

逐次的動的モデル選択の線形時間アルゴリズム

櫻井 瑛一, 山西 健司

東京大学大学院情報理工学系研究科

2010年6月15日

1 背景

- 動的モデル選択とは
- 既存の一括型 DMS

2 提案手法

- 既存手法の問題点
- 提案手法:逐次型 DMS
- 理論上界

3 数値実験

- 実験結果:計算時間
- 実験結果:一致率
- 実験結果:遅れ評価

4 まとめ

1 背景

- 動的モデル選択とは
- 既存の一括型 DMS

2 提案手法

- 既存手法の問題点
- 提案手法:逐次型 DMS
- 理論上界

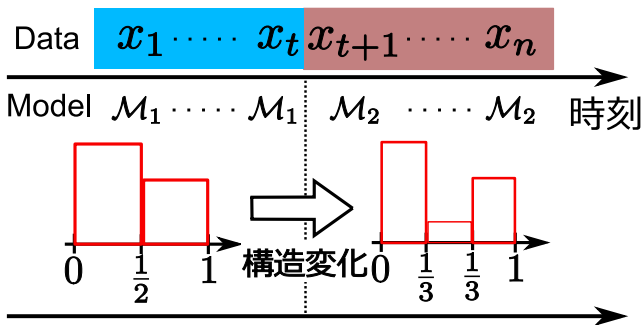
3 数値実験

- 実験結果:計算時間
- 実験結果:一致率
- 実験結果:遅れ評価

4 まとめ

動的モデル選択問題とは

設定: 非定常な状況から取得された時系列データ x_1, \dots, x_n



目標: モデルの変化を捉えることを考える

$\Leftrightarrow x_1, \dots, x_n$ を説明する **モデルの系列** を求める

\Rightarrow **本発表では特に逐次的な構造変化検出を目的**

\Leftrightarrow 従来の情報量規準 (ex. MDL) では **単一** のモデルを選択

関連研究

静的なモデル選択:

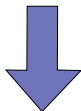
定常性を仮定

データに対し最適なモデルを 1 つ選択

- AIC (Akaike 1973)
- BIC (Schwarz 1978)
- MDL (Rissanen 1978)

動的な状況での予測 (モデル選択なし):

- Tracking Best Experts (Herbster & Warmuth 1998)
- Derandomizing Stochastic Prediction Strategies (V. Vovk 1999)
Best Expert が時間とともに変わる場合の逐次的予測アルゴリズム
- Switching 分布 (Erven, Grunwald, Rooij, 2008)



動的なモデル選択:

- 情報理論に基づく動的モデル選択 (Yamanishi & Maruyama 2007)
← 今回注目した研究

DMS(Dynamic Model Selection)での 問題設定

K 個のパラメトリックな確率モデル: $C_k = \{P_k(\cdot|\theta_k)\} (k = 1, \dots, K)$
(モデル次数が増大) $\dim(\theta_1) < \dim(\theta_2) < \dots < \dim(\theta_K)$

逐次的なパラメータ推定: $\hat{\theta}_k(x^{t-1}) = \hat{\theta}_k^{(t)}$

(最尤推定, **Bayes** 推定, 忘却推定等)

としたときにモデル選択規準により以下を選択
従来のモデル選択



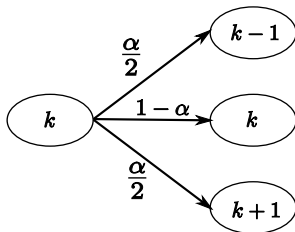
動的モデル選択



DMSでのモデル系列の評価規準

情報論的立場から，モデル系列を評価
→MDL(Minimum Description Length) 原理
に従い記述長最小となるモデル系列を選択

ただし，モデル遷移は隣接するモデルに
マルコフ的に確率遷移としてモデル系列
の記述長を定義



Yamanishi and Maruyama の DMS 規準

$$\ell(x^n, k^n) = \ell(x^n | k^n) + \ell(k^n) = \underbrace{\sum_{t=1}^n -\log P(x_t | \hat{\theta}_{k_t}^{(t)})}_{\text{データの記述長}} + \underbrace{\sum_{t=1}^n -\log P(k_t | k_{t-1} : \hat{\alpha}_{t-1})}_{\text{モデル系列の記述長}}$$

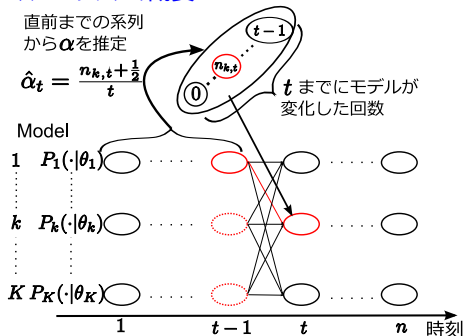
予測的記述長として計算

既存手法:一括型DMS

目標

モデルの遷移確率は一定だが未知，データが一括に与えられた状況での最適モデル系列を求める

アルゴリズム概要



ポイント

- Krichevasky & Trofimov 推定で遷移確率 α を推定 (Krichevasky & Trofimov 1981)

$$\hat{\alpha}_t = \frac{n_{k,t} + \frac{1}{2}}{t}$$

- 動的計画法を用いて DMS 規準を最小とする最適モデル系列を推定

全体計算量:

時刻 t において α の推定のため計算すべき系列の状態が t 個

∴ 全体計算量は $O(K \sum_{t=1}^n t) = O(Kn^2)$

目次

1 背景

- 動的モデル選択とは
- 既存の一括型 DMS

2 提案手法

- 既存手法の問題点
- 提案手法:逐次型 DMS
- 理論上界

3 数値実験

- 実験結果:計算時間
- 実験結果:一致率
- 実験結果:遅れ評価

4 まとめ

既存手法の問題点

問題点

- 逐次的な推定が出来ない
∴ 動的計画法で計算する際に計算すべき量が増え、
逐次的な推定が不能
- 計算量がデータ量の 2 乗かかる



目標

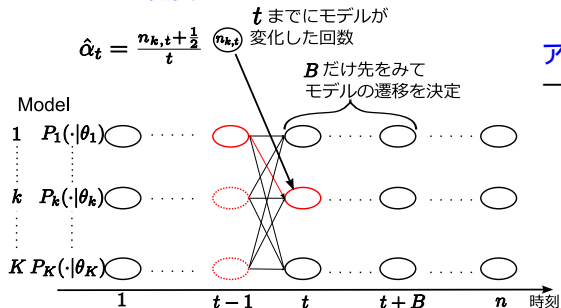
- **線形でかつ逐次的**なモデル系列の推定が可能なアルゴリズムへ
- 一括型 **DMS** で求まる最適な系列からあまり外れないように

提案手法:逐次型 DMS

アルゴリズムを構成する上での着眼点

- 逐次的に幅 B のスライディングウィンドウでモデル系列を順々に接続
- 接続していくモデル系列の端を $1, \dots, K$ の K 種類保持

アルゴリズム概要



アルゴリズム上のポイント 一括型 DMS との相違点

- α の推定が単一に
- B だけ先読みする

全体計算量:

各時刻 t で計算すべき量が各モデルで 1 個

B だけ先読みをするコストは DP を用いると $O(B^2)$

\therefore 計算量は $O(KB^2n)$

DMS 規準の上界評価

- モデル系列の変化点数 m のみにモデル記述長上界が依存
- 上界での \min の制約はモデル系列 k^n にデータからの制約を入れた形
- 一括型 DMS では \min の範囲が**全ての**モデル系列 k^n

逐次型 DMS での DMS 規準の上界評価

$$\ell(x^n, k^n) \leq \min_{k^n \in D(B, n)} \left\{ \sum_{t=0}^n -\log P_{k_t}(x_t | \hat{\theta}_{k_t}^{t-1}) + nH\left(\frac{m}{n}\right) + \frac{1}{2} \log n + m + o(\log n) \right\}$$

$D(B, n)$ は以下を満たすモデル系列の集合

$$D(B, n) = \{k^{n-B} \oplus k^B | k^{n-B} \in \hat{D}(B, n), k^B \in \mathcal{K}^B\}$$

$$\hat{D}(B, n) = \{k_{op}^{n-B-1}(k) \oplus k | k \in \mathcal{K}\}$$

$$\text{ただし } k_{op}^{n-B-1}(k) = \arg \min_{k^{n-B-1} \in \hat{D}(B, n-1)} \min_{k^B \in \mathcal{K}^B}$$

$$\left\{ \sum_{t=1}^n -\log_{k_t} P(x^n | \hat{\theta}_k^{(t)}) + l(k^{n-B-1} \oplus k \oplus k^B) \right\}$$

$$\mathcal{K} = \{1, \dots, K\}$$

目次

1 背景

- 動的モデル選択とは
- 既存の一括型 DMS

2 提案手法

- 既存手法の問題点
- 提案手法:逐次型 DMS
- 理論上界

3 数値実験

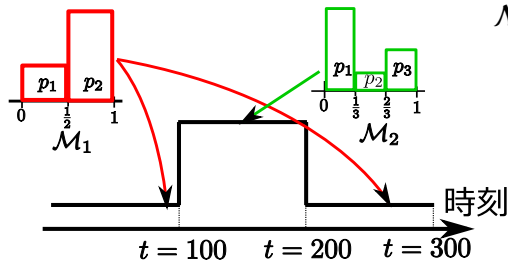
- 実験結果:計算時間
- 実験結果:一致率
- 実験結果:遅れ評価

4 まとめ

実験方法

目的

情報源が変化する人工データで逐次型 **DMS** が変化を検知できるかを実験
[0, 1] 区間上でヒストグラムの数が変化する人工データを使用
人工データ



\mathcal{M}_1 : 固定

\mathcal{M}_2 : KL-divergence を
変化させ実験

	D_{12}	D_{21}
\mathcal{M}_{21}	0.611	0.5007
\mathcal{M}_{22}	1.0192	0.7782
\mathcal{M}_{23}	1.1343	1.1594
\mathcal{M}_{24}	1.4115	1.5165

ただし $D_{ij} = D(\mathcal{M}_i || \mathcal{M}_j)$

比較対称

既存手法の一括型 **DMS** と逐次型 **DMS** を比較
逐次型 **DMS** は先を見る幅 B を 0, 1, 5 と変化

実験結果: 計算時間

実験に用いた機械: C2D 1.86GHz, Memory 2G

使用した言語: Mathematica

平均計算時間 (秒)

一括型 DMS	逐次型 DMS		
	$B = 0$	$B = 1$	$B = 5$
38.73	0.25	3.70	29.03



逐次型 DMS は計算時間が非常に少なくなっている

実験結果:一致率

一致率: 一括型 DMS と逐次型 DMS で求めたモデル系列の一致した割合

	逐次型 DMS		
	$B = 0$	$B = 1$	$B = 5$
\mathcal{M}_{21}	93/100	93/100	93/100
\mathcal{M}_{22}	100/100	100/100	100/100
\mathcal{M}_{23}	96/100	97/100	98/100
\mathcal{M}_{24}	91/100	91/100	91/100

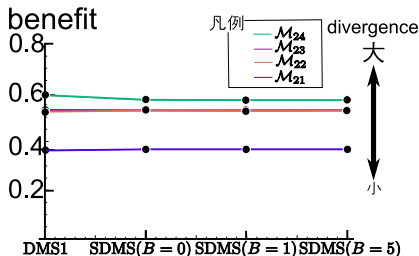
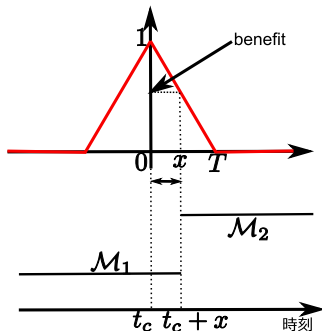


9 割以上の場合で一括型 DMS と同じ系列を求めている

実験結果:遅れ評価

benefit: 求まったモデル系列での変化点の
データ生成での変化点 $t = 100, 200$
からの外れ度合い
($T = 10$ を使用)

$$\text{benefit}(x, T) = \max \left\{ 0, 1 - \frac{|x|}{T} \right\}$$



- benefit は一括型 DMS とほぼ同じ
- benefit の大小は divergence の大小に依存

目次

1 背景

- 動的モデル選択とは
- 既存の一括型 DMS

2 提案手法

- 既存手法の問題点
- 提案手法:逐次型 DMS
- 理論上界

3 数値実験

- 実験結果:計算時間
- 実験結果:一致率
- 実験結果:遅れ評価

4 まとめ

まとめ

- 動的モデル選択の問題枠組みで逐次的な入力に対応でき、かつ線形なアルゴリズムを提案
- 逐次型 **DMS** は **90%**を超える割合で一括型 **DMS** と同様のモデル系列が得られ、正確さ、**benefit** も大幅に悪化することはない

補足:

人工データのパラメータ

$\mathcal{M}_1 = \{0.2, 0.8\}$:固定

$\mathcal{M}_2 = \{p_1, p_2, p_3\}$:以下のように変化

	\mathcal{M}_2 のパラメータ	D_{12}	D_{21}
\mathcal{M}_{21}	{0.5, 0.05, 0.45}	0.611	0.5007
\mathcal{M}_{22}	{0.15, 0.8, 0.05}	1.0192	0.7782
\mathcal{M}_{23}	{0.8, 0.05, 0.15}	1.1343	1.1594
\mathcal{M}_{24}	{0.9, 0.05, 0.05}	1.4115	1.5165

