Technical Report of the 3rd Workshop on Latent Dynamics (Sep 24, 2012, Tokyo, Japan)

潜在変数の分布推定誤差に関する漸近解析

山崎 啓介*
Keisuke Yamazaki

Abstract: 潜在変数を含むパラメトリックモデルの使われ方にはデータの「予測」と「分析」の２つの側面がある．予測精度を測る汎化誤差はこれまで様々な統計的性質が明らかにされてきた．一方でクラスタリングなどの潜在変数推定は分析にあたるが理論的な評価は十分に行われていない．本稿では分布推定の精度について誤差関数を定式化し, 最近得られた漸近解析の結果を紹介する．特に最尤法とベイズ法の誤差を比較することで潜在変数推定が予測とは異なる性質をもつことを示す．

Keywords: 教師無し学習, 階層モデル, 最尤推定, ベイズ推定, 推定精度

1 はじめに

混合分布や隠れマルコフモデル, ベイジアンネットワークなど階層構造をもつパラメトリックモデルは機械学習やデータマイニングなどで広く用いられている．これらのモデルは観測データを表現する変数と隠れた構造を表す変数を有する．本稿では前者を観測変数, 後者を潜在変数とよぶ．モデルの使われ方は将来のデータを推定する「予測」と, 潜在変数を用いて表されるモデル内部の構造や状態を推定する「分析」に大別される．例としてラベル無しデータが与えられたときの混合分布の用途を考える．このモデルでは潜在変数がラベルを表す．データの分布を推定し次に出現するものを言い当てるのが予測であり, 手持ちのデータのラベルを推定するのが分析である．

観測変数の予測については様々な統計的性質が知られている．特にデータ分布の推定精度を KL ダイバージェンスで評価した汎化誤差は多くの条件下でその漸近形が導出されている．潜在変数の次元や範囲はデータから直接決定することができないため, 真のものに比べ (1) 不足している場合, (2) 過不足がない場合, (3) 冗長な場合が考えられる．最尤推定や MAP 推定では (1) と (2) の冗長性がない場合において漸近形が知られており, ベイズ推定ではそれらに加え代数幾何学を用いることで (3) の漸近形が近年明らかになった [1].

このような状況に対し, 潜在変数の推定はクラスタリングに代表される教師無し学習のタスクとして重要であるにも係わらず, その精度の理論的な評価は十分に行われていない．本稿では分布推定に焦点を絞りし, 真の潜在変数分布からの KL ダイバージェンスを誤差関数として推定精度を考察する．観測変数の予測と推定対象が異なるため最尤推定とベイズ推定を改めて定義し, それぞれの推定法に対し最近得られた誤差関数の漸近形を紹介する [2, 3]．これにより推定法の性能比較が可能となり, さらに推定精度が予測の場合と異なる性質をもつことを明らかにする．

2 潜在変数の同時分布推定

ここでは潜在変数の同時分布について最尤推定とベイズ推定の定義を与える． 観測されたデータを $X^n = \{x_1, \ldots, x_n\}$ とし, これに対応するラベルを $Y^n = \{y_1, \ldots, y_n\}$ とする． $\{X^n, Y^n\}$ を完全データと呼び, これに対し X^n を不完全データと呼ぶ．学習モデルを $p(x, y|w)$ とする. ここで x, y はそれぞれ観測変数と潜在変数であり w はパラメータである. 本稿では潜在変数は離散とする. 観測変数の分布は

$$p(x|w) = \sum_{y=1}^{K} p(x, y|w)$$

で与えられる. つまり学習モデルがもつ潜在変数を $y \in \{1, \ldots, K\}$ とした．混合正規分布の場合では混合比 a_k とパラメータ b_k をもつ正規分布 $\mathcal{N}(x|b_k)$ を用いて

$$p(x|w) = \sum_{k=1}^{K} a_k \mathcal{N}(x|b_k)$$

で表されるため, $p(x, y = k) = a_k \mathcal{N}(x|b_k)$ とするモデルである．他にも多くの階層型モデルが同様の形式で表現される．

*東京工業大学大学院 知能システム科学専攻, 〒 226-8503 横浜市緑区長津田 4259 G5-19, e-mail k-yam@math.dis.titech.ac.jp, Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku, Yokohama

完全データと不完全データの同時確率はそれぞれ

$$p(X^n, Y^n|w) = \prod_{i=1}^{n} p(x_i, y_i|w),$$

$$p(X^n|w) = \prod_{i=1}^{n} p(x_i|w)$$

で表される.

真の分布 $q(x, y)$ を仮定し $\{X^n, Y^n\}$ はこれから独立に生成されるとする. $1 < K^* \leq K$ となる整数 K^* を用いて真の分布における潜在変数を $y \in \{1, \ldots, K^*\}$ とする. 観測データが与えられた下での真の潜在変数の同時分布は

$$q(Y^n|X^n) = \prod_{i=1}^{n} q(y_i|x_i) = \prod_{i=1}^{n} \frac{q(x_i, y_i)}{\sum_{y_i=1}^{K^*} q(x_i, y_i)}$$

と表現される. 本稿では真の分布が学習モデルで表現可能であるとする. つまり $q(x, y) = p(x, y|w^*)$ を満たす w^* の集合が空でないとする.

潜在変数の推定は学習モデルを用いて潜在変数の同時分布 $p(Y^n|X^n)$ を構成することである. まず最尤推定を以下のように定義する. 観測データに対する最尤推定量は

$$\hat{w} = \arg\max_w P(X^n|w)$$

で与えられる. これを用いて同時分布を

$$p(Y^n|X^n) = \prod_{i=1}^{n} p(y_i|x_i, \hat{w}) = \prod_{i=1}^{n} \frac{p(x_i, y_i|\hat{w})}{p(x_i|\hat{w})}$$

とする. 分散をパラメータにもつ混合正規分布などいくつかのモデルにおいて最尤推定量が発散する場合があるが, 本稿では真のパラメータ w^* に収束する場合に限定して議論を進める. 次にベイズ推定の定義を述べる. ハイパーパラメータ η を有する事前分布を $\varphi(w|\eta)$ とし, 完全データに対する周辺尤度を

$$Z(X^n, Y^n) = \int p(X^n, Y^n|w)\varphi(w|\eta)dw$$

とする. この周辺尤度を用いて潜在変数の同時分布を

$$p(Y^n|X^n) = \frac{Z(X^n, Y^n)}{\sum_{Y^n} Z(X^n, Y^n)}$$

とする. 定義より分母は不完全データの周辺尤度

$$Z(X^n) = \int p(X^n|w)\varphi(w|\eta)dw$$

に等しい.

潜在変数の推定を可能とする条件として以下のものを考える. データの生成過程において, 観測変数の分布 $p(x|w^*)$ は w^* 全ての影響を受けると仮定する. つまり次のモデルは本稿の議論から除外する.

定義 1 (潜在変数推定として不適切な生成モデル) パラメータが $w = \{w_1, w_2\}$ と分離可能であり, 観測変数と潜在変数の同時分布が

$$p(x, y|w) = p(x|w_1)p(y|x, w_2)$$

と表現される.

このモデルは観測変数のみの分布 $p(x|w) = p(x|w_1)$ がパラメータ w_2 の情報を含まないため, 観測データ（不完全データ）のみから潜在変数を推定することができない.

3 推定精度の定式化

ここでは推定精度を評価するための誤差関数を定式化しその特徴を述べる. 本稿では潜在変数の真の同時分布と推定された同時分布を比較し誤差関数とする. 分布の違いを示す量として KL ダイバージェンスを用いると, 誤差関数は以下のように定義できる.

$$D(n) = \frac{1}{n} E_{X^n}\left[\sum_{Y^n} q(Y^n|X^n) \ln \frac{q(Y^n|X^n)}{p(Y^n|X^n)}\right].$$

ここで $E_{X^n}[\cdot]$ は観測データの出方での平均を意味する. データ数 n で正規化されているため潜在変数 1 つあたりの平均誤差となる. 学習モデルの潜在変数が $K^* > K$ では誤差関数は無限大に発散する. 以降では $K^* \leq K$ の場合を考える.

階層構造を有するモデルは潜在変数に入れ替え対称性が存在するが, 誤差関数 $D(n)$ は真の分布における変数順序を基準としている. 本稿では順序を含めて最良の推定を行った場合の誤差を解析する.

また $K^* < K$ では誤差関数における潜在変数の和が K^* までしかないため, $K^* + 1$ 以上の変数値を用いた推定は精度を悪化させる. この場合, 誤差関数は学習モデルの冗長性の影響を受けると予想される.

4 誤差関数の漸近形

ここでは $D(n)$ の漸近形を紹介し, 推定法や学習モデルの冗長性など異なる条件下での誤差を比較する.

前章までの仮定よりデータ数 n が増えるにしたがって最良の予測結果 $p(Y^n|X^n)$ は真の分布 $q(Y^n|X^n)$ へ収束する. つまり誤差関数 $D(n)$ は $n \to \infty$ において零となる. 以下に紹介する誤差関数の漸近形は収束のオーダを示すものである.

まず $K^* = K$ の場合を考える. 不完全データと完全データのフィッシャー情報行列 I_X, I_{XY} を次のように定

義する.

$$\{I_X\}_{ij} = E_{xy}\left[\frac{\partial \ln p(x|w^*)}{\partial w_i}\frac{\partial \ln p(x|w^*)}{\partial w_j}\right],$$
$$\{I_{XY}\}_{ij} = E_{xy}\left[\frac{\partial \ln p(x,y|w^*)}{\partial w_i}\frac{\partial \ln p(x,y|w^*)}{\partial w_j}\right].$$

ここで平均は

$$E_{xy}[f(x,y)] = \int \sum_{y}^{K^*} f(x,y)p(x,y|w^*)dx$$

とした. このとき以下の 2 つの定理が成立する.

定理 2 最尤推定において誤差関数 $D(n)$ の漸近展開は次式で表される.

$$D(n) = \frac{1}{2n}\mathrm{Tr}\left[\{I_{XY} - I_X\}I_X^{-1}\right] + o\left(\frac{1}{n}\right).$$

定理 3 ベイズ推定において誤差関数 $D(n)$ の漸近展開は次式で表される.

$$D(n) = \frac{1}{2n}\ln\det\left[I_{XY}I_X^{-1}\right] + o\left(\frac{1}{n}\right).$$

これらの結果より誤差の比較が可能となる.

系 4 最尤推定とベイズ推定における誤差関数をそれぞれ $D_{ML}(n)$, $D_{Bayes}(n)$ とする. 行列 $I_{XY}I_X^{-1}$ の正定値性を仮定すると,

$$D_{ML}(n) \geq D_{Bayes}(n)$$

が成り立つ.

観測変数の予測は $p(x|X^n)$ を構成することに対応し, 最尤推定とベイズ推定でそれぞれ

$$p(x|X^n) = p(x|\hat{w}),$$
$$p(x|X^n) = \int p(x|w)\frac{p(X^n|w)\varphi(w|\eta)}{Z(X^n)}dw$$

と定義される. 分布推定についての汎化誤差は

$$G(n) = E_{X^n}\left[\int q(x)\ln\frac{q(x)}{p(x|X^n)}dx\right]$$

で与えられる. 汎化誤差は最尤推定, ベイズ推定ともに次式の漸近形をもつことが知られている.

$$G(n) = \frac{\dim w}{2n} + o\left(\frac{1}{n}\right).$$

誤差関数 $G(n)$ が 2 つの推定法で同じ漸近形となるのに対し, $D(n)$ は推定法によって異なる. さらに $G(n)$ はパラメータ次元のみに依存するが, $D(n)$ はフィッシャー情報行列に含まれるモデルの式 $p(x|w)$ や真のパラメータ w^* によって主要項の係数が変化することがわかる.

次に $K^* < K$ の場合を考える. 学習モデルに冗長性があるとパラメータ空間に特異点が生じることが知られている. この特異点の影響でフィッシャー情報行列が縮退するため, 最尤推定量の漸近挙動は未だ解明されていない. ここではベイズ推定の結果のみを紹介する.

2 つの KL ダイバージェンスを次式で定義する.

$$H_{XY}(w) = \int \sum_{y=1}^{K^*} q(x,y)\ln\frac{q(x,y)}{p(x,y|w)}dx,$$
$$H_X(w) = \int q(x)\ln\frac{q(x)}{p(x|w)}dx.$$

これらのダイバージェンスが実解析関数のとき, ゼータ関数

$$\zeta_{XY}(z) = \int H_{XY}(w)^z\varphi(w|\eta)dw,$$
$$\zeta_X(z) = \int H_X(w)^z\varphi(w|\eta)dw$$

の全ての極は実軸上, 負の有理数となることが知られている. ここで z は一変数複素数である. これらゼータの最大極とその多重度の組をそれぞれ $(-\lambda_{XY}, m_{XY})$ と $(-\lambda_X, m_X)$ とする.

定理 5 ベイズ推定において誤差関数 $D(n)$ の漸近展開は次式で表される.

$$D(n) = (\lambda_{XY} - \lambda_X)\frac{\ln n}{n} - (m_{XY} - m_X)\frac{\ln\ln n}{n} + o\left(\frac{\ln\ln n}{n}\right).$$

真の分布では潜在変数の空間が K^{*n} 次元であるのに対し学習モデルでは K^n である. 正しい推定を行うためには $K^n - K^{*n}$ 次元の冗長な空間に対する確率を零にする必要がある. $K^* = K$ の場合と比べ主要オーダが $1/n$ から $\ln n/n$ へ増加しているのはこのためのコストが大きいことを示している.

5 おわりに

本稿では潜在変数の分布推定について誤差関数を定式化し, その漸近形を紹介した. 観測変数の場合と異なり不完全データから推定誤差を計算することは原理的に不可能であるため, 漸近形を導出しその統計的振る舞いを知ることは重要と思われる. 現在は主に静的な状況を計算しているが, 今後はダイナミクスの解析に発展させたい.

謝辞

本研究の一部は栢森情報科学振興財団研究助成金, 倉田財団倉田奨励金および科研費 (24700139, 23500172) の助成を受けたものである.

参考文献

[1] S. Watanabe, Algebraic Geometry and Statistical Learning Theory, Cambridge University Press, New York, NY, USA, 2009.

[2] K. Yamazaki, “A theoretical analysis of KL-type generalization error on hidden variable distribution,” Technical Report NC2010-165, IEICE, 2011.

[3] K. Yamazaki, “An accuracy analysis of latent variable estimation with the maximum likelihood estimator,” Technical Report IBISML2011-55, IEICE, 2011.