F-050

GESによる補正を行った情報量に基づくパラメタ推定法の評価

Evaluation of Parameter Estimation Methods based on Divergence corrected with GES

藤本 悠 † 村田 昇 †
Yu Fujimoto Noboru Murata

1. まえがき

自然言語処理では，様々な単語や文章などを変数とした共起頻度に基づく同時確率表のパラメトリックなモデリングがしばしば行われる．一般に単語の種類が多くなることで，各変数の状態数が非常に多くなり確率表は巨大になる．このような表を少ないパラメタで表現するための代表的な方法としては，PLSA(Probabilistic Latent Semantic Analysis)[7] といったものが提案されている．PLSA では巨大な同時確率表を対象とした具体的なモデルとして，例えばアスペクトモデル [8] や潜在クラスモデル [1] と呼ばれる混合モデルが用いられる．例として 2 つの離散変数 A と B の組み合わせによって表現される $X=(A,B)$ を考えた時，X のある状態 $x=(A=a,B=b)$ においてデータが観測される確率を K 個の独立モデルの混合で表すと，

$$q(x;\theta)=\sum_{k=1}^{K}\pi_k p_k(a)p_k(b) \tag{1}$$

のように与えられる．ただしここで $p_k(a)$ と $p_k(b)$ は k 番目の独立モデルにおける A と B の周辺確率，π_k は混合率，θ は混合モデルのパラメタで $\theta=(\pi_1,\cdots,\pi_{K-1},p_1(A),p_1(B),\cdots,p_K(A),p_K(B))$ を表す．式 (1) の混合モデルのパラメタは EM アルゴリズム [10, 13] などを用いて推定することが可能となる．しかしこの混合モデルは各変数の状態数に応じてパラメタの数も増加するため，相対的に十分な数のデータが無い状況でそのまま推定してしまうと，オーバーフィットが深刻となる場合がある．これを緩和するための単純な方法としては，経験分布に微小な定数を足して正規化する方法や EM アルゴリズムが収束する前に止める方法などがあるが，近年，EM アルゴリズムの目的関数をオーバーフィットが緩和されるように改変する方法 [12, 5, 4] がいくつか提案されている．これらの方法では目的関数を従来のものから様々に変化させることによって柔軟なオーバーフィットの緩和を実現している．

このように同じモデルの推定を行うためにいくつかのアルゴリズムが存在している場合，種々の推定方法・推定結果の善し悪しは，一般のモデル選択における議論の枠組みで評価することが可能である．しかし多くのモデル選択の指標では，適切な評価を行うための前提として充分大量のサンプルがある状況を仮定している．そのため，推定時のオーバーフィットが深刻となるような少数のサンプルの下では適切な評価が行うことができない．本稿ではデータが少ない時の離散モデルの推定結果の評価を目的として，推定のロバスト性という観点から信頼のおける評価指標を導出し，この指標の適切さを実験によって確認する．

†早稲田大学先進理工学部

2. モデルの評価

標本空間 $\mathcal{X}$ に属する 1 つのサンプル $d\in\mathcal{X}$ が観測されたとき，この経験分布を

$$\delta_d(x)=\begin{cases}1 & (x=d)\\ 0 & (x\neq d)\end{cases} \tag{2}$$

と定義する．ある分布 p に従って観測される N 個のデータ $x_{1:N}=\{x_1,\ldots,x_N\}$ の経験分布 $\tilde{p}$ は，δ_d を用いることで

$$\tilde{p}(x)=\sum_{n=1}^{N}\frac{1}{N}\delta_{x_n}(x) \tag{3}$$

と表すことができる．定義域を Θ とするパラメタ θ で記述されるモデル $q(x;\theta)$ を考えた時，データに基づく θ の推定とは，経験分布 $\tilde{p}(x)$ をうまく再現できる θ を選ぶことである．モデルを評価する際には，推定結果が分布をどの程度再現しているか，推定結果がどの程度安定しているか，などが重要な指標となる．以下では前者を評価するための指標として情報量，後者の指標として SC，GES を紹介し，さらに一般に用いられるモデル選択規準について触れる．

2.1 情報量

モデル q の良さを評価するための考え方として，ある分布 p に対する近さを表す情報量 $\mathrm{D}(p,q)$ が重要な役割を果たす．

代表的な情報量としては KL 情報量

$$\mathrm{D}_{\mathrm{KL}}(p,q)=\sum_{x\in\mathcal{X}}p(x)\log\frac{p(x)}{q(x)} \tag{4}$$

が挙げられる．KL 情報量は p と q に関して非対称であり，$p(x)>0$，$q(x)=0$ において発散してしまうなどの特徴を持っている．また p を経験分布 $\tilde{p}$ とした時に

$$\hat{q}=\operatorname*{argmin}_{q}\mathrm{D}_{\mathrm{KL}}(\tilde{p},q) \tag{5}$$

となるような推定モデル $\hat{q}$ が最尤推定モデルとなることも広く知られている．

KL 情報量の他にも様々な情報量があるが，特に Bregman 情報量 [11, 4] に属する情報量の中で KL 情報量の 1 つの一般化となる β-情報量

$$\mathrm{D}_{\beta}(p,q)=\sum_{x\in\mathcal{X}}\left\{\frac{p(x)^{\beta+1}}{\beta(\beta+1)}-\frac{p(x)q(x)^{\beta}}{\beta}+\frac{q(x)^{\beta+1}}{\beta+1}\right\} \tag{6}$$

は適切な β の値を選ぶことによって，次節で述べる推定のロバスト性を実現することができる [5, 3] .

ここに示した2つの情報量はどちらも $\mathrm{D}(p,q) \geq 0$ であり，小さい値を示す程 q が p に近いことを表す．本稿では分布 p が与えられた時のモデル q の良さを，$\mathrm{D}_{\mathrm{KL}}(p,q)$ の小ささで評価する．

2.2 ロバスト性

例えばオーバーフィットが問題視される状況では，データのわずかな違いによってモデルの推定結果が敏感に変化してしまう．データの違いに対して推定結果の揺らぎが少なければ推定結果の信頼性は高いと考えられ，この性質のことをロバスト性と呼ぶ．ロバスト性を評価するためによく使われる考え方としては IF(Influence Function) や GES(Gross-Error Sensitivity) が挙げられる [6]．本稿ではモデルをデータで評価するための指標としてこれらの考え方を用いる．

N 個のデータで構成される経験分布 $\tilde{p}$ が，未知のサンプル d が加わることによって $\tilde{p}_d$ へ変化したとすると，式 (2) と (3) を用いることで $\tilde{p}_d$ は

$$\tilde{p}_d(x) = \frac{N\tilde{p}(x) + \delta_d(x)}{N+1} \tag{7}$$

と表せる．以下では分布 $\tilde{p}$, $\tilde{p}_d$ に基づいて推定されたモデルを $\hat{q}$, $\hat{q}_d$ とし，それぞれの推定パラメタを $\hat{\theta}$, $\hat{\theta}_d$ と表す．すると $x_{1:N}$ に，未知のサンプル d が加わることによって推定量が変化する度合いは

$$\mathrm{SC}(d) = (N+1)(\hat{\theta} - \hat{\theta}_d) \tag{8}$$

のように表すことができる．式 (8) で表される SC(Sensitivity Curve)[6] は経験分布に摂動が入った時の推定結果への影響の大きさとして解釈することができ，この値の絶対値が巨大になる推定法ほどそのデータ d に対するオーバーフィットの影響が大きいことになる．この考え方に基づいて，以下では経験分布に基づく GES を

$$\gamma = \max_{d \in \mathcal{X}} ||\mathrm{SC}(d)|| \tag{9}$$

のように定める．ただし $||\cdot||$ は2つのパラメタベクトルを比較するための適切なノルムで，例えば単純な比較としては L2 ノルム，L1 ノルムなどを使うことができる．ここで γ は分布が最悪な歪み方をした時のオーバーフィットの影響の大きさを表現していることになる．つまり式 (9) が小さい程，分布の歪みに対して推定量が変化しにくく，高いロバスト性を持つことになる．

2.3 一般的な評価手法とその問題点

例えば一般の情報量に基づくモデル $q(x;\theta)$ のパラメタの推定では，式 (4) や (6) を用いて

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \mathrm{D}(\tilde{p}, q(\theta)) \tag{10}$$

を実現することによって，対応する推定パラメタ $\hat{\theta}$ を求めることができる．一方本当に興味があるのは真の分布との距離を KL 情報量で測った $\mathrm{D}_{\mathrm{KL}}(p, q(\hat{\theta}))$ の小ささであるが，p が未知の状況ではこのような評価を直接行うことは不可能となる．一般的なモデル選択の手法では，経験分布 $\tilde{p}$ が真の分布 p の周囲でどのようにばらつくかを考慮することで，真の分布 p のどの程度良い近似となっているかを推定し，それを比較している．

例えば一般的な情報量規準は，

$$\mathrm{D}^{\mathrm{IC}} = \mathrm{D}_{\mathrm{KL}}(\tilde{p}, \hat{q}) + \text{バイアス} \tag{11}$$

の小ささで平均的な $\hat{q}$ の良さを評価していると解釈することができる．ここでバイアスは，

$$\text{バイアス} = \mathrm{E}[\mathrm{D}_{\mathrm{KL}}(p, \hat{q}) - \mathrm{D}_{\mathrm{KL}}(\tilde{p}, \hat{q})] \tag{12}$$

の意味であり，様々な経験分布を考えた時の情報量の平均的な差を表している．AIC[2] や GIC[9] などでは，サンプル数が充分多い時の $\hat{q}$ のばらつき方の漸近的な性質を利用して，このバイアス項を導出している．しかしこのような情報量規準では，サンプル数が少ない時には漸近論が破綻し，適切な評価を行うことができない．

また，モデルの評価方法としてクロスバリデーションやブートストラップなどのリサンプリング法も用いられる．例えば one-leave-out クロスバリデーションでは N 個のデータ $x_{1:N}$ の中から n 番目のデータ x_n を抜いたものを用いて推定したモデル $\hat{q}_{-n}$ を δ_{x_n} で評価するということを繰り返し，

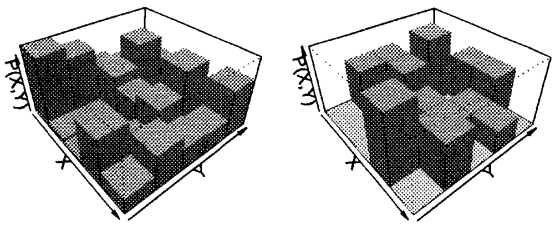
$$\mathrm{D}^{\mathrm{CV}} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{D}_{\mathrm{KL}}(\delta_{x_n}, \hat{q}_{-n}) \tag{13}$$

というような評価値の小ささでモデルの良さを評価する．このようなリサンプリング法では漸近的な性質を利用せずに，データに基づいて経験分布のばらつきによる $\hat{q}$ の平均的な良さを測ることができる．しかしこうしたリサンプリング法でも，サンプルが少ない時にはうまく評価できない場合がある．これはある状態 x に着目した時に，真の分布が $p(x) > 0$ であるにも関わらず経験分布が $\tilde{p}(x) = 0$ となってしまう，サンプリングゼロ [1] の状況が頻出することで，リサンプリングによって状態 x が本来生み出すばらつきを再現できないためである．

3. 最悪値評価に基づく評価

情報量規準やリサンプリング法では経験分布のばらつきを考慮した時の推定結果 $\hat{q}$ の平均的な良さを測っているが，少数サンプルの時にはこれらの指標は適していない．一方，サンプルが少数であるような状況でも，ロバスト性を測る指標を用いるとモデルの安定性の議論は行うことができる．以下では少数サンプルへのオーバーフィットが深刻となるような状況でロバストなモデルを選択することを目的として，GES に基づく評価指標を導入する．

式 (8), (9) に代表されるようなロバスト性の評価指標を導入することで，オーバーフィットの影響を論じることが可能になる．式 (7) から分かるように，データ数 N が十分に大きくなるにつれて $\tilde{p}_d$ は $\tilde{p}$ に近づいて行き，その推定結果 $\hat{q}_d = q(\hat{\theta}_d)$ も $\hat{q} = q(\hat{\theta})$ からの乖離が少なくなる．一方で N が少ない場合には，逆に $\hat{q}_d$ の $\hat{q}$ からの乖離が大きくなる傾向にある．ここでデータ $x_{1:N}$ に基



(a) 真の分布 p　　(b) 経験分布 $\tilde{p}$ の例

図 1: 実験で用いた分布．実際には $p(x) > 0$ の部分であっても $\tilde{p}(x) = 0$ となっているセルがいくつか存在する．

づいて推定したモデルが仮想的なサンプル d によって変化し得る度合いを情報量 $\mathrm{D_{KL}}(\hat{q}, \hat{q}_d)$ で測ることを考える．もし $\hat{q}_d$ が $\hat{q}$ から大きく乖離する場合，推定方法や推定モデルはデータに大きく依存しているという意味で信頼性が低いと言える．そこで本稿では評価指標として次のような量を提案する．

$$\mathrm{D}^{\gamma} = \mathrm{D_{KL}}(\tilde{p}, \hat{q}) + \max_{d \in \mathcal{X}} \mathrm{D_{KL}}(\hat{q}, \hat{q}_d) \qquad (14)$$

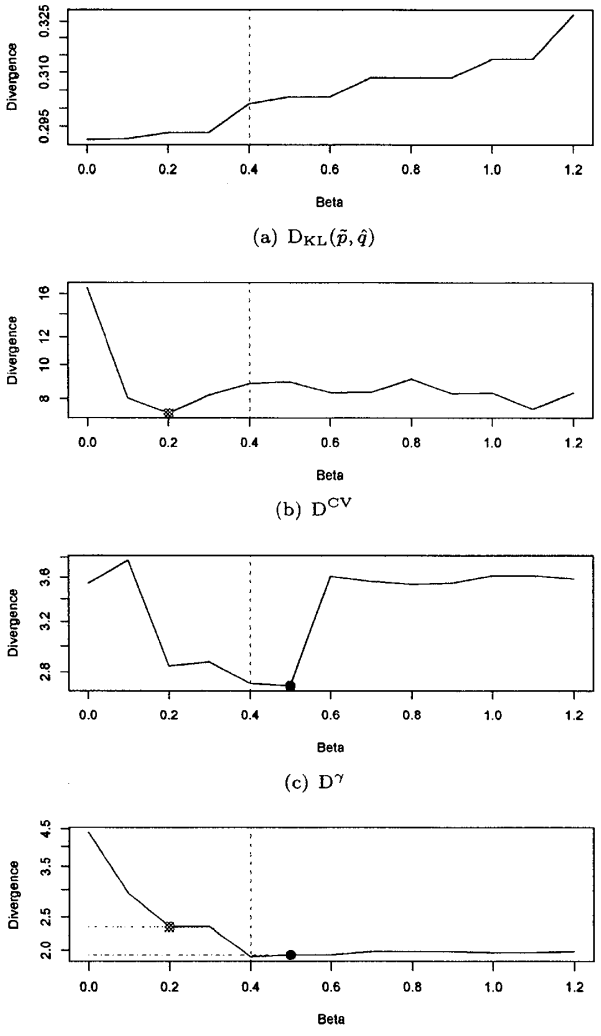
式 (14) の第 2 項は式 (9) に対応している項であり，D^{γ} を小さくする $\hat{q}$ は，情報量の意味で経験分布をうまく再現できていて，オーバーフィットも少ないと解釈できる．

4. 評価実験

提案する評価方法の適切さを見るために，式 (1) に示した潜在クラスモデルを異なる方法によって推定し，その推定結果の中から適切なものを提案指標によって選択する実験を行った．

それぞれ 5 状態を持つ離散 2 変数の同時確率 p（図 1(a) 参照）を用意し，そこから $N = 25$ 個のサンプルを生成することで，図 1(b) に示すような経験分布 $\tilde{p}$ を用意した．図から分かるようにサンプル数が少ない状況では，真の分布において $p(x) > 0$ であるにも関わらず $\tilde{p}(x) = 0$ となる部分が頻出する．そのため，クロスバリデーションなどではモデルの推定結果を適切に評価できない可能性がある．この経験分布に基づいて式 (1) に示した潜在クラスモデル ($K = 2$) の推定を，式 (6) に示した β 情報量 ($\beta = \{0, 0.1, \ldots, 1.2\}$) に基づく EM アルゴリズム [4] によって行った．なお，通常の EM アルゴリズムによる推定結果が，$\beta = 0$ における推定となっている．

推定に用いた β の値と推定結果の関係を示したのが図 2 である．図 2(a) を見ると $\beta = 0$ の時に $\tilde{p}$ に最も近づき，大きい β の時には，経験分布 $\tilde{p}$ からは乖離することが分かる．しかし，経験分布に対して良い結果を示す $\beta = 0$ の推定結果を用いてしまうと，真の分布に対する良いモデルにはなっていないことが図 2(d) と比較することで確認できる．これはサンプル数が少ないことでオーバーフィッティングが深刻となる典型的な例と言える．ここで式 (13) によって計算されるクロスバリデーション値を用いると図 2(b) より $\beta = 0.2$ の時の結果が良いと判断できる．一方，式 (14) に示した最悪値評価を用いると，図 2(c) より $\beta = 0.45$ が選ばれる．実際に推定



(a) $\mathrm{D_{KL}}(\tilde{p}, \hat{q})$

(b) D^{CV}

(c) D^{γ}

(d) $\mathrm{D_{KL}}(p, \hat{q})$

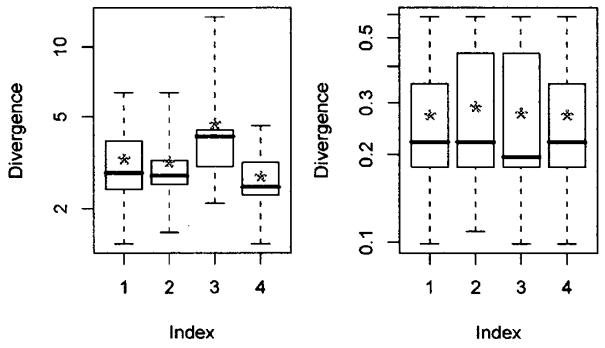
図 2: 評価値と β の関係．$\mathrm{D_{KL}}(p, \hat{q})$ を最小にする β を破線で，D^{CV} で選択される β を ■ で，D^{γ} で選択される β を ● で示している．

結果が真の分布 p に対してどのようになっているかを示す $\mathrm{D_{KL}}(p, \hat{q})$ を見てみると，この実験状況においてはクロスバリデーションでは選ぶことができないような良い推定結果を提案指標によって選択できることが分かる．

図 1(a) に示した p からデータを生成し直すことで同様の実験を 10 回繰り返し，実際にそれぞれの経験分布から推定したモデルを，表 1 で示す指標で選択した結果が図 3 である．図 3(a) を見ると，クロスバリデーション（指標 1）では選ぶことのできなかった $\mathrm{D_{KL}}(p, \hat{q})$ を小さくする推定結果を，提案指標（指標 4）を用いることで選ぶことができていることが分かる．また，クロスバリデーションを行う際に平均値評価を行う代わりに最悪値評価を行った場合（指標 2）や，GES に基づく評価を行う際に $\tilde{p}$ から $\hat{q}_d$ へ直接測った KL 情報量の最悪値を用いた場合（指標 3）には CV よりも悪くなってしまう場合があることも結果より確認できる．このことから，サンプル数が少ない時のモデル推定方法の選択に D^{γ} を

表 1: 実験で用いた選択指標

指標 1:	$\frac{1}{N}\sum_{n=1}^{N} \mathrm{D}_{\mathrm{KL}}(\delta_{x_n}, \hat{q}_{-n}) \quad (= \mathrm{D}^{\mathrm{CV}})$
指標 2:	$\max_{n} \mathrm{D}_{\mathrm{KL}}(\delta_{x_n}, \hat{q}_{-n})$
指標 3:	$\max_{d \in \mathcal{X}} \mathrm{D}_{\mathrm{KL}}(\tilde{p}, \hat{q}_d)$
指標 4:	$\mathrm{D}_{\mathrm{KL}}(\tilde{p}, \hat{q}) + \max_{d \in \mathcal{X}} \mathrm{D}_{\mathrm{KL}}(\hat{q}, \hat{q}_d) \quad (= \mathrm{D}^{\gamma})$



(a) $N = 25$ の時　　(b) $N = 250$ の時

図 3: 各選択指標によって選ばれたモデルの $\mathrm{D}_{\mathrm{KL}}(p, \hat{q})$. ボックスプロットは最大値，最小値，四分位値を表し，* は平均値を表している．

用いることでより良いモデルの選択が可能なことが確認できる．

なお，クロスバリデーションが適切に働くようなサンプル数がある程度多い状況では，これらの指標による選択結果の違いはあまり無くなってくることが確認できる (図 3(b) 参照)．

5. まとめ

本稿ではデータが少数の時に離散モデルの推定方法の適切な評価を行うための指標として，ロバスト性の観点から評価規準を提案した．実験結果より，データが少数であっても推定方法の評価を適切に行えることが確認できた．

また，現時点では $\hat{q}_d$ を得るために $\tilde{p}_d$ に基づく推定を全ての $d \in \mathcal{X}$ に対して実際に行っているため，クロスバリデーションと同じように計算コストが高い評価法となっている．この計算量の高さを回避するためには，例えば $\mathrm{D}_{\mathrm{KL}}(\hat{q}, \hat{q}_d)$ を大きくしそうな標本空間の部分集合 $\mathcal{X}' \subset \mathcal{X}$ を考え，実際に $\hat{q}_d$ の推定を行う範囲を $d \in \mathcal{X}$ に限定するなどの工夫を行うことで，より実用的な評価指標になると考えている．

なお，本稿では提案指標を推定方法の評価のために提案しているが，サンプル数が少ない状況で行われる一般的なモデル選択に用いることも考えられる．

参考文献

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

[2] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[3] Y. Fujimoto and N. Murata. Robust estimation and for mixture of probability tables based on β-likelihood. In *the Sixth SIAM Conference on Data Mining*, pages 519–523, 2006.

[4] Y. Fujimoto and N. Murata. A modified EM algorithm for mixture models based on Bregman divergence. *Annals of the Institute of Statistical Mathematics*, 59(1):3–25, 2007.

[5] H. Fujisawa and S. Eguchi. Robust estimation in the normal mixture model. *Journal of Statistical Planning and Inference*, 136:3989–4011, 2005.

[6] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics*. Wiley, 1986.

[7] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of Uncertainity in Artificial Intelligence, UAI'99*, 1999.

[8] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196, 2001.

[9] S. Konishi and G. Kitagawa. Generalised information criteria in model selection. *Biometrika*, 83(4):875–890, 1996.

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

[11] N. Murata, T. Takenouchi, T. Kanamori, and S. Eguchi. Information geometry of U-boost and Bregman divergence. *Neural Computation*, 16:1437–1481, 2004.

[12] N. Ueda and R. Nakano. Deterministic annealing EM algorithm. *Neural Networks*, 11:271–282, 1998.

[13] M. Watanabe and K. Yamaguchi, editors. *The EM Algorithm and Related Statistical Models*. Marcel Dekker, Inc., 2004.