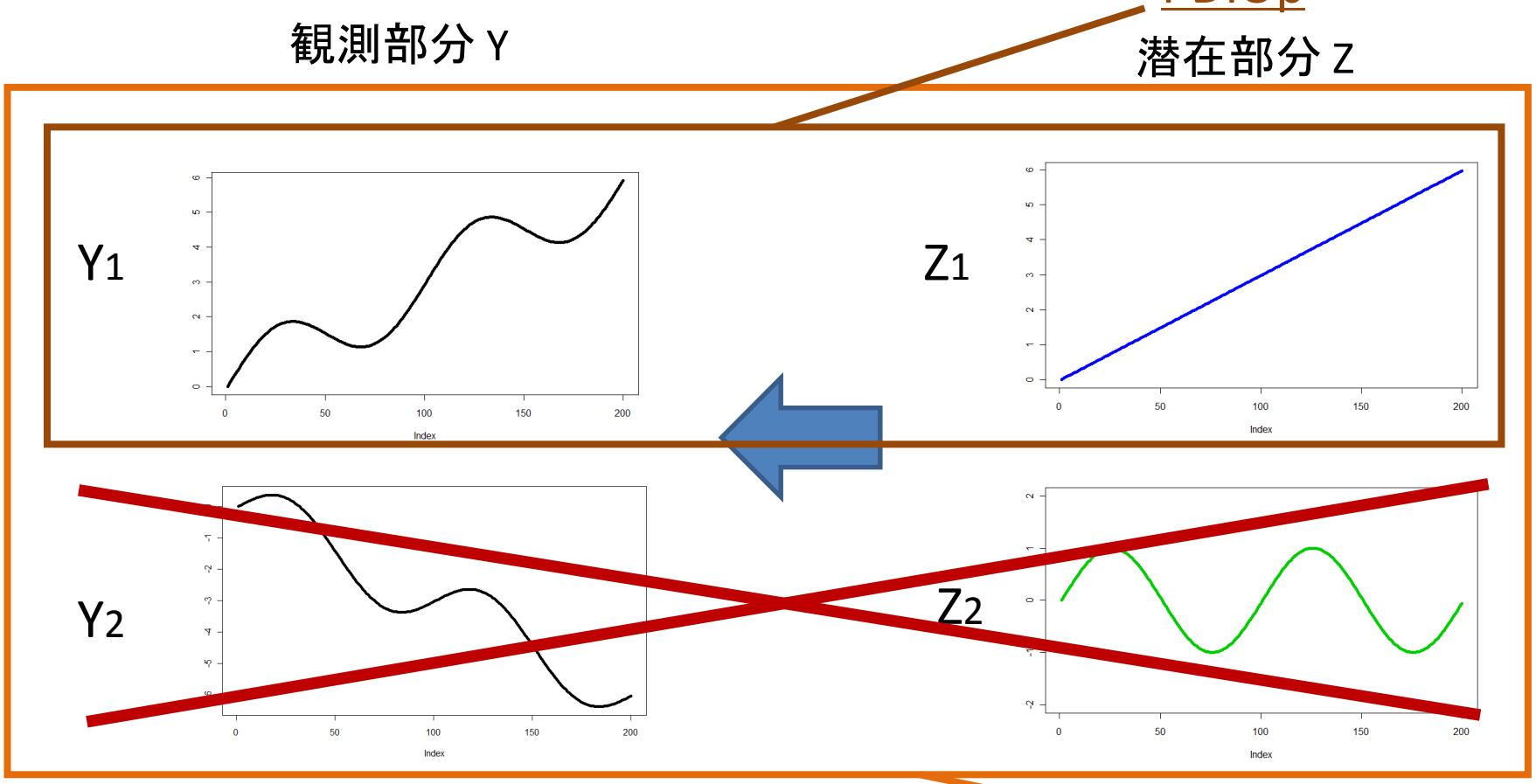


興味がない部分まで評価してしまう

PDIO

PDIOpは興味ある部分を評価できる

- 本研究で導出したPDIOp が重視する部分



情報量規準の計算

$$\mathrm{AIC} = -2l(\hat{\theta}(y)) + 2m$$

$$\mathbf{PDIO} = -2l(\hat{\theta}(y)) + 2tr(H_X H_Y{}^{-1})$$

$$\mathrm{PDIO_P} = -2l_1(\hat{\theta}(y)) + 2tr(H_{X_1} H_Y{}^{-1})$$

- $l(\hat{\theta}(y))$ は最尤推定値での対数尤度、mはパラメタ数
- $l_1(\hat{\theta}(y))$ は周辺分布に関する最大対数尤度
- H_X, H_Y, H_{X_1}はフィッシャー情報行列

数値実験

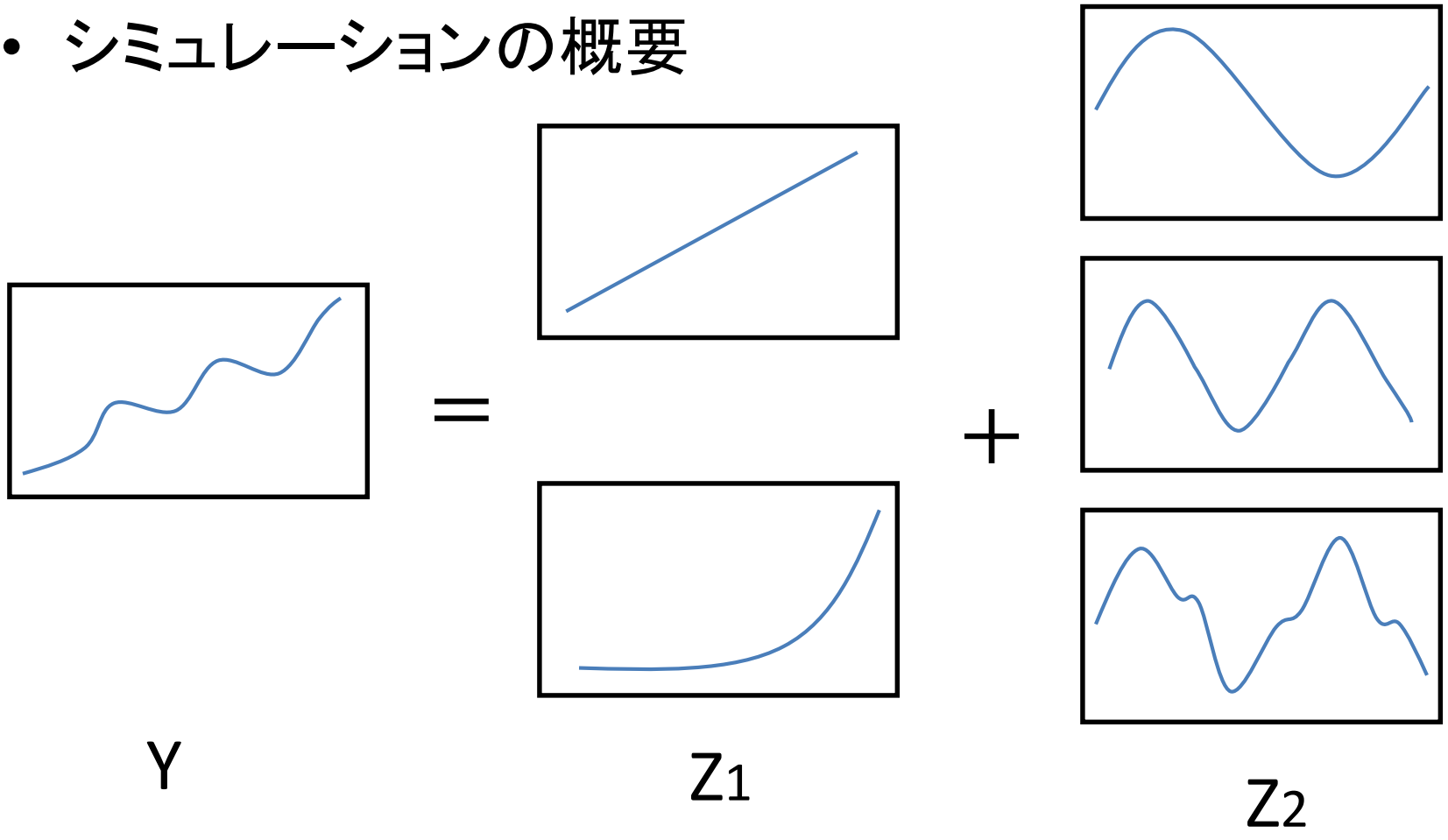
- 回帰モデルによる予測のシミュレーションを行った。

- 観測データYは2つの潜在データから生成されるとする。
  - Z_1: 単調増加（減少）する成分 （トレンド）
  - Z_2: 周期性のある成分 （季節成分）
  - $Y = Z_1 + Z_2 + \varepsilon$,　Z_1, Z_2 は 直接観測できない
  - YとZ_1 の性質を特に知りたいとする

- AIC, PDIO, PDIOpについての予測誤差を比較する

数値実験

- シミュレーションの概要



- ２×３＝６通りのモデルから選ぶ
- 100期分のデータからモデルを決定し、30期分の予測誤差を測定した

実験結果

- 10000回シミュレーションを行い、予測二乗誤差を計測した。

AIC	PDIO	PDIOp
62.8(0.55)	42.9 (0.54)	**38.5**(0.34)

  - 数字はZ_1の予測二乗誤差の平均
  - ( )内は標準誤差

- 予測二乗誤差はPDIOpが一番小さい。

まとめと今後の課題

- PDIOpは、完全データの一部を重視したモデル選択を行う情報量規準である

- 今後の課題としては、PDIOpを用いて回帰モデル以外のシミュレーションを行うこと、実データに応用することなどが挙げられる

参考文献

- Shimodaira H (1994), *A new criterion for selecting models from partially observed data, Selecting Models* from Data: AI and Statistics IV(eds. P. Cheeseman and R. W. Oldford), Lecture Notes in Statistics, 89:21-30, 1994