1I1-3

複雑な構造の経時データ解析におけるモデル選択についての考察

Discussion to longitudinal data analysis and model selection with complicated structured data

原田 奈弥[*1*2] 山下 和也[*2] 黄冬陽[*2*3] 本村 陽一[*2*3]
Nami Harada Kazuya Yamashita Toyo Ko Yoichi Motomura

[*1] 株式会社豊田自動織機 TOYOTA INDUSTRIES CORPORATION
[*2] 産業技術総合研究所 National Institute of Advanced Industrial Science and Technology
[*3] 東京工業大学 Tokyo Institute of Technology

When we analyze some data with pLSA (Probabilistic Latent Semantic Analysis) and decide its model structure, usually we use information criteria, such as AIC or BIC. ID-POS data observe longitudinally and so frequently with certain number of customers or items that we need some technical discussion about deciding its appropriate model. Because it is difficult to calculate each model's log likelihood with such data for its so complicated data structure. Our suggestion is that if we have too complicated structured data to decide its model with AIC or BIC, but we have some information about its structure in advance, how to understand the meaning of the data.

1. ID 付 POS データとその活用

POS データとは、スーパーや百貨店などにおいて、商品が買われた店舗や日時、値段を示すデータである。個人を識別する ID が付いた、ID 付 POS データにすることで、個々の購入者が、いつ、どこで、何をいくらで買ったかを知ることができる。ID 付 POS データを活用して顧客の購買行動や、その時間軸上での変化を知ることはサービス向上に不可欠である。ここで重要な点は、顧客の購買行動の多様性を認めることである。ライフスタイルの多様化や経時的な変化に伴い、顧客の購買行動のパターンも多様であり、それが時間と共に変化していくと考えるのが自然である。ID 付 POS データの分析においては、顧客の購買行動の多様性と、時間に伴う変化を考慮することが不可欠である。

また、ID 付 POS データは、顧客数や売り上げに比例して膨大なデータとなる。データの大きさも大きく、顧客へのアンケートデータなども統合すると変数の項目も多く、複雑な構造をしていることが考えられる。分析を行う上ではビッグデータ活用のための専門的な技術や知識が必要になる。ID-POS データの分析には、pLSA(Probabilistic Latent Semantic Analysis)を用いた事例が多く提案されている。 [石垣, 2010]では、pLSA を用いて商品と顧客のクラスタリングを行い、どのような商品がどのような顧客に購入やされやすいのか、百貨店における顧客クラスター別の売れ筋について分析を行った。[原田, 2016]では、時間や季節の変化に伴う顧客の購買行動の変化について、pLSA を用いてモデリングを行った。兵庫県を主な事業エリアに約 150 店舗を展開する総合スーパーの、2009 年 9 月から 2010 年の 8 月までの 1 年間の ID 付 POS データから商品と顧客について、月別での季節感あり/なしの、pLSA によるクラスタリングのモデリングを行った。

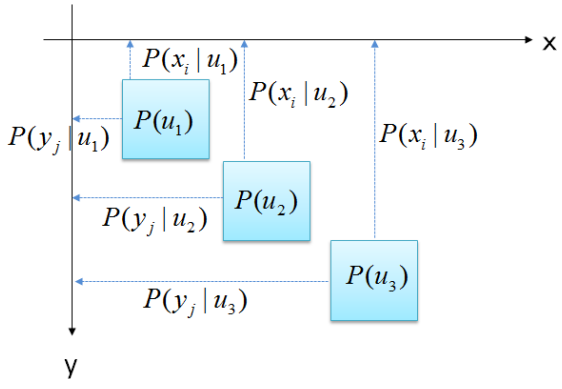

図 1:pLSA と潜在クラス

このように、ID-POS データの活用に pLSA は欠かせない。本研究では、ID 付 POS データに対して pLSA を用いた購買行動の統計的な分析を行う際、統計的確率モデリングとして記述するかについて考える。

2. 問題提起

2.1 pLSA と情報量規準

pLSA とは、Probabilistic Latent Semantic Analysis の略称で、確率的潜在意味解析とも呼ばれ、2 つの変数の変化について同時に意味付けを行う潜在クラスを仮定し、事後確率とベイズの公式からその潜在クラスを推測する、確率モデリングの手法である。

pLSA の潜在クラスとは、図 1 中のu_lで、これが二つの変数x_iとy_jに、データ解析上意味を与えるものであると考える。

モデルは潜在クラスのクラス数によって異なり、モデルの選択は通常、(1)の AIC や(2)の BIC などの情報量規準に基づいて行う。クラス数を k としたときの尤度 L とサンプルサイズ N を用いて求められる。すなわち、AIC や BIC などの情報量規準を用いてモデル選択をするということは、データに対して統計的に最も当てはまりのいいクラス数を選択している。

$$\mathrm{AIC} = -2\ln L + 2k \quad (1)$$

$$\mathrm{BIC} = -2\ln L + k\ln N \quad (2)$$

pLSA において尤度は、潜在クラスu_lを与えた下での、x_iやy_jが観測される確率から、(3)のように求められる。

連絡先:原田奈弥, 株式会社豊田自動織機/産業技術総合研究所, 135-0064 東京都江東区青海 2-3-26, 03-3599-8224, nami.harada@aist.go.jp

$$\mathrm{L} = \sum_i \sum_j N_{ij} \log p(x_i, y_j)$$
$$= \sum_i \sum_j \left[N_{ij} \log\{p(x_i) \sum_l p(y_j|u_l) p(u_l|x_i)\} \right] \quad (3)$$

2.2 pLSA とモデル選択

ID-POS データのように縦断的に観測されるビッグデータに対して pLSA を用いる際の問題の一つは、2 つの変数のクラス数が大きく異なることである。例えば、商品と販売日で pLSA を行った場合、商品の種類数に対して販売日の日数が非常に多い、長方形のデータがその典型である。図 1 において、販売日x_iと商品y_jに対して、ある潜在クラスu_lについて、あるP$(y_j|u_l)$が 0 となる可能性がある。このとき、(3)の尤度 L が妥当な値にならず、適切なモデルを選択できない可能性がある。

2.3 先行研究

(1) [Brants, 2005]の研究

$$p(x_i, y_j|\theta) = \prod_i \prod_j p(x_i, y_j|\theta) \quad (4)$$

[Brants, 2005]では、観測値の同時確率から尤度を求めるというアプローチを行った。その尤度関数 L は、(4)としている。

(2) [Ishigaki, 2010]の研究

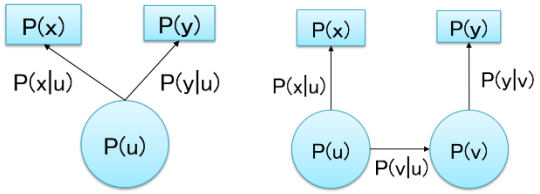

図 2: pLSA と PdLSI の潜在クラス

[Ishigaki, 2010]では、2 つの観測変数 x_i と y_j にそれぞれ、ことなる潜在クラスを仮定したモデリングを行い、パラメータを推定する、PdLSI という手法を提案している。

通常、pLSA は潜在クラスを図 2 の左図のように仮定しているが、[Ishigaki, 2010]の PdLSI では、図 2 の右図のようなモデルを仮定して推定を行っている。

$$p(x_i, y_j, u_k, v_l) = p(u_k) p(x_i|u_k) p(v_l|u_k) p(y_j|v_l) \quad (5)$$

PdLSI では、尤度は、図 2 の右図とベイズの公式より、(5)の確率を用いて、尤度や情報量基準を求める。

2.4 実データを用いた検証

先行研究として紹介したモデリングの手法や情報量規準について、実データを用いてその有用性を検証する。

(1) 用いたデータ

[原田, 2016]でも用いた、関西を主な事業エリアとする総合スーパーの ID 付 POS データと、[原田, 2016]の結果を元に検証を行った。

(2) 考え方

複雑な構造のデータへの対応の考え方として、[Ishigaki, 2010]の、x_iとy_jで潜在変数が多層になっている構造を想定した。2 つの変数のそれぞれでの周辺尤度が、クラス数が異なるモデルそれぞれでの推移を検証した。この考え方問題提起の 2.2 で述べた長方形のデータにおけるモデル選択に関する先行研究との関係について次の通りである。

[Brants, 2005]では、潜在クラスを条件付で与えた下で観測値が観測される確率、例えばP$(x_i|u_l)$が 0 となるときの問題点を指摘している。長方形のデータで 2 変数のそれぞれでの周辺尤度を求めることで、長方形の短辺の方の変数に関しては情報量規準から妥当なモデルを選択することができると考えられる。

[Ishigaki, 2010]のような構造を持ったデータでも、周辺尤度を求めることで、図 2 の右図における潜在クラスu_lとv_kのそれぞれのクラス数を求めることができる。

$$l_{pLSA|x} = \sum_i N_{i.} \log p(x_{i.})$$
$$= \sum_i \{N_{i.} \log \sum_l p(u_l) p(x_{i.}|u_l)\} \quad (6)$$

このような考え方により、[Watanabe, 2010]で指摘された非特異なデータについて、pLSA のクラス数に関するモデル選択を行う上での妥当性について、簡易的に検証できると考えた。周辺尤度は(6)から求めた。

この(6)の式と、[原田, 2016]において、APOSTOOL で推定した潜在クラスの確率P(u_i)と、潜在クラスを与えた下での観測変数の確率P$(x_i|u_i)$を用いて(6)の周辺尤度を求めた。

なお APOSTOOL とは、産業技術総合研究所が所有する pLSA を行う分析ツールである。

(3) 検証結果

[原田, 2016]での分析における、各クラスの情報量規準と、顧客(id)と商品(item)それぞれでの周辺尤度に基づく情報量規準の推移を図 3 に示す。なお、図 3 では便宜上、情報量規準の桁を丸めて表示している。

[原田, 2016]では、図 3 の上図で示す通り AIC と BIC がクラス数の変化に伴い、異なる挙動をしたため、BIC の改善が最も大きいクラス数 10 を採択している。

2.4(2)で述べた通り、先行研究の [Brants, 2005]や [Ishigaki, 2010]で指摘された問題があったとしても、周辺尤度に基づく情報量規準を求めれば、顧客(id)か商品(item)の少なくとも一方では、妥当なクラス数以上のモデルでは情報量規準の改善が止まり、正しいクラス数のモデルが選択できると考えられた。[原田, 2016]でも用いた実データで検証した結果、顧客(id)と商品(item)の両方で、図 3 の中図と下図の通り、クラス数が大きくなるほど当てはまりが悪くなり、妥当なクラス数の選択はできなかった。

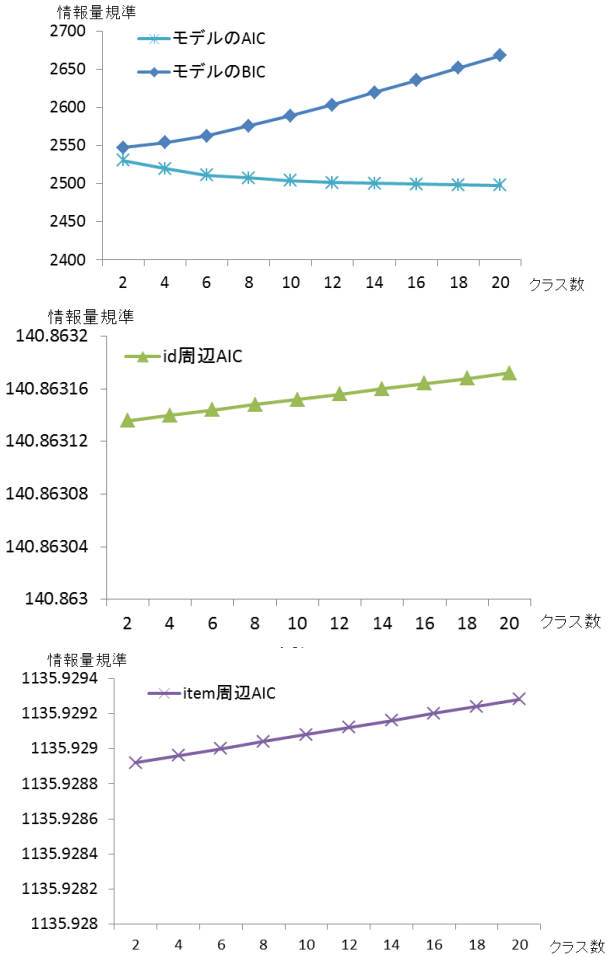

図 3: 情報量規準の推移

(5) 問題まとめ

以上のように、実データでは先行研究で提案された考え方では妥当なモデル選択ができなかった。この理由は、扱ったデータが多くの顧客の長期的な購買行動を記録したビッグデータであるため、変数間や変数内のクラス間の構造が先行研究で提案された考え方よりも更に複雑であることが考えられる。複雑な構造を持っているため、モデル間の尤度の比較による探索的なモデル選択ができない場合が考えられる。

3. 提案

2 で示したとおり、ビッグデータを扱う上では、変数間や、変数内のクラス間の構造が複雑なため、情報量規準に基づいた最も当てはまりのいいモデルの探索が難しい場合がある。このような場合でも、データや、データが観測される現場に関する既知の情報に基づき、pLSA によるクラスタリングを行い、データから知識や情報を得る方法を本研究にて提案する。

3.1 提案の詳細

過去の研究・調査や、データが観測される現場における固有の知識から既に妥当な潜在クラスの数が分かっている場合、変数のクラス内での着目すべき潜在クラスを、当てはまりのよさから選び出すことを考えた。モデル全体の尤度に対する、各クラスでの尤度の比を求めることでそのクラスを明らかにする。

$$L_{all} = \sum_i \sum_j \left[N_{ij} \log\{p(x_i) \sum_l p(y_j|u_l) p(u_l|x_i)\} \right] \quad (7)$$

$$L_{u_l|x} = \sum_j N_{\cdot j} \log p(x_{ij|u_l}) \quad (8)$$

$$\mathrm{LLR}(u_{l|x}) = \frac{L_{u_l|x}}{L_{all}} \quad (9)$$

(9)の、変数x_iの各潜在クラス間での当てはまりのよさを比較できる指標LLR($u_{l|x}$)を本研究では提案する。LLR とは、Log Likelihood Ratio の頭文字である。観測値をクラスタリングする潜在クラスが、どのクラスで当てはまりがいいかを、LLR($u_{l|x}$)の値の相対的な比較から知ることができる。この指標を用いることで、pLSA の潜在クラスの構造があらかじめ分かっているデータについて、着目するべきクラスを知ることができる。

4. 考察

提案した(9)の指標について、2.4 でも用いた実データを用いて、有用性を検証する。これは、データの複雑な構造について、あるデータの中の変数x_iやy_jの、あるモデルについての当てはまりのよさが、潜在クラス間で変わっていることを想定している。発表当日、その結果を示し、考察を行う。

5. まとめ

5.1 提案内容のまとめ

ビッグデータは、単にデータの容量が大きくなるだけでなく、変数間や変数内のクラス間の構造が複雑になるため、このようなデータに対して妥当なモデリングを行うことは難しい。そこで、データのモデル構造が以前の研究などから既知の場合について、データから知識や情報を得る方法を、本研究では提案した。

5.2 今後の課題

ビッグデータに対してモデリングを行う上での課題を 2 点考えている。

1 点目は、先行研究でも指摘されたような、階層的な構造や、観測変数のクラス間で異質な構造のような、複雑な構造のデータについて、探索的なモデリングを行うことができる、情報量規準などの指標の検討である。 [Watanabe, 2010]では、特異な構造のデータでのモデル選択のための情報量規準として、WAIC を提案している。WAIC はベイズ汎化誤差にも漸近的に一致する情報量基準で、AIC では妥当なモデル選択ができない特異構造のデータでも、ベイズ推定の考え方に基づいた妥当なモデル選択をすることができると述べられている。データの構造が複雑になることで特異な構造になることが考えられるので、そこでいかに妥当なモデル選択を行うのかを考える必要がある。

2 点目の問題は、実用的な指標の検討である。データ解析の結果を基に小売業などで業務改善を行う上で、どのような改善が必要かを、人が理解できることが必要である。人工知能の学習や出力について意味のある解釈が可能であるだけなく、誤りがなく、情報量を落とさない範囲で、簡易であるべきではないか。データが、階層や潜在クラスについて、より複雑な構造をもっていたとしても、出力や可視化を工夫することで、「人と協調する AI」が実現できると考えている。

6. 謝辞

本研究は NEDO 委託事業「人間と相互理解できる次世代人工知能技術の研究開発」の支援を受けたものです。

参考文献

[Brants, 2005] Thorsten Brants: Test Data Likelihood for PLSA Models, Information Retrieval, 8, 181-196, 2005.

[Ishigaki, 2010] Tsukasa Ishigaki, Takeshi Takenaka and Yoichi Motomura: Category Mining by Heterogeneous Data Fusion Using PdLSI Model in a Retail Service, IEEE International Conference on Data Mining , 2010.

[Watanabe, 2010] Sumio Watanabe: Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory, Journal of Machine Learning Research 11 (2010) 3571-3594, 2010

[稲垣, 2003] 稲垣宣生: 数学シリーズ 数理統計学 改訂版, 裳華房, 2003

[石垣, 2010] 石垣司, 竹中毅, 本村陽一: 百貨店 ID 付き POS データからのカテゴリ別状況依存的変数間関係の自動抽出法, オペレーションズ・リサーチ, 56, 2, 2010

[白旗, 1992] 白旗慎吾: 統計解析入門, 共立出版株式会社, 1992

[原田, 2016] 原田奈弥, 山下和也, 本村陽一: ID 付 POS データによる購買行動の季節変化の分析と視覚化, 人工知能学会 社会における AI 研究会 27 回研究会,2016.