情報論的学習理論テクニカルレポート 2009
Technical Report on Information-Based Induction Sciences 2009 (IBIS2009)

線形時間異種混合モデル選択のための期待情報量基準最小化法

Linear Time Model Selection for Mixture of Heterogeneous Components via Expectation Minimization of Information Criteria

藤巻遼平* 森永聡 門馬道也 青木健児 中田貴之
Ryohei Fujimaki Satoshi Morinaga Michinari Momma Kenji Aoki Takayuki Nakata

Abstract: Our main contribution is to propose a novel model selection methodology, expectation minimization of information criterion (EMIC). EMIC makes a significant impact on the combinatorial scalability issue pertaining to the model selection for mixture models having types of components. A goal of such problems is to optimize types of components as well as the number of components. One key idea in EMIC is to iterate calculations of the posterior of latent variables and minimization of expected value of information criterion of both observed data and latent variables. This enables EMIC to compute the optimal model in linear time with respect to both the number of components and the number of available types of components despite the fact that the number of model candidates exponentially increases with the numbers. We prove that EMIC is compliant with some information criteria and enjoys their statistical benefits.

Keywords: Mixture of Heterogeneous Components, Expectation Minimization of Information Criteria

1 はじめに

複雑多様なデータ分析へのニーズの高まりと共に，異種多様な分布を混合したモデルクラス (我々は、このモデルを異種混合モデルと呼ぶ) の重要性が高まっている．異種混合モデルのクラスには，正規分布と指数分布の混合分布のような比較的単純なモデルから，異なる内部次元を持つ混合因子分析モデル [6]，異なる次数を持つ混合多項曲線モデル，異なる属性を利用する混合判別モデルといった複雑なモデルまで含まれ，このような複雑な分布は実応用の様々な場面で遭遇する．

異種混合モデル (の一部) を推定するための素朴な方法としては、混合正規分布によって近似をする事が挙げられる。これは、混合正規分布がコンポーネント数の増加と共に任意の精度で分布を近似可能であるという点で根拠を持つ [14]。一方で、異種混合モデルを精度よく近似するために多数のコンポーネントを利用した場合にはモデルの解釈が困難となり、解釈性を高めるためにコンポーネント数を少なくすると近似が粗くなるという問題がある。

本稿では，異種のコンポーネントを混合させて，異種混合モデルを「真っ向から」学習するための，一般的な枠組みを確立する事を目的とする．異種のコンポーネントを混合させるためには，事前に適切なコンポーネント数と各コンポーネントの種類を決める必要がある．しかしながら，実応用においてこれらを事前に知る事は不可能であり，データを人手で分析してコンポーネントを一つずつ設定する事も非現実的である．したがって，異種混合モデルに関する学習問題では，データからコンポーネントの数，種類およびそのパラメータを推定する事が重要となる．我々は，この問題を “異種混合モデル選択”，可能なコンポーネントの種類を “コンポーネント候補” と呼ぶ事にする．

異種混合モデル選択問題の重要な課題として，

(A) 任意のモデルをコンポーネント候補として扱う，

(B) コンポーネント数およびコンポーネント候補数に対して線形の計算量を保証する，

という 2 点が挙げられる．(A) は，このクラスの問題を

*NEC 共通基盤ソフトウェア研究所，211-8666 神奈川県川崎市中原区下沼部 1753，e-mail r-fujimaki@bx.jp.nec.com，URL http://www.nec.co.jp/rd/datamining/
NEC Common Platform Software Research Laboratoies, 1753, Shimonumabe, Nakahara-ku, Kawasaki-shi, Kanagawa

一般に広く解決するために必要な課題である．また，モデルの候補数はコンポーネント候補の組み合わせ数によって決まり膨大となるため，(B) は異種混合モデル選択をコンポーネント候補の組み合わせ数に依存せずに，現実的な時間で解決するために必要となる課題である．

(A) を解決するための一般的な方法として，情報量基準 (最小記述長 (MDL) [10]，赤池情報量基準 (AIC) [1]，ベイズ情報量基準 (BIC) [13] など) を目的関数としてモデル空間を探索する方法が挙げられる．しかし，モデルの候補数は，コンポーネント候補の組み合わせの数に依存し、全探索はもちろんパタン探索 [9] などの探索アルゴリズムを利用したとしても (B) の問題を解決する困難である．より洗練された方法として，変分ベイズ法に基づく手法 [6] や階層的クラスタリングタイプの手法 [8] が提案されているが，双方ともコンポーネントの分布クラスに制限があり，線形の計算量を保証する事はできていない．

本稿は，異種混合モデル選択に対して，期待情報量基準最小化法 (Expectation Minimization of Information Criterion; EMIC) を提案する．EMIC は，隠れ変数を含むモデルのパラメータ推定として一般的な EM アルゴリズム [5] をモデル選択へ拡張した方法で，隠れ変数の事後分布に関する完全データ[1]の情報量基準の期待値を繰り返し最小化する．通常の EM アルゴリズムとの最大の相違点は，M ステップにおいて，各コンポーネントのパラメータだけでなく，各コンポーネントの種類そのものを更新可能な事である．これによって，コンポーネント数とコンポーネント候補数に対して線形時間の異種混合モデル選択が実現される (すなわち課題 (B) が解決される)．また，第 3.4 節で，MDL/BIC, AIC に対して，期待情報量基準の繰り返し最小化が，元の情報量基準の値が単調に減少させることを示す．これは，EMIC による期待情報量基準の探索が，元の情報量基準に対する最適解を探索可能である事を意味している．MDL/BIC, AIC などの情報量基準は，任意の分布をコンポーネント候補として扱う事が可能なため，これによって EMIC が課題 (A) を解決可能である事を示す．

本稿は次節以降，以下のように構成される．まず，第 2 節で異種混合モデルの定義を与え，提案手法の土台となる情報量基準に関する説明を行なう．第 3 節で，提案手法の定式化，最適化，理論的正当性などを説明し，第 4 節で人工データおよび UCI データ [2] を利用して，提案手法が (A) および (B) を解決可能である事を実証する．

[1]観測データとそれに対応する隠れ変数.

2 数学的準備

$|\circ|$ は，集合 $\circ$ の要素数，$P(\bullet;\star)$ は $\star$ によってパラメトライズされた $\bullet$ の確率密度 (質量) 関数，$\hat{\star}$ は $\star$ の最尤推定量を表すとする..

2.1 異種混合モデル

パラメトリックな確率分布族 $V_j = \{P(X;\boldsymbol{\phi}^{V_j})|\boldsymbol{\phi}^{V_j} \in \Phi_j\}$ を考える．ただし，$\boldsymbol{\phi}^{V_j} = (\phi_1^{V_j}, \ldots, \phi_{J_{V_j}}^{V_j})$ は V_j に対応するパラメータ集合 Φ_j の要素とする．ここで，$P(X;\boldsymbol{\phi}^{V_j})$ は j に対して異なるパラメータ次元 J_{V_j} を持つだけでなく分布系が異なってもよい．また，コンポーネント候補を，$\mathcal{S} = \{V_j|j = 1, \ldots, |\mathcal{S}|\}$ と定義する．

本稿では，異種混合モデルを $\mathcal{S}$ に対する混合分布，

$$\left\{P(X;\boldsymbol{\theta}) = \sum_{c=1}^{C} \pi_c P(X;\boldsymbol{\phi}_c^{S_c})\right\}, \tag{1}$$

と定義し[2]，$\mathcal{H} = \{H_i|i = 1, \ldots, |\mathcal{H}|\}$ と表記する[3]．また，$\mathcal{H}$ をモデル候補と呼ぶ．(1) では，C はコンポーネント数，$S_c \in \mathcal{S}$ $(c = 1, \ldots, C)$ はコンポーネント c の種類を表し，$\mathcal{H}$ の要素ごとに異なる．またパラメータを，$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_C)$，$\boldsymbol{\phi} = (\boldsymbol{\phi}_1^{S_1}, \ldots, \boldsymbol{\phi}_C^{S_C})$，$\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\phi}\}$ と表記する．

2.2 情報量基準

一般に，情報量基準は観測データ $\boldsymbol{x}^N = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ および $H \in \mathcal{H}$ に対して，(負の) 対数尤度[4]とモデルの複雑性のトレードオフとして，

$$IC(\boldsymbol{x}^N; H) = -\sum_{n=1}^{N} -\log P(\boldsymbol{x}_n; \hat{\boldsymbol{\theta}}^H) + \ell(\hat{\boldsymbol{\theta}}^H), \tag{2}$$

と表現される[5]．ただし，$\ell(\hat{\boldsymbol{\theta}}^H)$ はモデルの複雑性を表す．例えば，MDL/BIC[6]，AIC は，

$$MDL/BIC(\boldsymbol{x}^N; H) = \sum_{n=1}^{N} -\log P(\boldsymbol{x}_n; \hat{\boldsymbol{\theta}}^H) + \frac{J_H}{2} \log N,$$

[2]階層隠れ変数モデル [3] や階層混合エキスパートモデル [7] など，より複雑な混合分布に関しては今後の課題とする．

[3]混同の恐れがない場合には，表記の簡単化のため $\boldsymbol{\theta}^H$ の上付き文字 H は省略する．

[4]拡張型確率的コンプレキシティ [15] のように対数損失 (尤度) 以外の適合度を利用する基準，予測的確率的コンプレキシティ [11] や交差検定のように，データとモデルの複雑性が加算的な形の分解されていな基準など，より広いクラスのモデル適合度を利用する基準に対する議論は今後の課題とする．

[5]交差検定法によるモデル選択は，一般には異なる形式となるため本稿の議論は必ずしも当てはまらない．

[6]このタイプの MDL は crude MDL と呼ばれ，近似精度が $\mathcal{O}(\log N)$ となりうる事が知られている．最尤符号化を利用して近似精度が $\mathcal{O}(1)$ である refined MDL が Rissanen [12] によって提案されている．本稿の議論は，直接 refined MDL へも適用可能であるが，簡単のため crude MDL に対して議論を行なう．

$$AIC(\boldsymbol{x}^N;H) = 2\sum_{n=1}^{N} -\log P(\boldsymbol{x}_n;\hat{\boldsymbol{\theta}}^H) + 2J_H,$$

と表現される．一般的に，これらの情報量基準を利用してモデル選択を行なう場合，各モデル候補に対して EM アルゴリズム [5] によって最尤推定量 $\hat{\boldsymbol{\theta}}^H$ を計算し，次に情報量基準を最小化するモデルを選択する．しかし，異種混合モデル選択では，コンポーネント数とコンポーネント候補数と共に，モデル候補の数 $|\mathcal{H}|$ が膨大となる点に問題がある．

3 期待情報量基準最小化

3.1 完全データの期待情報量基準

まず，X に対する隠れ変数を $Z = (Z_1, \ldots, Z_C)$ を考える．Z_c は X が c 番目のコンポーネントから生成された場合に 1 ，それ以外で 0 をとる．また，$\boldsymbol{x}_n$ に対応する隠れ変数の値を $\boldsymbol{z}^N = \boldsymbol{z}_1, \ldots, \boldsymbol{z}_N$ とする．一般に，(X, Z) の組は完全変数，$(\boldsymbol{x}^N, \boldsymbol{z}^N)$ は完全データと呼ばれる．

完全変数に関する同時分布 $P(X, Z|\boldsymbol{\phi})$ は，$P(X|Z, \boldsymbol{\phi})$ および $P(Z|\boldsymbol{\pi})$ に分解する事が可能である．ここで，$P(X|Z, \boldsymbol{\phi})$ は隠れ変数 Z が条件として与えられているため，$P(X|Z_c = 1, \boldsymbol{\phi}) = P(X; \phi_c^{S_c})$ である．したがって，完全データの情報量基準 $IC(\boldsymbol{x}^N, \boldsymbol{z}^N; H)$ は，

$$\begin{aligned} IC(\boldsymbol{x}^N, \boldsymbol{z}^N; H) = &-\sum_{n=1}^{N}\sum_{c=1}^{C} z_{nc} \log P(\boldsymbol{x}_n; \hat{\boldsymbol{\phi}}_c^{S_c}) \\ &- \sum_{n=1}^{N}\sum_{c=1}^{C} z_{nc} \log \hat{\pi}_c + \sum_{c=1}^{C} \ell(\hat{\boldsymbol{\phi}}_c^{S_c}) + \ell(\hat{\boldsymbol{\pi}}) \end{aligned} \tag{3}$$

と表す事が可能である．右辺の $\sum_{c=1}^{C} \ell(\hat{\boldsymbol{\phi}}_c^{S_c}) + \ell(\hat{\boldsymbol{\pi}})$ は，完全変数に対する同時分布のモデル複雑性であり，MDL/BIC および AIC に対しては，

$$\begin{cases} \sum_{c=1}^{C} \frac{J_{S_c}}{2} \log\left(\sum_{n=1}^{N} z_{nc}\right) + \frac{C-1}{2} \log N & \text{for MDL/BIC} \\ \sum_{c=1}^{C} J_{S_c} + (C-1) & \text{for AIC} \end{cases} \tag{4}$$

と計算される．

(2) に対する (3) の重要な利点は，目的関数が各コンポーネントに対して分離されているため，各コンポーネントの種類やパラメータを独立に最適化可能な点にある．これによって，コンポーネント候補数およびコンポーネント数に対して線形の計算オーダーで，異種混合モデル選択を実行する事が可能である．

実際には，隠れ変数の値は観測する事ができないため，$IC(\boldsymbol{x}^N, \boldsymbol{z}^N; H)$ を計算する事ができない．そこで，EM アルゴリズムと同様に，隠れ変数の事後分布，

$$P(Z|\boldsymbol{x}, \boldsymbol{\theta}, H) \propto \prod_{n=1}^{N}\prod_{c=1}^{C}\left(\pi_c P(\boldsymbol{x}; \boldsymbol{\phi}_c^{S_c})\right)^{Z_c}. \tag{5}$$

に対する，$IC(\boldsymbol{x}^N, \boldsymbol{z}^N; H)$ の期待値 $E_Z[IC(\boldsymbol{x}^N, \boldsymbol{z}^N; H)]$ を最適化することにする．完全変数の期待情報量基準を目的関数とする事の正当性は，第 3.4 節で説明する．

$E_Z[\bullet]$ は，$\bullet$ の $P(Z|\boldsymbol{x}, \boldsymbol{\theta}, H)$ に関する期待値を表すとすると，$E_Z[IC(\boldsymbol{x}^N, \boldsymbol{z}^N; H)]$ は，

$$\begin{aligned} E_Z[IC(\boldsymbol{x}^N, \boldsymbol{z}^N; H)] = &-\sum_{n=1}^{N}\sum_{c=1}^{C} E_Z[z_{nc}] \log P(\boldsymbol{x}_n; \hat{\boldsymbol{\phi}}_c^{S_c}) \\ &- \sum_{n=1}^{N}\sum_{c=1}^{C} E_Z[z_{nc}] \log \hat{\pi}_c + E_Z\Big[\sum_{c=1}^{C} \ell(\hat{\boldsymbol{\phi}}_c^{S_c}) + \ell(\hat{\boldsymbol{\pi}})\Big] \end{aligned} \tag{6}$$

と計算される．(6) と通常の EM アルゴリズムの期待対数尤度との違いは，モデルの複雑性の隠れ変数の事後分布に対する期待値を含んでいる点で，これによって最適化の各ステップでコンポーネントの種類を最適化する事が可能となる．

3.2 最適化アルゴリズム

EMIC は，各コンポーネント数 C に対して，以下で説明する EM ステップを繰り返す事でコンポーネントの種類 S_c およびパラメータ $\boldsymbol{\phi}_c^{S_c}$ を最適化し，最適なコンポーネント数 C^* を情報量基準によって選択する．以下では，(t) は EM ステップの t 回目の繰り返しである事を表すとし，混乱の恐れがない場合には省略する場合がある．

3.2.1 E ステップ

E ステップでは，隠れ変数の事後分布 $P(Z|\boldsymbol{x}, \boldsymbol{\theta}, H)$ を計算する．具体的には，M ステップで必要となる $E_Z[z_{nc}]$ および $E_Z[\sum_{c=1}^{C} \ell(\hat{\boldsymbol{\phi}}_c^{S_c}) + \ell(\hat{\boldsymbol{\pi}})]$ の期待値計算を行なう．

まず，$E_Z[z_{nc}]$ に関しては EM アルゴリズムと同様に，

$$E_Z^{(t)}[z_{nc}] = \frac{\pi_c^{(t-1)} P(\boldsymbol{x}_n; \boldsymbol{\phi}_c^{S_c^{(t-1)}})}{\sum_{c=1}^{C} \pi_c^{(t-1)} P(\boldsymbol{x}_n; \boldsymbol{\phi}_c^{S_c^{(t-1)}})}. \tag{7}$$

と計算される．$\boldsymbol{z}_n$ はお互いに独立であるため，この処理にかかる計算量は $\mathcal{O}(N)$ である．

次に，$E_Z[\sum_{c=1}^{C} \ell(\hat{\boldsymbol{\phi}}_c^{S_c}) + \ell(\hat{\boldsymbol{\pi}})]$ に関しては，基準によって具体的な計算方法が異なるため，ここでは MDL/BIC および AIC の場合を説明する．MDL/BIC に関しては，$E_Z^{(t)}[\log(\sum_{n=1}^{N} z_{nc})]$ を直接計算するためには，$z_{1c}, \ldots, z_{Nc}$ に関する 2^N の組み合わせを考える必要があるため，指数時間の計算量が必要となる．これでは第 1 節で述べた

(B) の課題を解決できたとしても，高速な異種混合モデル選択が実現できない事となる．そこで，第 3.3 節で N に対する線形オーダーかつ任意精度で近似が可能なアルゴリズムを説明する．AIC に関しては，モデルの複雑性 (4) が隠れ変数に依存しないため，特別な期待値の計算は必要ない．

3.2.2 M ステップ

M ステップでは，$E_Z[IC(\boldsymbol{x}^N, \boldsymbol{z}^N; H)]$ を最小化するコンポーネントの種類およびパラメータを推定する．EMIC では，完全変数に対する情報量基準を考える事で，目的関数がコンポーネント毎に分解されているため，

$$\underset{S_c, \boldsymbol{\phi}_c^{S_c}}{\operatorname{argmin}} \Big\{ -\sum_{n=1}^{N} E_Z^{(t)}[z_{nc}] \log P(\boldsymbol{x}_n; \boldsymbol{\phi}_c^{S_c}) + \ell(\hat{\phi}_c^{S_c}) \Big\}, \tag{8}$$

なる最適化問題を解くことによって，各コンポーネントごとに S_c および $\boldsymbol{\phi}_c^{S_c}$ を最適化することが可能である．また，混合比に関しては，

$$\hat{\pi}_c^{(t)} = \frac{\sum_{n=1}^{N} E_Z^{(t)}[z_{nc}]}{N}, \tag{9}$$

と推定される．

EMIC のモデル選択の鍵は，各 EM ステップにおいて，コンポーネントごとに S_c および $\boldsymbol{\phi}_c^{S_c}$ の推定が可能な点である．これによって，コンポーネントの種類に関する全ての組み合わせに関して情報量基準を計算する事を回避し，コンポーネント候補数の線形オーダーでの異種混合モデル選択が可能となっている．

EM ステップの収束は，元の情報量基準 $IC(\boldsymbol{x}^N; H^{(t-1)}) - IC(\boldsymbol{x}^N; H^{(t)})$ に対して判定を行う．通常の EM アルゴリズムでは，しばしばパラメータの値 $\hat{\boldsymbol{\theta}}^{(t)}$ に対して収束判定が行われるが，EMIC では各ステップで $\hat{\boldsymbol{\theta}}^{(t)}$ の次元が変動するため $\hat{\boldsymbol{\theta}}^{(t)}$ に対して収束判定は難しい点で注意が必要である．

Algorithm 1 に EMIC の疑似コードを示す．各コンポーネント数 C に対して，コンポーネント種類とパラメータを EM ステップによって最適化する．最適なコンポーネント数 $\overline{IC}$ は $IC(\boldsymbol{x}^N; H)$ の最小化によって選択される．Algorithm 1 では，コンポーネント数を無限大まで探索する事ができないため，入力として最大のコンポーネント数 $C_{\max}$ が必要となる．この問題は，Ghahramani ら [6] の手法のようにコンポーネントの分割併合といった経験則を利用することで解決する事が可能と考えられる．

Algorithm 1 Expectation Minimization of Information Criterion

Input: $\boldsymbol{x}^N$, $\mathcal{S}$ and $C_{\max}$
Initialization : $\bar{H} \leftarrow$ NULL and $\overline{IC} \leftarrow \infty$
for $C = 1, \ldots, C_{\max}$ **do**
  $t \leftarrow 1$ and initialize $H^{(t)}$ and $\hat{\boldsymbol{\theta}}^{(t)}$.
  repeat
    $t \leftarrow t + 1$.
    Evaluate $E_Z[z_{nc}]$ and $E_Z[\sum_{c=1}^{C} \ell(\hat{\boldsymbol{\phi}}_c^{S_c}) + \ell(\hat{\boldsymbol{\pi}})]$.
    for $c = 1, \ldots, C$ **do**
      Calculate $S_c^{(t)}$, $\hat{\boldsymbol{\phi}}_{S_c}^{(t)}$, $\hat{\pi}_c^{(t)}$ by (8) and (9).
    end for
    Update: $H^{(t)} \leftarrow \{C, S_1^{(t)}, \ldots, S_C^{(t)}\}$ and $\hat{\boldsymbol{\theta}}^{(t)} \leftarrow \{\hat{\boldsymbol{\pi}}^{(t)}, \hat{\boldsymbol{\phi}}^{(t)}\}$.
    Evaluate $IC(\boldsymbol{x}^N; H^{(t)})$ and convergence.
  until $IC(\boldsymbol{x}^N; H^{(t)})$ converges
  if $IC(\boldsymbol{x}^N; H^{(t)}) < \overline{IC}$ **then**
    $\overline{IC} \leftarrow IC(\boldsymbol{x}^N; H^{(t)})$, $\bar{H} \leftarrow H^{(t)}$ and $\bar{\boldsymbol{\theta}} \leftarrow \hat{\boldsymbol{\theta}}^{(t)}$.
  end if
end for
Output: the optimal model $\bar{H}$ and parameter $\bar{\boldsymbol{\theta}}$

3.3 MDL/BIC に関する高速近似アルゴリズム

第 3.2.1 節で説明したように，MDL/BIC の E ステップにおいて，素朴に $E_Z^{(t)}[\log(\sum_{n=1}^{N} z_{nc})]$ を計算しようとすると $\mathcal{O}(2^N)$ の計算量が必要となる．これは，場合によってはコンポーネント候補の組み合わせを考慮する以上の計算量が必要となる．そこで，本節では，任意の近似精度かつ N の線形オーダーの計算量で $E_Z^{(t)}[\log(\sum_{n=1}^{N} z_{nc})]$ を近似する方法を説明する．

まず，$\mu_{nc} = E_Z[z_{nc}]$ および $\mu_c = \sum_{n=1}^{N} \mu_{nc}$ とする．$E_Z^{(t)}[\log(\sum_{n=1}^{N} z_{nc})]$ を $\sum_{n=1}^{N} z_{nc}$ に関して μ_c の周りでテイラー展開し，M 次より大きい項を無視すると，

$$E_Z[\log(\sum_{n=1}^{N} z_{nc})] \approx \sum_{m=0}^{M} b_m E_Z\Big[(\sum_{n=1}^{N} z_{nc})^m\Big], \tag{10}$$

を得る．ここで，b_m は適当な係数とする．M は近似の精度を制御する定数である．

(10) を高速に計算する鍵は，z_{nc} に関する $z_{ic}^k = z_{ic}$ $(k \geq 1)$ および $E_Z[z_{n_1 c} z_{n_2 c}] = \mu_{n_1 c} \mu_{n_2 c}$ $(n_1 \neq n_2)$ なる性質を利用する事である．この性質を利用すると，任意の m に対して，$E_Z[(\sum_{n=1}^{N} z_{nc})^m]$ が $\mathcal{O}(N)$ で計算できる．例として，$m = 2$ に関しては，

$$E_Z[(\sum_{n=1}^{N} z_{nc})^2] = \mu_c^2 + \mu_c - \sum_{i=1}^{N} \mu_{ic}^2 \tag{11}$$

と計算される．

3.4 EMIC の正当性

本節では，MDL および AIC に対して，完全データに対する期待情報量基準の最小化処理によって計算される最適解が，元の情報量基準に関する局所最適解となる事を証明する．BIC に関しては，漸近的には MDL と同じ形式となるため，MDL に関する証明によって正当性を示す．なお，各証明は紙面の都合上、骨子のみ記述する。

まず，MDL に関して以下の定理を与える．

定理 1 *EMIC* において *MDL* を情報量基準として利用した場合，データ数 N が大きい場合，第 *3.2* 節の *EM* ステップによって，$MDL(\boldsymbol{x}^N; H)$ が単調に減少する．

証明 1 *MDL* のモデル複雑性 $J_H/2 \log N$ は，パラメータ $\boldsymbol{\theta}^H$ を記述するために必要な符号長という意味を持つが，これはクラフトの不等式 *[4]* を満たし，パラメータ $\boldsymbol{\theta}^H$ に関する確率分布[7] *(*の対数尤度*)* と解釈する事が可能である．よって，$\boldsymbol{\theta}$ を確率変数とみなし $MDL(\boldsymbol{x}^N; H) = -\log P(\boldsymbol{x}^N, \hat{\boldsymbol{\theta}})$，$MDL(\boldsymbol{x}^N, \boldsymbol{z}^N; H) = -\log P(\boldsymbol{x}^N, \boldsymbol{z}^N, \hat{\boldsymbol{\theta}})$ と解釈する事が可能である．

したがって，

$$MDL(\boldsymbol{x}^N; H^{(t)}) = E_Z^{(t)}\Big[-\log \frac{P(\boldsymbol{x}^N, \boldsymbol{z}^N, \hat{\boldsymbol{\theta}}^{(t)})}{P(\boldsymbol{z}^N|\boldsymbol{x}^N, \hat{\boldsymbol{\theta}}^{(t)})}\Big]. \tag{12}$$

が成立し，$E_Z^{(t)}[MDL(\boldsymbol{x}^N, \boldsymbol{z}^N; H^{(t)})]$ の最適性および *Jensen* の不等式 *[4]* を考慮すると，

$$MDL(\boldsymbol{x}^N; H^{(t+1)}) \leq MDL(\boldsymbol{x}^N; H^{(t)}) \tag{13}$$

となる事がわかる．□

定理 1 では，MDL のモデル複雑性が漸近的にパラメータに関する確率値の対数と等価となる事を利用しているため，小さい N に対しては必ずしも成立しない．しかし，本稿での実験では単調性が常に成立していたため，実用上はある程度小さな N までは定理 1 が成立し，EMIC によって正しく MDL/BIC の最小化が実施できる．

次に，AIC に関して以下の定理を与える．

定理 2 *EMIC* において $\ell(\hat{\boldsymbol{\theta}}^H) = E_Z[\sum_{c=1}^{C} \ell(\hat{\boldsymbol{\phi}}_c^{S_c}) + \ell(\hat{\boldsymbol{\pi}})]$ を満たす情報量基準は，第 *3.2* 節の *EM* ステップによって，$IC(\boldsymbol{x}^N; H)$ が単調に減少する．

証明 2 まず、*EMIC* の最適化の性質より、

$$E_Z^{(t+1)}[IC(\boldsymbol{x}^N, \boldsymbol{z}^N; \boldsymbol{\theta}^{(t+1)}] \leq E_Z^{(t+1)}[IC(\boldsymbol{x}^N, \boldsymbol{z}^N; \boldsymbol{\theta}^{(t)}], \tag{14}$$

が成立する。*Jensen* の不等式より、

$$E_Z^{(t+1)}[\log P(\boldsymbol{z}^N|\boldsymbol{x}^N; \boldsymbol{\theta}^{(t+1)}] \leq E_Z^{(t+1)}[\log P(\boldsymbol{z}^N|\boldsymbol{x}^N; \boldsymbol{\theta}^{(t)}] \tag{15}$$

が成立するため[8]、両辺を *(14)* の両辺に加えて、$\ell(\hat{\boldsymbol{\theta}}^H) = E_Z[\sum_{c=1}^{C} \ell(\hat{\boldsymbol{\phi}}_c^{S_c}) + \ell(\hat{\boldsymbol{\pi}})]$ を考慮すると、

$$IC(\boldsymbol{x}^N; H^{(t+1)}) \leq IC(\boldsymbol{x}^N; H^{(t)}) \tag{16}$$

となる事がわかる．□

定理 2 によって、AIC を含む $\ell(\hat{\boldsymbol{\theta}}^H) = E_Z[\sum_{c=1}^{C} \ell(\hat{\boldsymbol{\phi}}_c^{S_c}) + \ell(\hat{\boldsymbol{\pi}})]$ を満たす情報量基準のクラスに対して，EMIC における $IC(\boldsymbol{x}^N; H)$ の単調減少性が証明された．

定理 1 および定理 2 より，MDL/BIC，AIC に関しては，適当な正則条件の元で，EMIC の最適化処理によって計算されるモデルおよびパラメータ $(H^{(t)}, \hat{\boldsymbol{\theta}}^{(t)})$ が，$IC(\boldsymbol{x}^N; H)$ の停留点へ収束する事がわかる．これらの定理によって，EMIC が，対応する情報量基準に関する最適化という意味において正当性を持つことが示された．

4 実験と考察

本稿では，2 つの異種混合モデルに関して、EMIC の評価を実施した．1 つ目のモデルは，異なる独立性を持つ 3 次元正規分布[9]の混合モデルで GAUSS と表記する．2 つ目のモデルは，異なる次数を持つ多項曲線の混合モデルで POLY と表記する．また，以下の結果は全て 10 回の実験の平均 (または 10-fold 交差検定) の結果であり，全ての手法が 10 個の異なる初期値に対して最適化を実施し，最適解を探索した．また，各実験における真のモデルを H^*，真のコンポーネント数を C^* と表記する．

4.1 人工データに対する性能評価

まず，人工データを利用して EMIC の計算速度およびモデル選択の性能に関して評価を行った．本実験では，真のモデルに関する一致性を評価するために，MDL を基準として利用した[10]．

4.1.1 比較手法

MDL を利用した EMIC (EMIC_{MDL}) に関する比較手法として，全モデル候補に対して 1 回ずつ探索を行う

[7] 正確には劣確率分布であるが，確率分布とみなしても議論に差はないため，簡単のため確率分布として議論をする．

[8] $E_Z^{(t+1)}[\bullet] = \sum_Z \bullet P(Z|\boldsymbol{x}, \boldsymbol{\theta}^{(t)})$ である点に注意。

[9] コンポーネントの候補は独立性の違いで 8 種類。3 次元と限定した理由は、高次元とした場合に比較手法の計算量が膨大となり実験が難しいため。

[10] 混合モデルは，特異モデルであるため MDL/BIC、AIC の導出における漸近展開が発散してしまい，各基準が正しく機能しない場合がある事が知られている．しかし，実用上で真のモデルへの一致性能を評価する事は意味がある．なお，AIC は真のモデルへの一致性はないため，本実験では MDL/BIC を採用している．

FULL$_{MDL}$ と，パタン探索 [9] によってモデル候補を探索する PAT$_{MDL}$ を評価した．各手法ともパラメータ推定には EM アルゴリズムを用いた．なお，EM アルゴリズムによるパラメータ推定は，大域的な最適解を得ることができないため，EMIC$_{MDL}$ および PAT$_{MDL}$ の推定精度が FULL$_{MDL}$ よりも良い可能性がある (EMIC$_{MDL}$ および PAT$_{MDL}$ は、同じモデルを複数回探索する可能性があるため)．

4.1.2 評価基準

モデル選択の性能を評価するために，1) 情報量基準の値 $MDL(\boldsymbol{x}^N;\bar{H})$，2) $\bar{H}=H^*$ となった割合 (R_H)，3) $\bar{C}=C^*$ となった割合 (R_C)，4) 最適化の CPU 時間，を評価基準とした．定性的には，1) は目的関数に対する最適解の探索精度，2) はコンポーネント数およびコンポーネントの種類まで含めたモデル選択の精度，3) はコンポーネント数に関するモデル選択の精度を表す．

4.1.3 真のモデルとデータの生成

GAUSS の真のモデルの各コンポーネントは，平均ベクトルを $[-5,5]$ より，共分散行列の要素を $[0,1]$ よりランダムに生成し，各次元の独立性もランダムに決定した．GAUSS に関しては真のコンポーネント数を $C^*=5$ とし，探索するコンポーネントの最大数 $C_{\max}$ を変化させた場合の性能の変化を評価した．

POLY の真のモデルの各コンポーネントは，図 1 に示される多項曲線からランダムに 4 つを選択した[11]．データは各曲線に対して平均が 0 の正規分布にしたがうノイズを加えて生成し，分散は $[1,4]$ でランダムに決定した．POLY に関しては、$C_{\max}=5$ と固定し，探索する次数の最大値 $D_{\max}$ を変動させた場合の性能の変化を評価した．

なお，両方のモデルで混合比はランダムに設定した．また，各結果ともデータ数は $N=500$ とした．また，最適化に関しては EMIC では情報量基準に対して，EM アルゴリズムに関しては対数尤度の値に対して 10^{-6} の閾値を設けて最適化を打ち切っている．

4.1.4 結果と考察

図 2 に，GAUSS に関する結果を示す．$C_{\max}$ の増加とともに，PAT$_{MDL}$ および FULL$_{MDL}$ の CPU 時間が急速に増加するのに対して，EMIC では線形で CPU 時間が増加している事が確認できる (左上図)．これは，探索するコンポーネントの最大数 $C_{\max}$ に対して EMIC が

[11] 多項曲線の次数と係数をランダムとした実験では，複数の曲線が重なってしまいモデル選択がほぼ不可能なケースが多く発生したため，本稿の実験では評価のために図 1 に示される曲線を真のコンポーネントとした．

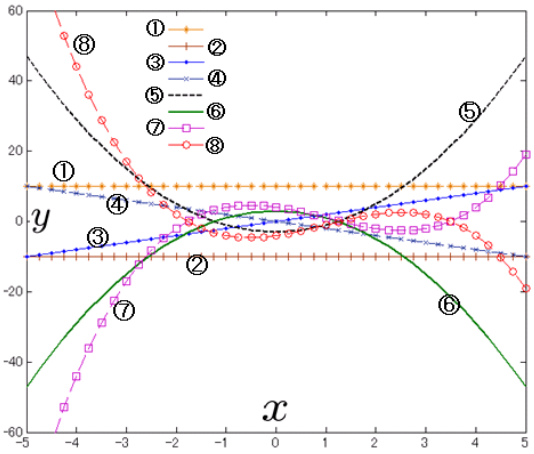

図 1: POLY の真のモデルに関する 8 つコンポーネント．①:$y=10$, ②:$y=-10$, ③:$y=2x$, ④:$y=-2x$, ⑤:$y=2x^2-3$, ⑥:$y=-2x^2+3$, ⑦:$y=0.5x^3-1.5x^2-2x+4$, ⑧:$y=-0.5x^3+1.5x^2+2x-4$.

線形時間の計算量で異種混合モデル選択を実施できる事を裏付ける結果である．さらに，真のモデルの推定性能では，R_H および R_C の両方で，EMIC は他の手法を凌駕している．これは，同一基準を利用した通常の EM アルゴリズムに基づく手法と比較して，EMIC の探索が局所解に対してロバストにモデルを推定可能である事を示している．EMIC はパラメータとモデルの空間を同時に探索するため，パラメータ空間で性質の悪い局所解に落ちた場合 (あるいは落ちる前) に，モデルを変更して探索を継続可能であるため，よりよい解が得られたと考えられる．

図 3，POLY に関する結果を示す．CPU 時間に関しては，$D_{\max}$ の増加とともに，PAT$_{MDL}$ および FULL$_{MDL}$ では探索に必要な時間が急速に増加するのに対して，EMIC では線形に増加している事が確認できる (左上図)．これは，コンポーネントの候補数[12]の増加に対して，EMIC が線形時間の計算量で異種混合モデル選択を実施できる事を裏付ける結果である．また，GAUSS の場合と同様に，$IC(\boldsymbol{x}^N;\bar{H})$，$R_H$，$R_C$ に関しても，EMIC$_{MDL}$ がその他の手法よりも優れた性能を発揮している事が確認できる．

次に，EMIC の EM ステップを詳しく分析するために，POLY に関して，各ステップにおけるモデルの推定結果の例を，図 4 に示す．t は EM ステップの回数，F_i は i 番目のコンポーネントに対応する曲線，D_i は F_i の次数をそれぞれ表している．$t=1$ (左上) は，隠れ変数をランダムに初期化した際に推定されたコンポーネントで，真

[12] 0 次まで合わせて $D_{\max}+1$.

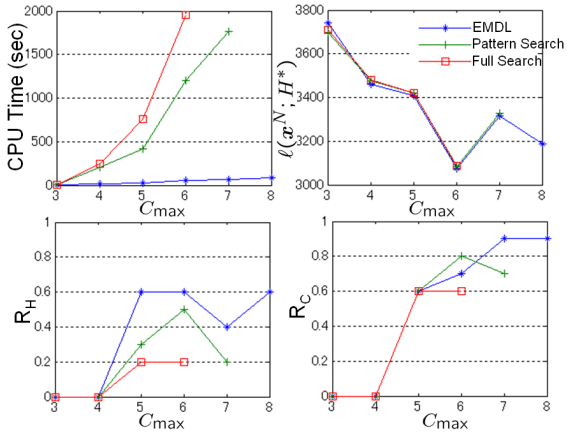

図 2: GAUSS に対する EMIC_{MDL}, PAT_{MDL}, FULL_{MDL} のモデル選択性能の比較.

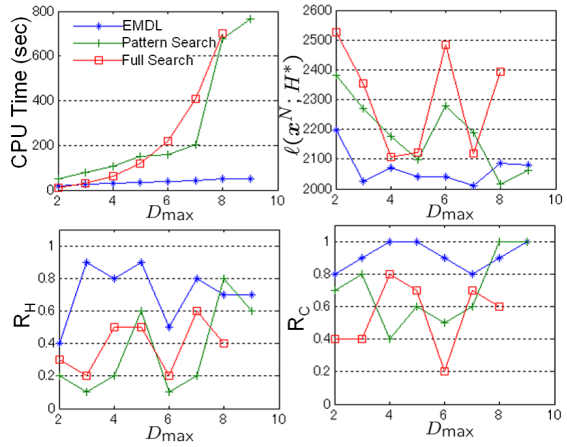

図 3: POLY に対する EMIC_{MDL}, PAT_{MDL}, FULL_{MDL} のモデル選択性能の比較.

の曲線を全く再現できていない．EMIC による最適化が $t = 3$, $t = 10$, $t = 20$ と進むとともに，推定された曲線が徐々にデータの適合していく様子が確認できる．ここで，各コンポーネントに対する最適な次数 $(D_1 \sim D_4)$ が最適化とともに頻繁に変動している点は注目に値する．変動の様子も，真のコンポーネントへ次数が単調に近づくもの (例えば F_1) や，真のコンポーネントよりも複雑なモデルを推定してから真のコンポーネントへ収束するもの (例えば F_4 の $t = 10$) など，コンポーネントによってモデル空間の探索の様相が異なる点で興味深い．これは，期待されるデータ数 $\sum_{n=1}^{N} E_Z[z_{nc}]$ が小さい場合には単純なモデルが，大きい場合には複雑なモデルが選ばれやすいため，各ステップにおける隠れ変数の事後分布の推定 (E ステップ) とコンポーネントごとのモデル選択 (M ステップ) が密接に関連しているためであろう．なお，この例では $t = 20$ で最適化が収束し，推定されたモデルは真のモデルと一致した．

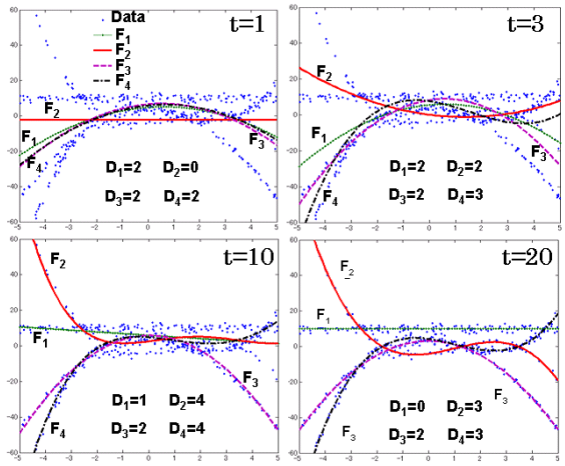

図 4: EMIC における POLY モデル推定プロセス.

4.2 UCI データに対する性能評価

次に，UCI データセット [2] の ecoli, housing, iris, wine, yeast, vowel context に関して，各手法の比較実験を実施した．各データについて，真のモデルはわからないため，評価指標としては交差検定尤度[13]と CPU 時間を行った．また，本評価では EMIC，パタン探索，全探索の全てに対して情報量基準として MDL および AIC の両者の比較を行った．実験では，交差検定の各繰り返しにおいて，ランダムに 3 つの属性を選択し，3 次元の GAUSS モデルを学習させた．属性を 3 つと限定したのは，それ以上の場合にはモデル候補の数が増大し，比較手法の計算が終了しなかったためである．同様に，比較手法の計算コストの問題からコンポーネントの最大数は $C_{\max} = 5$ とした．

表 1 に，比較結果を示す．交差検定尤度に関しては各手法とも大きな差はなく，データによって最も性能がよかった手法が異なり、予測精度の意味でよいモデルおよびパラメータの探索性能に関しては，手法による違いを明確に観察する事はできなかった．一方で，推定にかかった CPU 時間をみると，MDL および AIC 共に，EMIC は他の手法と比較して圧倒的に短い時間で解を探索できている事が確認できる．この結果は，真のモデルがわからない実応用においても，EMIC によって全探索などと同程度の精度かつ高速に異種混合モデルを推定可能である事を示している．

[13] 交差検定に関するテストデータに対して計算された尤度.

表 1: UCI データに関する交差検定尤度と CPU 時間 (尤度/CPU 時間) の比較

	ecoli	housing	iris	wine	yeast	vowel context
EMIC_{MDL}	70.6/39.8	-219.6/25.6	-46.5/27.8	-73.3/35.5	558.6/85.5	-553.8/69.4
PAT_{MDL}	72.3/723.0	-239.1/346.7	-47.7/408.2	-70.6/691.3	556.2/1252.9	-550.2/800.7
FULL_{MDL}	72.3/856.2	-239.3/533.5	-49.9/612.7	-69.2/780.9	556.9/1630.9	-555.2/1317.5
EMIC_{AIC}	70.1/35.5	-228.6/23.5	-52.3/23.4	-71.7/30.0	557.8/81.0	-548.8/64.0
PAT_{AIC}	67.4/607.6	-250.1/342.9	-51.7/429.4	-74.8/526.7	558.8/1158.4	-553.1/1099.4
FULL_{AIC}	65.0/855.9	-239.2/538.3	-50.8/610.6	-75.1/779.2	558.6/1635.8	-550.6/1313.2

5 おわりに

本稿では，複数の異種多様なモデルが混合する，異種混合モデルに対するモデル選択問題に対する一般的な枠組みとして，期待情報量基準最小化法 (EMIC) を提案した．EMIC は，任意のコンポーネントを扱う事が可能であり，コンポーネント数とコンポーネント候補数の双方に対して線形時間でモデルを推定が可能，という二つの利点を持つ．本稿では，一部の情報量基準に対して EMIC の理論的な正当性を証明するとともに，実証実験によってその有効性を確認した．

本稿の中心として議論した MDL/BIC や AIC を含む漸近展開から導出される情報量基準は，混合分布のような特異モデルに対して，実用上は重要であるが基準としての理論的正当性は失ってしまう．今後，漸近展開に基づかない基準[14]など，その他のモデル選択基準に関しても EMIC を拡張し議論を発展させる必要がある．この際に，特に重要なのは，期待情報量基準の最小化プロセスが，元の情報量基準の最小化につながるかという点であろう．

また，本稿では EMIC の基礎的な性質を理解するために，比較的単純な異種混合モデルに対して実証実験を実施したが，より複雑な異種混合モデルに対する評価が今後の重要な課題として挙げられる．

参考文献

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Caski, editors, *Proceedings of the 2nd International Symposium on Information Theory*, pages 267–281, 1973.

[2] A. Asuncion and D. Newman. UCI machine learning repository, 2007.

[3] C. Bishop and M. Tipping. A hierarchical latent variable model for data visualization. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 20(3):281–293, 1998.

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from imcomplete data via the em algorithm. *Journal of the Royal Statistical Society*, B39(1):1–38, 1977.

[6] Z. Ghahramani and M. J. Beal. Variational inference for bayesian mixtures of factor analysers. In *In Advances in Neural Information Processing Systems 12*, pages 449–455. MIT Press, 2000.

[7] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214, 1994.

[8] Y. Ma, H. Derksen, W. Hong, and J. Wright. Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1546–1562, 2007.

[9] M. Momma and K. P. Bennett. A pattern search method for model selection of support vector regression. In *In Proceedings of the SIAM International Conference on Data Mining*. SIAM, 2002.

[10] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[11] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific Pub Co Inc, 1989.

[12] J. Rissanen. Fisher information and stochastic complexity. *IEEE Transaction on Information Theory*, 42(1):40–47, 1996.

[13] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[14] R. A. Tapia and J. R. Thompson. *Nonparametric Probability Density Estimation*. Johns Hopkins Univ. Press, 1978.

[15] K. Yamanishi. Advances in minimum description length: Theory and applications. In P. D. Grunwald, editor, *Extended Stochastic Complexity and Its Applications to Learning*. The MIT Press, 2005.

[14]例えば MDL 基準における予測確率的コンプレキシティ [11] や，交差検定によるモデル選択基準など。