N-030

ベイズ推定を利用した対応分析による提案型オンラインデーティングサービスの分析
Analysis of a recommendation-based online dating service by correspondence analysis using Bayesian estimation

北原 洋一
Yoichi Kitahara
株式会社リブセンス
Livesence Inc.

1. はじめに

1.1 背景

近年、オンラインデーティングサービスは、新たな出会いの場として注目されている。現在主流のオンラインデーティングサービスでは、ユーザーが希望の条件を指定して検索することで、交際候補を見つける仕組みとなっている。お互い相手のプロフィールが好みに合うことが交流機会の条件となっているところに特徴がある。希望の条件に合った相手を見つけられるといったメリットがある一方で、プロフィールへの過度な依存が問題となることがある。

knew(https://knew.jp)は、データに基づいて選定した交際候補とビデオチャットを行うオンラインデーティングサービスである。検索機能を備えておらず、プロフィールでの選別をほとんどせずに、サービス提供者が紹介した人と日程調整を行って5分間のビデオチャットを行う。ビデオチャット後にお互い連絡先交換を了承すると、メッセージを通じて交流を深めることができる。

knewではサービス提供者が相手を紹介するためデータ活用の重要性が高い。サービス登録時に年齢等の属性や相手の希望条件の情報を収集し、ビデオチャット後に満足度や特徴などの相手への評価情報と次回紹介の希望を収集する。交際候補はこれらの情報に基づいて選定される。これらの収集データは、入力しやすさを考慮し選択式回答となっているものが多い。

1.2 研究目的

knewでは選択式回答を多用していることからカテゴリカルデータが多く対応分析も利用されている。従来の対応分析ではカテゴリーの座標を点で示している。そのためデータ量に応じた信頼性を定量化しにくいことから解釈がしにくいことがあった。

本研究の目的は、ベイズ推定を利用して対応分析結果の信頼性を定量化することで対応分析の解釈を容易にすることである。

2. ベイズ推定を利用した対応分析

2.1 対応分析

対応分析とは、度数の二元分割表から行と列のスコアを算出し、行と列の関連性を把握させやすくする手法である。主に、探索的分析に利用される。通常は固有値の大きい順に各カテゴリーの2軸のスコアをプロットして可視化し、関連性を直感的に把握するのに使われる。対応分析で示される距離はカイ二乗距離となっており、プロットされたカテゴリー間の距離が短いほど関連があると解釈される。

対応分析のスコア導出過程の定式化はいくつか知られているが同じ結果を算出する。スコアの分散が最大となるよう最適化したもの[1, 2]、相関比最大化や相関係数最大化によるもの[3]がある。本研究では[1, 2]の方法を使う。

行と列のスコアは次のように算出される。度数の分割表をXとし、度数の総和をNとする。さらに$Z = \frac{1}{N}X$に関する行の総和のベクトルをr、列の総和のベクトルをcとし、$D_r = diag(r)$、$D_c = diag(c)$とした対角行列を使う。ここで、$diag()$はベクトルを対角行列にする関数とする。このとき、標準化残差行列

$$S = D_r^{-\frac{1}{2}}\left(Z - rc^{\mathrm{T}}\right)D_c^{-\frac{1}{2}} \quad (1)$$

を特異値分解

$$S = PDQ \quad (2)$$

する。ここで、特異値ベクトルをdとし$D = diag(d)$である。対応分析の行スコアRと列スコアCは

$$R = D_r^{-\frac{1}{2}}P\mathrm{D} \quad (3)$$

$$C = D_c^{-\frac{1}{2}}Q\mathrm{D} \quad (4)$$

として算出できる。可視化を行う場合は、1番目と2番目に大きい特異値に対応した列を二次元にプロットする。

対応分析の具体例を示す。図1は、knew女性ユーザーの「自分の学歴」と「希望する相手の学歴」の対応分析の結果を示したものである。

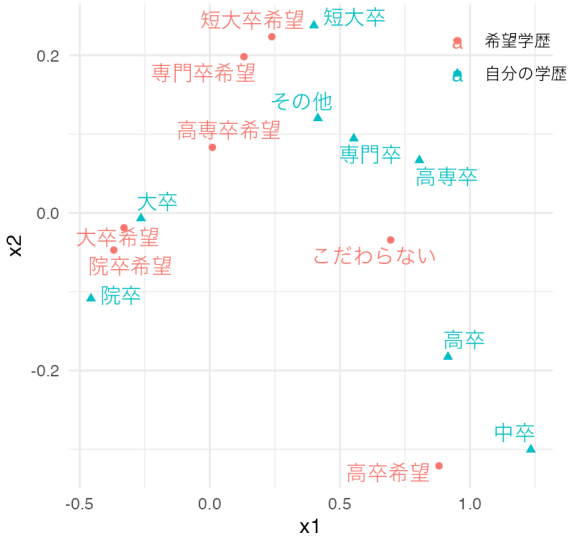

図1 女性ユーザーの自分の学歴と希望する相手の学歴の対応分析（青三角：自分の学歴、赤丸：希望する相手の学歴）

図1ではカテゴリー間の距離を直感的に把握できるものの、結果が点座標のみとなっているため座標の信頼性が不明である。例えば、図1のデータでは、大卒は多いものの高卒、中卒は少ない。また、院卒希望、大卒希望は多いものの、高卒希望は少ない。そのため、大卒と、大卒希望もしくは院卒希望との関連は

高いと確度高く言えるはずである。同様に、高卒および中卒と高卒希望との関連は高いものの確度は低いはずである。しかし、図1を見るだけではそのような解釈をするのは難しい。元の分割表を確認すれば、度数が非常に大きいカテゴリーと小さいカテゴリーとでは座標の信頼性に違いがあることは定性的に判断できる。しかし、どの程度信頼性が異なるかはわからない。

2.2 ベイズ推定

ベイズ推定はベイズの定理を利用した推定方法である。近年は計算機の発展と確率的プログラミング言語の登場で複雑なモデルのベイズ推論が容易になったため、柔軟な分析を可能にする手法として利用されている。本研究でも、ベイズ推定には確率的プログラミング言語Stan[4]を利用している。

ベイズ推定を利用した対応分析では、二元分割表の度数をベイズ推定し、得られた事後分布のサンプルを使って対応分析の行と列のスコアを計算する。度数のモデルは扱う問題に応じて適切なものを用いる。スコアの計算方法は、2.1節で示したものを用いる。

二元分割表のモデルは、総度数や行和、列和が与えられるかに応じて4つある[3]。総和が与えられる場合、行和もしくは列和のいずれかが与えられる場合、行和と列和のいずれもが与えられる場合、総度数も与えられない場合である。後者二つは特殊な問題設定のときに使われるものなので、前者二つを説明する。

特定の属性をもつユーザーが何らかの選択をするケースにおいて、属性と選択とに関連があるかを調べたいことが多い。このようなケースは、行和あるいは列和が与えられた場合のモデルが適している。

行和が与えられる場合は、行ごとに行和と各列の生起確率から各列の度数を推定するモデルを使う。i番目の行の度数ベクトルX_iは、パラメータをi番目の行の行和N_iとベクトルθ_iとする多項分布$Multinomial(\theta_i, N_i)$を使って

$$X_i \sim Multinomial(\theta_i, N_i) \tag{5}$$

と表現できる。ここで列サイズがmのとき

$$\sum_{j=1}^{m} \theta_{ij} = 1 \tag{6}$$

である。

扱う問題に応じてθ_iの事前分布を設定する。なるべく仮定を置かずに関連を調べたい場合は無情報事前分布を使う。しかし、何らかの事前知識を仮定しても関連が見られるかを調べる場合は事前分布を設定する。ディリクレ分布を使うことが多いと思われるが、列カテゴリーが順序尺度となっている場合にはガンマ分布や正規分布を使うことも考えられる。knewの場合、年収情報が順序尺度かつ年収帯に応じた特徴的な分布となることがあるので、このような事前分布が有用なことがある。なお、行ごとに事前分布を設定することもできる。

特定ユーザーに生じた結果に関連があるかを調べることがある。このようなケースでは総度数が与えられる場合のモデルが適している。knewの場合、あるユーザーの自分の評価と相手からの評価の関連を調べるケースがこのケースに該当する。

総度数が与えられる場合は、総度数と全てのセルの生起確率から各セルの度数を推定するモデルを使う。度数ベクトルXは、パラメータを総度数Nとベクトルθとする多項分布を使って

$$X \sim Multinomial(\theta, N) \tag{7}$$

と表現できる。ここで行サイズがn、列サイズがmのとき

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \theta_{ij} = 1 \tag{8}$$

である。

2.3 分析結果

本節ではベイズ推定を利用した対応分析による分析結果の一例を示す。

図2は、図1と同じデータを使ってベイズ推定を利用した対応分析をした結果である。モデルは行和が与えられたモデルを用い、事前分布は無情報とした。点は事後分布の中央値、エラーバーは5%と95%の分位数を結んだものである。バーが短いほど信頼性が高いことがわかる。

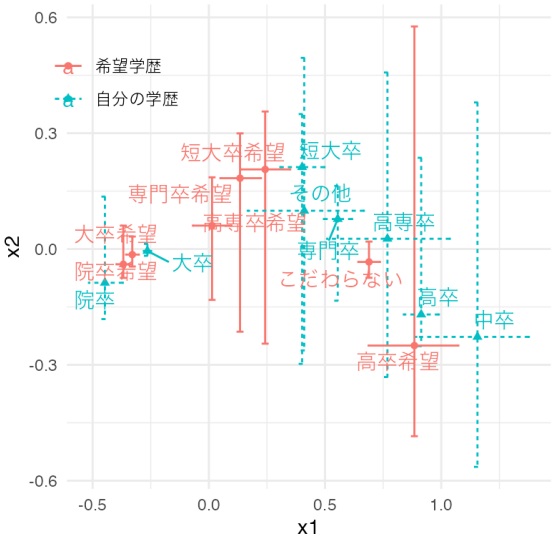

図2 女性ユーザーの自分の学歴と希望する相手の学歴のベイズ推定を利用した対応分析（青三角破線：自分の学歴、赤丸実線：希望する相手の学歴）

図2では座標の信頼性を確認できる。例えば、大卒と、大卒希望もしくは院卒希望との関連は信頼性が高いことが確認できる。また、学歴に「こだわらない」の座標の信頼性は高い一方で、周辺に院卒、大卒以外の学歴カテゴリーが分布していることがわかる。このことから、院卒、大卒以外は学歴に「こだわらない」と関連がある可能性が考えられる。なお、エラーバーの長さは分割表のモデルによって異なることに留意が必要である。

3. まとめ

本研究では、二元分割表の度数をベイズ推定し、得られた事後分布を用いて対応分析を行うことで、行スコアと列スコアの信頼性を算出した。ベイズ推定を利用した対応分析の一例として、knewユーザーのデータを使って分析を行い信頼性を可視化できることを示した。

参考文献

[1] M. Greenacre, “Correspondence Analysis in Practice, 3nd ed.”, Chapman and Hall/CRC (2017).

[2] H. Abdi, M. Béra, “Correspondence Analysis”, Encyclopedia of Social Network Analysis and Mining, pp. 1–12 (2017).

[3] 宮川雅巳, “分割表の統計解析：二元表から多元表まで”, 朝倉書店 (2018)

[4] A. Gelman, D. Lee, and J. Guo, Stan: A Probabilistic Programming Language for Bayesian Inference and Optimization, Journal of Educational and Behavioral Statistics, vol. 40, no. 5, pp. 530–543 (2015)