1B-01

離散一般化ベータ分布を仮定した研究分野マッピングの導出

蔵川　圭† 孫　媛‡
国立情報学研究所†‡

はじめに

本研究は，エビデンスベースの研究開発戦略や政策立案に必要なツールとして，科研費の分野分類とトムソン・ロイター社の論文データベース Web of Science (WoS)の分野カテゴリのマッピングテーブルを構築する．マッピングテーブルは，科研費と WoS のデータベースを利用して共通要素である論文をカウントして作成した研究分野の対応関係を示す分割表に対し，ランク順分布の一つである離散一般化ベータ分布を非線形最小二乗法で当てはめて観測値から理論値を導出して作成する．我々のこれまでの研究から，この非線形最小二乗法の適用で解を得るために重要なパラメータ初期値の与え方によって解が導出できる場合とそうでない場合があることがわかった[1]．本報では，パラメータ初期値に着目し，実際のデータ処理を通して考察した結果を示す．

分割表

論文と対応づけられた科研費研究分野と WoS カテゴリの 2 種の分類に対しクロス集計して分割表を作成する（図 1）．分割表は，一論文に対し対応した研究分野の出現度数をそのまま 1 とする整数カウントと案分する分数カウントの 2 種類である．科研費研究分野は，用いたデータでは 4 系・10 分野・67 分科・284 細目からなり，系から細目に詳細化されるような階層構造になっている．一方，WoS カテゴリはフラットに 251 分類となっている．

離散一般化ベータ分布の当てはめ

科研費のある研究分野 B_i ごとに WoS カテゴリ S の度数 $f_{i1},\ldots,f_{ij},\ldots,f_{in}$ を降順に並べ替え，$f'_{i1} > \cdots > f'_{ij} > \cdots > f'_{in}$ となるよう順序を付けた WoS カテゴリ S' を定める．すなわち，WoS カテゴリ S のランク順の分布に対し理論的モデルを当てはめる．様々な自然現象の中で出現頻度

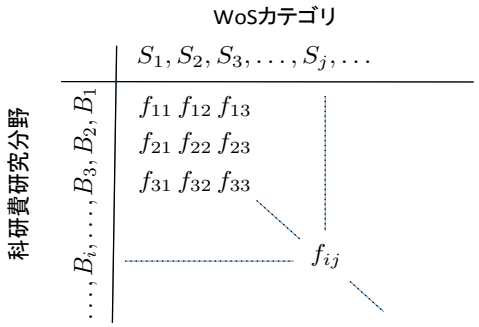

図 1 科研費研究分野と WoS カテゴリを二軸とする分割表

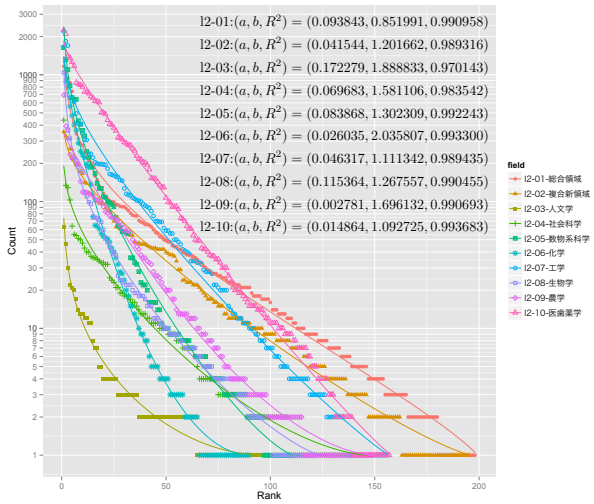

図 2 離散一般化ベータ分布の当てはめ例

をランク順に並べた分布は，log-log スケールにおいて直線になるという Zipf の法則への当てはまりを見る．観測数→∞のときに頻度→∞，ランク→∞であるようなスケールフリーの時であるが，我々の分布においてランクは有限であるため Zipf の法則へは当てはまらない．

スケールフリーの性質を持ちつつも，頻度→∞，ランク→Rとなるような制約のある特徴を表した分布として，離散一般化ベータ分布(discrete generalized beta distribution: DGBD) が提案されている[2][3]．この分布は，

$$f(r) = K\frac{(R-r+1)^b}{r^a}$$

で定義され，r はランク，R はランクの最大値，K は正規化定数，(a,b) はパラメータである．

Mapping subject categories based on discrete generalized beta distribution
† Kei Kurakawa, National Institute of Informatics
‡ Yuan Sun, National Institute of Informatics

分割表に対し，科研費の分野ごとに R の非線形最小二乗ソルバーnlmrt パッケージを用いてDGBD をフィッティングした一例を図 2 に示す．図では，科研費研究分野分類の 10 分野ごとに 251 の WoS カテゴリへのランク-頻度分布を同時にプロットした．科研費分野分類は，系・分野・分科・細目が包含される階層構造を持つため，得られた細目と WoS カテゴリの分割表の度数を集計することで上位のレベルの科研費分野分類と WoS カテゴリとの分割表を導出できる．得られたパラメータ値と決定係数 R^2 も記した．決定係数は，0.99~0.97 を得た．

パラメータ初期値による当てはめの成否

ここに図示した科研費研究分野 B_i 以外で，パラメータの初期値を$(a, b, \mathrm{K}) = (1,1,1)$と与えるのでは R の非線形最小二乗ソルバーnlmrt パッケージでは解決できないものが存在した．

一般に，非線形最小二乗法は，局所的に線形近似して残差が最小となる方向にパラメータ x を漸近更新して、推定値 $\hat{\mathrm{x}}$ を求める[4]．もっとも簡単な解法は，最急降下法(method of steepest descent)であり，他にはNewton 法，Gauss-Newton 法がある．モデル関数値と 1 階偏微分からヘシアン行列を推定する方法は準Newton 法と呼び，変形は，Gill-Murray 法，BFGS 法，Biggs 法などがある．また，Gauss-Newton 法の変形としては、Marquardt 法、Powell の最小二乗法，Powell のハイブリッド法などがある．

推定値を得られるかどうかは、解法の選択と初期値の与え方に依存する．基本的戦略は以下の２点に集約される．

- モデル関数の形から、解法を選択する
- 推定値の近傍がわからない場合は、数多くの初期値を用意する

ここでは初期値の与え方に着目して，パラメータ初期値$(a, b, \mathrm{K}) = (1,1,1)$で解が導出できなかった分科「心理学」を例に，幾つかのパラメータ初期値から解が導出できるかどうか実験を行った．パラメータ$\mathrm{a,b}$は0.1から1.0まで0.1刻みの 10 点づつ，パラメータKは1に固定して，合計 100 通りの組み合わせに対し非線形最小二乗法ソルバーnlmrt を適用した．その結果を図 3 に示す．83 通りのパラメータ初期値で推定値を得ることができ，17 通りで推定値を得ることができなかった．また，得られた推定値の分布を図 4 に示す．図の左はパラメータaの推定値のヒストグラムであり，図の右はパラメータbの推定値のヒストグラムである．

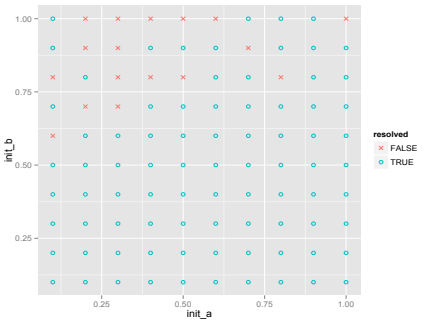

図 4 パラメータ初期値による解の導出結果

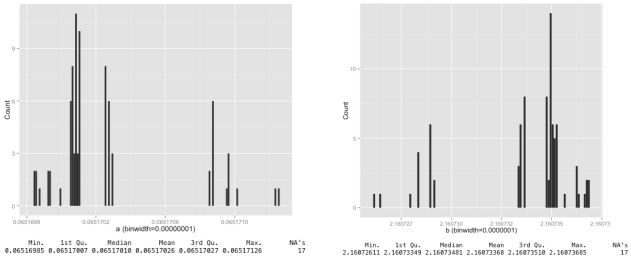

図 3 パラメータ推定値（左：a，右:b）

考察

100 通りのパラメータ初期値を与えた際に，推定値を導出できなかった部分は図の左上に偏在している．nlmrt では，ランク落ちなどの原因で推定値の得られない場合があり，計算過程における丸め誤差があることを考慮すると，推定値の導出できない初期値は近傍に偏在する可能性はある．

推定値の導出できたパラメータa,bそれぞれを見ると，必ずしも一致はしていないが，ほぼ等しい．しなしながら，その分布は解の平均に対して正規分布をなしてはいない．むしろ，平均を避けるように 2 極に分離して推定値が得られている．非線形最小二乗法における解の収束過程が漸近更新であるため，収束方向によって 2 極に分離したと考えられる．

まとめ

非線形最小二乗法においてパラメータ初期値を数多く用意することによって，パラメータ推定値を導出できることを示した．今後の展望として，残った事例についても実験を行う．

参考文献

[1] 蔵川圭，孫媛 〝レコードリンケージに基づく研究分野マッピングの導出〞 日本計算機統計学会第 28 回シンポジウム論文集，pp.183-186 (2014).

[2] Naumis, G.G., Cocho, G.: Tail universalities in rank distributions as an algebraic problem: The beta-like function. Phys. A Stat. Mech. its Appl. 387, 1, 84-96 (2008).

[3] Martínez-Mekler, G. et al.: Universality of rank-ordering distributions in the arts and sciences. PLoS One. 4, 3, e4791 (2009).

[4] 中川徹、小柳義夫：最小二乗法による実験データ解析、東京大学出版会、206 pages (1982).