LH-004

少数サンプルを基にした独立な確率表の混合表現
Mixture of Independent Tables Estimated with Small Samples

藤本 悠 † 村田 昇 †
Yu Fujimoto Noboru Murata

1. まえがき

多値の質的データの解析を行なう際，度数表を基にした対数線形モデルを考えることで有用な知見をしばしば得ることができる [1]．モデル化する対象の詳細な記述が求められる際には度数表は巨大なものになるが，それに伴うデータの数は一般に不足する場合が多い．我々は各変数の取り得る状態数が多いために度数表が巨大化する場合に，対数線形モデルの代わりに独立な変数関係の混合表現を用いることで，少ないパラメタで同時分布を表現することを提案してきた [5]．

しかし，少数サンプルの下で対数尤度を指標としたEMアルゴリズムを用いてモデル推定を行うとデータの少なさに起因する分布の歪みをそのまま反映してしまう．本稿ではこの歪みに対してロバストな性質を持つ β 尤度を指標とした推定を行うことで少数データを用いたモデル推定結果が改善されることを実験的に示す．

2. 確率表の混合表現と問題点

2.1 確率表の混合表現

各変数の状態数の多さが原因で巨大になっている 2 元の度数表から変数間の同時確率分布を推定することを考える．一般に用いられる対数線形モデルでは候補となる 2 つのモデル，独立モデルと飽和モデルの間には著しいパラメタ数の差が存在する．そのため場合によってはどちらのモデルを選択しても適度なパラメタ数で適切な推定精度を実現することが不可能になってしまう．そこで我々は独立な変数関係を複数混合することで，より適切なモデルを得るという提案を行った [5]．

2 変数 $X \in \{x_1, \ldots, x_I\}$,$Y \in \{y_1, \ldots, y_J\}$ の同時確率のモデル化を例にとる．2 変数が互いに独立である場合，

$$\mathrm{P}(x_i, y_j) = \mathrm{P}(x_i)\mathrm{P}(y_j) \quad (1 \le i \le I, 1 \le j \le J) \quad (1)$$

が成り立つ．独立な変数関係を複数混合するということは，式 (1) のような周辺確率の積で表現される確率表を K 枚重ね合わせることで元の同時確率表を

$$\mathrm{P}(x_i, y_j) = \sum_{k}^{K} \pi_k \overbrace{\mathrm{P}_k(x_i)\,\mathrm{P}_k(y_j)}^{k\text{ 枚目の独立な表}} \quad (2)$$

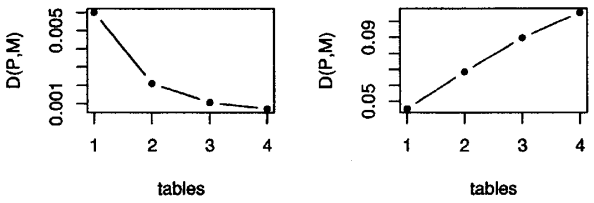
のように表現することを意味する．ただし π_k は混合比を表すパラメタで $\sum_{k=1}^{K} \pi_k = 1$，$\pi_k > 0$ $(k = 1, \ldots, K)$ である．変数の状態数 I，J が大きい場合，式 (2) において適切な K を実験的に設定することによって，適度なパラメタ数で，より高い精度を持つモデルの候補を獲得することができる．

度数表における (x_i, y_j) が同時に観測される頻度を T_{ij} とすると，パラメタは EM アルゴリズム [4] によって次式に示す対数尤度 $l(\theta)$ の最大 (極大) 値を探索することで推定が可能である．

$$l(\theta) = \sum_{i,j} \sum_{k} T_{ij} \log \pi_k \mathrm{P}_k(x_i)\mathrm{P}_k(y_j) \quad (3)$$

ここで θ は混合モデルのパラメタの集合で π_k，$\mathrm{P}_k(x_i)$，$\mathrm{P}_k(y_j)$ を表している．

†早稲田大学理工学研究科



(a) データが充分な場合　(b) データが少ない場合

図 1: D(P, M) と混合する表の数の関係

2.2 少数サンプルにまつわる問題

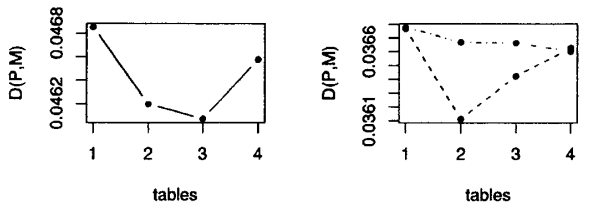
対数尤度を基にした EM 推定を行なうと，サンプル数 $n = \sum_{i,j} T_{ij}$ が度数表の大きさに対して相対的に少ない場合に良い推定結果が得られないことがある．簡単な例として，3 枚の独立な 30×30 の確率表の混合で表現される同時確率分布 P と，P に従うデータから構成される度数表を基に推定した分布 M の間の距離を KL 情報量 D(P, M) で測る．式 (2) の混合モデルにおいて $1 \le K \le 4$ で変化させた時の D(P, M) の推移を図 1 に示す．

推定に用いた度数表のサンプル数が $n = 9000000$ と充分に大きい場合，図 1(a) のように D(P, M) は K が大きくなるにつれて減少し，$K = 3$ の時に 0 付近へと収束する．一方で $n = 450$ とそれほど多くない時，図 1(b) のように D(P, M) は K の増加と共に減少しない場合がある．これは推定に用いている度数表のサンプルの少なさが原因となって，データの分布が真の分布 P を反映しきれずに歪みを含んでしまっているためである．

このように少数データを用いて EM 推定を行なう際には真の分布からかけ離れたモデルを構築してしまう可能性がある．

3. β 尤度に基づいた EM アルゴリズム

理想的な条件下ではモデル M のパラメタを対数尤度に基づく最尤推定によって求めることで，真の分布 P との間の D(P, M) の小さなモデルの発見を期待できる．一方でサンプルが少ない場合やデータが外れ値を含む場合には，D(P, M) が小さくなるとは限らない．推定時のデータの外れ値の影響を避けるために用いる尺度として対数尤度の最大化の代わりに，Hellinger 距離の最小化 [2] や，β 尤度の最大化 [3] といった基準が提案されてい



(a) $K=3$ が最適となる例　(b) 最適とならない例

図 2: β 尤度に基づく推定結果の典型的な例

る．いずれも外れ値の存在に起因する分布の歪みに対してロバストな尺度を導入することによって推定の精度の向上を計っている．我々はサンプルの少なさに起因する分布の歪みを同様に扱い，β 尤度の最大化を実現することで 2.2 節に挙げたような問題の回避を目指す．

混合表現における β 尤度 $l_\beta(\theta)$ は次のようになる．

$$l_\beta(\theta) = \frac{1}{n\beta}\sum_{i,j} T_{ij}\left(\sum_k \pi_k \mathrm{P}_k(x_i)\mathrm{P}_k(y_j)\right)^\beta - \frac{1}{1+\beta}\sum_{i,j}\left(\sum_k \pi_k \mathrm{P}_k(x_i)\mathrm{P}_k(y_j)\right)^{1+\beta} \quad (4)$$

ここで β は $0 < \beta \leq 1$ の値である．混合モデルの推定時には式 (3) の代わりに式 (4) の β 尤度に基づいた Q 関数を設計することで EM 推定が行える．

4. 実験結果

4.1 少数サンプルを用いた β 尤度に基づく推定実験

2.2 節で例示した実験設定において β 尤度を指標とすることで少数サンプルを用いた推定が改善されるかを確認する．2.2 節と同様に，3 枚の 30×30 の独立な確率表の混合と等価な P と，$n = 450$ の度数表から β 尤度に基づく EM 推定を行なったモデル M との間の $\mathrm{D}(P, M)$ の推移を図 2 に示す．なお，推定には $\beta = 0.5$ の値を用いている．

図 2(a) では $K = 3$ の時に $\mathrm{D}(P, M)$ が最小値を示している．これは少数サンプルを用いた確率表の推定の際に β 尤度を基にすることで，真の分布により近いモデルのパラメタを探索できていることを表している．一方，P から生成した $n = 450$ の別な度数表に対して同様の推定を行なうと，図 2(b) の 2 例のように $K \neq 3$ で $\mathrm{D}(P, M)$ が最小となる場合も確認された．このような最良モデルの K の違いの一因としてはデータ分布の違いに対する β の値の不適切さが考えられる．なお，いずれの度数表の場合も対数尤度を基に推定した場合には図 1(b) のように $K = 1$ が最小となった．

4.2 対数尤度，β 尤度に基づく推定結果の比較実験

対数尤度，β 尤度を用いて EM 推定を行なうことで推定に用いるサンプル数とどのような関係があるのか比較を行なった．まず 3 枚の独立な確率表の重ね合わせで表現される P に従って生成される総サンプル数 n の 10×10 の度数表を 20 枚用意する．n を 50 から 10000 まで変化させて度数表の生成を繰り返し，これらの度数

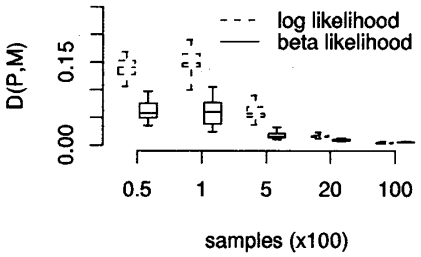


図 3: サンプルの数と推定モデルの $D(P, M)$ の関係

表から $K = 3$ の混合モデルのパラメタを対数尤度，β 尤度 ($\beta = 0.3$) に基づく EM 推定で求めた．n と $\mathrm{D}(P, M)$ の関係を記したものが図 3 である．

図 3 ではサンプル数が充分多い場合には対数尤度，β 尤度のどちらを用いても $\mathrm{D}(P, M)$ は 0 に収束していることが分かる．一方でサンプル数が少ない場合には対数尤度に基づく推定結果がそれほど良く無いのに対して，β 尤度に基づく推定結果は $\mathrm{D}(P, M)$ が改善されている．このことからも β 尤度を指標とすることで少数サンプルを用いた場合に良い EM 推定が可能であることが確認できる．また，充分なサンプルの下では尤度の種類に関わらず結果が一致することから，β 尤度を用いた EM 推定の精度の優位性が言える．

5. まとめ

本稿では少数サンプルを用いた独立な確率表の混合モデルのパラメタ推定を議論し，ロバストな性質を持つ β 尤度を指標とする EM 推定を行うことでモデルの推定結果が改善されることを示した．一方で P，データの分布，β の違いによって最良モデルの混合表数に揺らぎが確認されたため，引き続き異なる状況下での提案手法の効果について検証を行なう．

また実データに対して提案手法を用いる際にはさらに適切な β の選択や P が未知の場合の評価などを論じる必要がある．本研究が対象としている状況では充分なサンプル数を想定した AIC などの情報量基準によるモデル評価は不適切であるため，Cross Validation などを用いた実データへの適用例も今後示していく予定である．

参考文献

[1] A. Agresti. *Categorical Data Analysis.* Wiley Inc., 2 edition, 2002.

[2] R. Beran. Minimum Hellinger distance estimates for parametric models. *The Annals of Statistics*, 5(3):445–463, May 1977.

[3] H. Fujisawa and S. Eguchi. Robust estimation in the normal mixture model. Research memorandum, Feb. 2003.

[4] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions.* Wiley-Interscience, 1996.

[5] 藤本, 村田. ベイジアンネットによる即興音楽生成システム-少ない曲例からのモデル推定-. 第 3 回情報科学技術フォーラム, 2004.