重みつき最小二乗法で推定した非線形ARXモデルの評価と選択

九州大学
マス・フォア・インダストリ研究所
秦　攀

発表内容

1. 研究動機
2. 基底関数に基づく非線形ARXモデルと重みつき最小二乗法
3. 一般化情報量基準(GIC)によるモデルを評価
4. ロバストなモデル選択法
5. LASSO型の選択法
6. 数値例

1.研究動機

非線形システム同定

図1、非線形システム

- $u(t) \in R$: 外部の入力信号
- $w(t) \in R$: システムの出力信号
- $e(t) \in R$: 観測雑音
- $y(t) \in R$: は観測された出力信号
- **$u(t)$と$w(t)$は非線形の関係を持つ**
- **赤い信号$w(t)$と$e(t)$は観測できないもの**
- **離散時間信号$\{u(t), y(t)\}$のみに基づいて、システムをブラックボックスモデリング(Black-box modeling)する**

1.研究動機

非線形ARXモデル

- **非線形システム同定は離散時間信号$\{u(t), y(t)\}$に基づいて、非線形ARX(nonlinear autoregressive model with exogeneous variables, NARX)モデル**

$$M: y(t) = f(\boldsymbol{x}(t)) + e(t)$$

を求めることである。ここで、

$$\boldsymbol{x}(t) = [u(t-1), \cdots, u(t-l_u), y(t-1), \cdots, y(t-l_y)]^T$$

で定義する。

1.研究動機

モデルの選択：候補から、適切なモデルの構造を選択すること

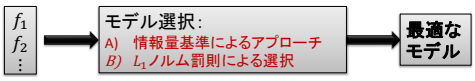


図2、モデル選択

本研究仮定

- $e(t) \sim N(0, \sigma^2)$が**独立同分布**
- モデルは**重みつき最小二乗法(weighted least squares method, WLS)**によって推定

2.基底関数に基づくNARXモデルとRWLS

基底関数に基づくNARXモデル

$$y(t) = \sum_{k=1}^{K} \theta_k b_k(\boldsymbol{x}(t)) + e(t) = \boldsymbol{b}(t)^T \boldsymbol{\theta} + \boldsymbol{e}(t)$$

- $\emptyset_k(\boldsymbol{x}(t)) \in R$は$\boldsymbol{x}(t)$ に関する基底関数(例えば、多項式、動径基底関数、スプラインなど)。
- $\theta_k \in R$は線形的回帰係数。
- $\boldsymbol{b}(t) = [b_1(\boldsymbol{x}(t)), \cdots, b_K(\boldsymbol{x}(t))]^T \in R^K, \boldsymbol{\theta}(t) = [\theta_1(\boldsymbol{x}(t)), \cdots, \theta_K(\boldsymbol{x}(t))]^T \in R^K,$
- **全基底関数候補の集合$\{\emptyset_k, k = 1, 2, \cdots, K\}$から適切な基底関数を選択する**

基底関数の例

- 動径基底関数: $\exp\left\{-\frac{\|x-\mu\|^2}{2h}\right\}$
- 3次スプライン: $\max\{0, (x - t_i)^3\}$

2.基底関数に基づくNARXモデルとWLS

重み付き最小二乗法(weighted least squares method, WLS)

$$\widehat{\boldsymbol{\theta}}_{WLS} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{t=1}^{N} w_t (y(t) - \boldsymbol{b}(t)^T\boldsymbol{\theta})^2 = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} (\boldsymbol{y} - \boldsymbol{B\theta})^T W (\boldsymbol{y} - \boldsymbol{B\theta})$$

$$\widehat{\boldsymbol{\theta}}_{WLS} = (B^T W B)^{-1} B W \boldsymbol{y}$$

$W = I$のとき、最小二乗法

- $w_t \geq 0$は時刻tの観測値につけられる**重み**
- $W = \mathrm{diag}(w_1, \cdots, w_N)$は**重み行列**(対角行列)
- $B = [\boldsymbol{b}(1), \cdots, \boldsymbol{b}(N)]^T$は基底関数による構成された**計画行列**

WLSを使う理由

- ある領域において、よりいい予測精度を期待するとき、その領域の観測値に大きい重みをつける。

例1:計測装置によって、いい計測精度である領域の観測値に大きい重みをつける。

2.基底関数に基づくNARXモデルとWLS

WLSで推定されたNARXモデルの選択問題

1. **最尤法に基づいたAICやBICなどの情報量基準が用いられなくなり**、新しい情報量基準が必要になる。
2. モデルの候補が大量であるとき、**複雑な現象を統一的に説明でき、簡潔なモデルを構築できるロバストな選択手法を提案する**。
3. **LASSO型の選択法をWLSに展開する**。

3.GICによるモデルを評価

情報量基準

$$\mathrm{IC} = -2\sum_{t=1}^{N} \log f(y(t)|\widehat{\boldsymbol{\theta}}) + 2\{\text{corrected-bias}\}$$

分布の種類や推定法などによって決める

一般化情報量基準(generalized information criterion、GIC)

[Konishi and Kitagawa (1996)]

$$\mathrm{GIC} = -2\sum_{t=1}^{N} \log f(y(t)|\widehat{\boldsymbol{\theta}}_M) + 2\mathrm{tr}\{R^{-1}Q\}$$

$$R = -\frac{1}{N}\sum_{t=1}^{N} \frac{\partial \boldsymbol{\psi}(y(t), \boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_M} \in R^k$$

$$Q = -\frac{1}{N}\sum_{t=1}^{N} \boldsymbol{\psi}(y(t), \widehat{\boldsymbol{\theta}}_M)\frac{\partial \log(y(t)|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_M} \in R^k$$

$\widehat{\boldsymbol{\theta}}_M \in R^k$は「M-estimator」であり、汎関数$\sum_{t=1}^{N} \boldsymbol{\psi}(y(t), \boldsymbol{\theta}) = \boldsymbol{0}$の解である。ここで、$\boldsymbol{\psi} \in \boldsymbol{R^k}$は$\boldsymbol{\psi}$関数である。

3.GICによるモデルを評価

WLSに対応するψ関数

$$\sum_{t=1}^{N} \boldsymbol{\psi}(y(t), \boldsymbol{\theta}) = \sum_{t=1}^{N} \frac{\partial}{\partial \boldsymbol{\theta}}\{w_t (y(t) - \boldsymbol{b}(t)^T\boldsymbol{\theta})^2\} = -2B^T W\boldsymbol{y} + 2B^T WB\boldsymbol{\theta} = \boldsymbol{0}$$

$\mathrm{GIC}_{\mathrm{WLS}}$

$$\mathrm{GIC}_{\mathrm{WLS}} = -2\sum_{t=1}^{N} \log f(y(t)|\widehat{\boldsymbol{\theta}}_{WLS}) + 2\mathrm{tr}\left\{(B^T WB)^{-1}\sum_{t=1}^{N} \frac{w_t \epsilon(t)^2}{\widehat{\sigma}^2{}_{WLS}} \boldsymbol{b}(t)\boldsymbol{b}(t)^T\right\}$$

- 対数尤度:$\log f(y(t)|\widehat{\boldsymbol{\theta}}_{WLS}) = -\frac{1}{2}\left\{N\log(2\pi\widehat{\sigma}^2{}_{WLS}) + \sum_{t=1}^{N} \frac{\epsilon(t)^2}{\widehat{\sigma}^2{}_{WLS}}\right\}$
- 推定されたノイズの分散:$\widehat{\sigma}^2{}_{WLS} = \frac{(\boldsymbol{y}-\boldsymbol{B\theta})^T W(\boldsymbol{y}-\boldsymbol{B\theta})}{\mathrm{tr}(W)}$
- 予測誤差:$\epsilon(t) = y(t) - \boldsymbol{b}(t)^T\boldsymbol{\theta}$
- **$\mathrm{GIC}_{\mathrm{WLS}}$を最小するモデルが望ましい**

4.ロバストなモデル選択法

***GIC_{WLS}*に基づく増加法(Forward stepwise method)**

- モデル候補の数は2^Kである。Kが大きいとき、全部の候補を評価することよって最適なモデルを選択するのは困難になる。

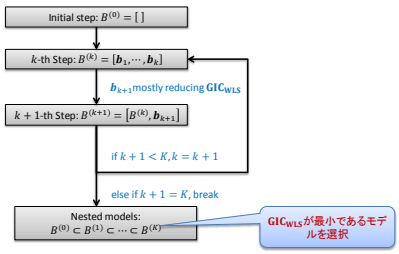


図3、増加法の流れ図

4.ロバストなモデル選択法

増加法のモンテカルロ実験

$\boldsymbol{S_1}$: $y(t) = -0.2y(t-1) + 0.75y(t-2) + 0.25u(t-1) + 0.3u(t-2) + e(t)$

- $u(t) \sim N(0,1)$ $e(t) \sim N(0,0.01)$
- 100個の独立なデータセットを生成し、各データセットの長さが1000である
- そのデータセットを用い、$\mathrm{GIC}_{\mathrm{WLS}}$に基づく増加法を行う(即ち、100回のモンテカルロ実験)
- $\{y(t-1), y(t-2), y(t-3), u(t-1), u(t-2), u(t-3)\}$の2次までの多項式を基底関数とする、全部は27個である。
- $w_t = |y(t)|$,大きい観測値に大きい重みをつける

図4、増加法のモンテカルロ実験

4.ロバストなモデル選択法

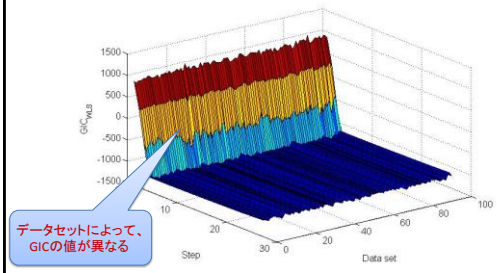


図5、各ステップのGICの値

12

4.ロバストなモデル選択法

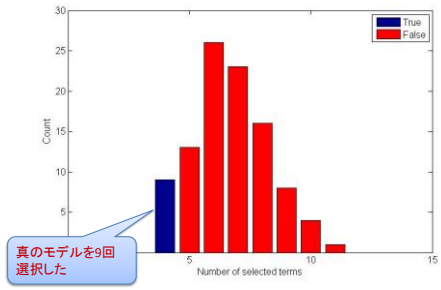


図6、選択された100個のモデルのサイズのヒストグラム

13

4.ロバストなモデル選択法

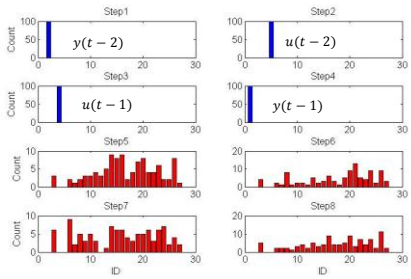
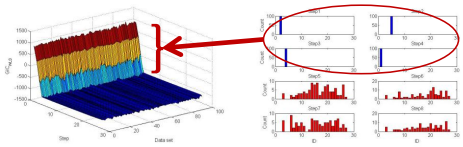


図7、最初の8個のステップで選択された基底関数のヒストグラム

14

4.ロバストなモデル選択法



- データセットによって、各ステップのGICの値が違う
- 必ずしも、真のモデルを選択していない
- Over-parameterizedモデルをよく選択された
- 増加法がロバストではない
- 選択法のロバスト性：独立のデータセットに対して、統一のモデルを選択する確率が高い

15

4.ロバストなモデル選択法

ロバストなモデル選択基準

- $\Pr_j(b(t))$はj番目のステップで基底関数$b(t)$**が選択された確率**を表す。この確率が

$$\Pr_j(b(t)) > \rho$$

を満たすとき、$b(t)$を選ぶ。$0 < \rho < 1$はuser-definedパラメータである。

モンテカルロ法によって$\Pr_j(b(t))$を近似

図8、モンテカルロ法によって$\Pr_j(b(t))$を近似

大量のデータセットを生成するのは困難である。

16

4.ロバストなモデルな選択法

Subsamplingによって$Pr_j(b(t))$を近似

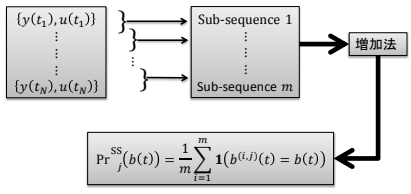


図9、Subsamplingによって$\Pr_j(b(t))$を近似

17

4.ロバストなモデルな選択法

Subsamplingの特徴

- 抽出：

$$X_1^{N_s} = \{y(t), u(t) | t = 1, \cdots, N_s\},$$
$$X_2^{N_s} = \{y(t), u(t) | t = p+1, \cdots, N_s+p\},$$
$$\vdots$$
$$X_m^{N_s} = \{y(t), u(t) | t = (m-1)p+1, \cdots, N_s+(m-1)p\},$$

- $1 \le N_s < N, p \ge 1$
- $p \le N_s$のとき、Sub-sequenceが重なっているので、お互いに相関になる
- モンテカルロ法に比べれば、一つのデータセットのみが必要
- チューニングパラメータN_sとpは結果に影響を与えられる

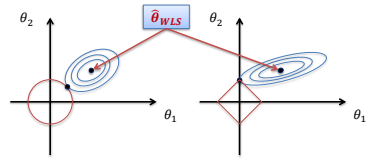
問題点：
- $e(t)$の分布を仮定する
- パラメータの推定とモデル選択を分けて行う
- Subsamplingのため、計算量が増やす

18

5.LASSO型の選択法

重みつきLASSO(least absolute shrinkage and selection operator)推定 [Tibshirani (1996)]

$$\hat{\boldsymbol{\theta}}_{\text{LASSO}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} (\boldsymbol{y} - B\boldsymbol{\theta})^T W (\boldsymbol{y} - B\boldsymbol{\theta}) + \lambda_l \sum_{k=1}^{K} |\theta_k|$$



(a) Ridge regression: $\theta_1^2 + \theta_2^2 < c_1$　　(b) Lasso: $|\theta_1| + |\theta_2| < c_2$

図10、2変数の縮小推定法

19

5.LASSO型の選択法

重みつきLASSO推定の計算

1. 観測データと計画行列を重み行列で調節：$\boldsymbol{y}^* = W^{\frac{1}{2}}\boldsymbol{y}, B^* = W^{\frac{1}{2}}B$

$$\hat{\boldsymbol{\theta}}_{\text{LASSO}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} (\boldsymbol{y}^* - B^*\boldsymbol{\theta})^T (\boldsymbol{y}^* - B^*\boldsymbol{\theta}) + \lambda_l \sum_{k=1}^{K} |\theta_k|$$

2. Shooting, LARSなどの最適化アルゴリズムで推定

λ_lの決め方

1. $\boldsymbol{y}^* = [\boldsymbol{y}_1^{*T}, \cdots, \boldsymbol{y}_P^{*T}]^T, B^* = [B_1^{*T}, \cdots, B_P^{*T}]^T$
2. k-fold cross validation (CV)を定義する

$$\text{CV}(\lambda_l) = \sum_{p=1}^{P} \frac{1}{N_p} \left\| \boldsymbol{y}_k^* - B^* \theta_{Lasso}^{(-p)} \right\|_2^2$$

3. CVを最小するλ_lを求める

20

5.LASSO型の選択法

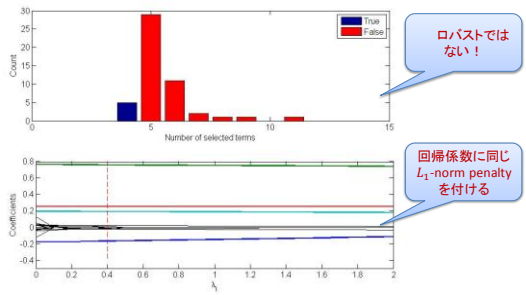


図11、重みつきLASSO推定の数値結果

21

5.LASSO型の選択法

Adaptive LASSO推定[Zou (2006)]

$$\hat{\boldsymbol{\theta}}_{\text{LASSO}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} (\boldsymbol{y} - B\boldsymbol{\theta})^T W (\boldsymbol{y} - B\boldsymbol{\theta}) + \lambda_l \sum_{k=1}^{K} v_k |\theta_k|$$

adaptive LASSO推定の計算

1. $\boldsymbol{\theta}^{\#} = [v_1\theta_1, v_2\theta_2, \cdots, v_K\theta_K]^T, \Phi^{\#} = \left[\frac{\Phi_1^*}{v_1}, \frac{\Phi_2^*}{v_2}, \cdots, \frac{\Phi_K^*}{v_K}\right]$
2. $\hat{\boldsymbol{\theta}}_{\text{LASSO}}^{\#} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} (\boldsymbol{y}^{\#} - B^*\boldsymbol{\theta})^T (\boldsymbol{y}^{\#} - B^{\#}\boldsymbol{\theta}) + \lambda_l \sum_{k=1}^{K} |\theta_k^{\#}|$
3. $\theta_{adaLSSO}^{\#} = \left[\frac{\theta_1^{\#}}{v_1}, \frac{\theta_2^{\#}}{v_2}, \cdots, \frac{\theta_K^{\#}}{v_K}\right]^T$

21

5.LASSO型の選択法

適応重みv_kのつけ方

1. Ridge regression推定を計算

$$\hat{\boldsymbol{\theta}}_{\text{ridge}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} (\boldsymbol{y} - B\boldsymbol{\theta})^T W (\boldsymbol{y} - B\boldsymbol{\theta}) + \lambda_r \sum_{k=1}^{K} \theta_k^2$$

2. $\hat{\boldsymbol{\theta}}_{\text{ridge}}$の$k$番目の要素$\frac{1}{|\hat{\theta}_k^{\text{ridge}}|}$を$v_k$とする

正規化パラメータの決め方

$\text{CV}(\lambda_l, \lambda_r) = \sum_{p=1}^{P} \frac{1}{N_p} \left\| \boldsymbol{y}_k^* - B^* \theta_{adaLasso}^{(-p)} \right\|_2^2$を最小する$\lambda_l, \lambda_r$

22

5.LASSO型の選択法

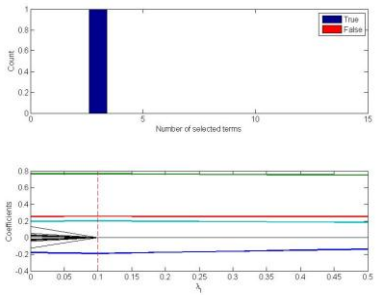


図11、重みつきAdaptive LASSO推定の数値結果

23

6.数値例

- データは多項式非線形システムによる生成する

$$y(t) = 0.5y(t-2) - 0.1y(t-1)u(t-3) - 0.3y(t-3)u(t-1) + 0.2u(t-2)^2 + e(t)$$

- $u(t) \sim N(0,1), e(t) \sim N(0,0.005)$
- 教師データ$\mathrm{D}_{\mathrm{Train}}$を700、テストデータ$\mathrm{D}_{\mathrm{Test}}$を700生成する
- 5%の教師データをランダムに欠損させ、その部分を$N(0,0.005)$によって生成されたノイズで替える
- 推定法：WLSとLS
- $w_t = \begin{cases} 10 & \text{if } |y(t)| > 0.2 \\ 1 & \text{else} \end{cases}$
- $\{y(t-1), \cdots, y(t-n), u(t-1), \cdots, u(t-n)\}$による2次までの多項式を用いて、基底関数を構成する。ここで、$n = 3,4,5,6$とする。したがって、基底関数の数が $K = 27,44,65,90$
- 選択手法：GICに基づくロバスト選択法、LASSO、adpative LASSO

24

6.数値例

ロバスト選択法の結果

$u(t)^2, y(t-2), y(t-3)u(t-2), y(t-1)u(t-3)$

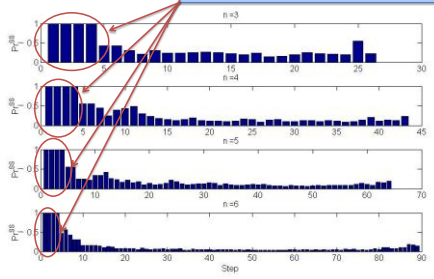


図12、各ステップで基底関数が選択される確率の最大値

25

6.数値例

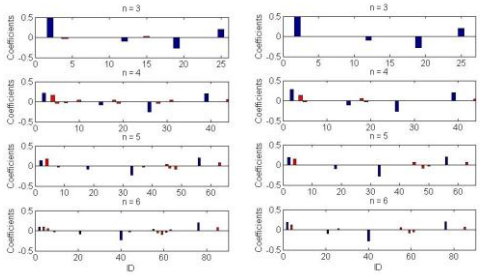


図13、LASSOとadaptive LASSOの選択結果

27

6.数値例

予測精度

- 予測精度がPMSE(predictive mean squared error)によって評価する

$$\mathrm{PMSE} = \frac{1}{N_{\mathrm{Test}}} \sum_{y(t) \in D_{\mathrm{Test}}} \left(y(t) - \hat{y}(t)\right)^2$$

表1、選択されたモデルの精度

n	Robust	LASSO	Adaptive LASSO	LS
3	**0.0073**	0.0075	**0.0073**	0.1818
4	**0.0073**	0.0085	0.0082	0.1818
5	**0.0073**	0.0078	0.0094	0.1818
6	**0.0073**	0.0093	0.0082	0.1818

28

6.数値例

選択法のロバスト性

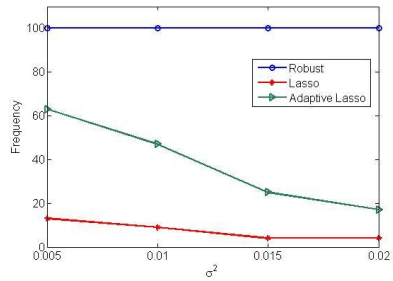


図14、三つの選択法のモンテカルロ実験

29

7、まとめ

- WLSで推定されたNARXモデルに対して、GICに基づくロバスト選択法、LASSO型の縮小選択法を提案した。
- GICに基づくロバスト選択法が予想のように、統一及び簡潔なモデルを選択できる。しかし、Subsamplingを用いるため、計算は複雑である。
- GICに基づくロバスト選択法に比べれば、LASSO型の選択法がモデルの推定と選択を同時に行える。
- Adaptive LASSOは確実にLASSOの選択パフォーマンスを改善できる。
- ノイズの分散を大きくすると、（すなわち、SN比(signal-noise-ratio)を小さくすると）、LASSO型のロバスト性が落ちる。それに比べれば、GICに基づくロバスト選択法がロバスト性を保てる。

30

参考文献

1. Konishi, S., Kitakawa, G.: 'Generalised information criteria in modle selection' Biometrika, **83**(4), 875-890 (1996)
2. Garatti, S., Bitead, R.R.: 'On resampling and uncertainty estimation in linear system identification', Automatica, **46**(55), 785-795 (2010)
3. Tibshirani, R.: 'Regression Shrinkage and Selection via the Lasso', J. Roy. Stat. Soc. B., **58**(1), 267-288 (1996)
4. Zou, H.: 'The adaptive Lasso and its oracle properties', J. AM. Stat. Assoc., **101**(476), 1418-1429 (2006)

30