東京学芸大学リポジトリ
Tokyo Gakugei University Repository

2PLモデルのパラメタ推定手法による語彙テストのパラメタ推定値の比較

メタデータ	言語: Japanese 出版者: 公開日: 2022-01-21 キーワード (Ja): キーワード (En): Parameter Estimation, Item Response Theory, Second Language Vocabulary 作成者: 江原,遥 メールアドレス: 所属:
URL	http://hdl.handle.net/2309/00173484

東京学芸大学紀要　自然科学系　73: 213－218，2021

2PLモデルのパラメタ推定手法による語彙テストのパラメタ推定値の比較

江　原　　遥*

技術科学分野

（2021年8月16日受理）

EHARA, Y.: Comparing the Estimated Values of the 2PL Model on a Vocabulary Test Dataset using Different Parameter Estimation Methods. Bull. Tokyo Gakugei Univ. Div. Nat. Sci., 73: 213–224. (2021)　ISSN 2434–9380

Abstract

Recently, interpreting the parameters of deep learning models has become an important research topic in deep learning studies, which have achieved remarkable predictive performance. The 2PL model, a model of the Item Response Theory (IRT), is a probabilistic model that is frequently used in the analysis of test-takers' abilities, questions' difficulties, and other aspects of testing. 2PL model is sometimes incorporated into deep learning models for improving its interpretability. However, there is a difference between the parameter estimation methods of the 2PL model and those commonly used for parameter estimation in deep learning models, in that the former is usually more robust and rigorous. If there exists a large difference in the estimates due to the parameter estimation methods, it is difficult to interpret the parameters. In this study, we compared the correlation between the estimates of the two parameter estimation methods using data from an actual vocabulary test result dataset. As a result of the experiment, we confirmed that the estimated values of both the ability and difficulty parameters showed high correlations between the two estimation methods.

Keywords: Parameter Estimation, Item Response Theory, Second Language Vocabulary

Department of Technology Sciences, Tokyo Gakugei University, 4-1-1 Nukuikita-machi, Koganei-shi, Tokyo 184-8501, Japan

要　旨

近年，顕著な予測性能を達成する深層学習モデルの研究においては，学習されたモデルのパラメタの解釈が課題になっている。項目反応理論における2PLモデルは，人間の能力や設問の難易度など，試験を対象にした分析において多用される確率モデルであり，深層学習モデルに組み入れて利用されることもある。しかし，2PLモデルのパラメタ推定法と，深層学習モデルのパラメタ推定に一般的に用いられるパラメタ推定法では，通常，前者の方が頑健・厳密であるなど，違いがある。パラメタ推定手法による推定値の違いが大きければ，パラメタの解釈も難しくなる。そこで，本研究では，実際の外国語学習者を対象にした語彙テストのデータを用いて，両パラメタ推定手法による推定値がどの程度一致するのかを比較した。実験の結果，能力パラメタや難易度などは，両推定手法で高い相関を示すことを確認した。

キーワード: パラメタ推定，項目反応理論，第二言語語彙

＊　東京学芸大学 技術・情報科学講座 技術科学分野（184-8501　東京都小金井市貫井北町 4-1-1）

1．はじめに

項目反応理論は，テストの分析に広く使われる確率モデルの総称である。被験者の設問に対する回答パターンを入力として，被験者の能力や設問（項目）の難しさ（困難度）をパラメタとして自動的に推定する。後に詳述する2PLモデルは，項目反応理論の一種である。

一方，近年の自然言語処理をはじめとする人工知能分野では，ニューラルネットワークを用いた深層学習モデルが盛んに利用されている。ニューラルネットワークは広範囲な確率モデルを記述することができるため，2PLモデルも，ニューラルネットワークで記述することが可能である。したがって，2PLモデルで想定している「能力パラメタ」や「困難度パラメタ」といったパラメタに相当するパラメタを持たせることで，2PLモデルの考え方を深層学習モデルに組み入れることが可能となっている。実際，このように2PLモデルの考え方を導入した深層学習モデルは過去に提案されている[1)]。一般的に，深層学習モデルはパラメタの値の解釈が困難であるといわれており，深層学習モデルのパラメタを2PLモデルの能力値や困難度といった考え方で解釈する事ができれば，説明性の高い深層学習モデルを構築できると注目されている。

このように，従来の2PLモデルと深層学習モデルは高い親和性を持つ一方で，そのパラメタ推定法は，大きく異なっている。一般に，項目反応理論では，被験者のパラメタと項目のパラメタを同時推定すると，被験者数を増やしたときに一致性が失われることが知られている。このため，被験者のパラメタについては積分消去して，項目パラメタのみをデータから推定するなど，一致性を考慮したパラメタ推定手法が用いられる[2)]。一方，深層学習モデルについては，通常，非常に多くのパラメタを持つため，確率的勾配をもとに，全パラメタを同時に更新していく手法が用いられる[3)]。しかし，深層学習モデルを2PLモデルの考え方を使って解釈しようとする際に，パラメタ推定手法の違いによって，本来の2PLモデルとはパラメタ推定値がかけ離れてしまうようでは，得られたパラメタの解釈が難しくなる。

本研究では，もっとも単純に，2PLモデルをニューラルネットワークとして実装しニューラルネットワークの標準的なパラメタ推定法によりパラメタ推定を行った場合のパラメタ値を，従来の項目反応理論による分析で標準的に使われるパラメタ推定法により得られたパラメタ値と比較する。対象とするテスト結果のデータには，公開されている英語の語彙テスト（単語テスト）の結果データを用いた。実験の結果，被験者のパラメタ，設問のパラメタともに，両推定手法で得られたパラメタ値がよく相関することが示された。

本研究の貢献は以下のとおりである。

（1）2PLモデルのパラメタ推定値をニューラルネットワークのパラメタ推定法と，項目反応理論で標準的に使われるパラメタ推定法で比較した。再現性のため，データセットには公開されている語彙テストの結果データを用いた。

（2）本研究による比較は，深層学習モデルに2PLモデルを組み入れてパラメタの値を解釈しようとする際，各パラメタの値がどの程度信頼できるのかの参考になる。

2．定式化

2PLモデルは，項目反応理論のモデルの１つであり，次の式で表される。被験者の数を J，設問（項目, item）の数を I とする。また，簡単のため，被験者の添字（index）と被験者，項目の添字と項目を同一視する。例えば，i 番目の項目についても，単に i と書くことにする。y_{ij} は，被験者 j が項目 i に正答するとき１，誤答であるとき０であるとする。試験結果データ $\{y_{ij} | i \in \{1, \ldots, I\}, j \in \{1, \ldots, J\}\}$ が与えられたときに，2PLモデルでは，被験者 j が項目 i に正答する確率を次の式でモデル化する。

$$P(y_{ij} = 1 | i, j) = \sigma(a_i(\theta_j - d_i)) \tag{1}$$

ここで，σ はロジスティックシグモイド関数であり，次の式で定義される。これは，$(0, 1)$ の範囲を動く単調増加関数であり，$\sigma(0) = 0.5$ である。任意の実数値を $(0, 1)$ の範囲に射影し，確率値として扱いたいときに用いられる。

$$\sigma(x) = \frac{1}{1 + \exp(-x)} \tag{2}$$

式１において，θ_j は能力パラメタ（ability parameter）と呼ばれ，被験者の能力を表すパラメタである。d_i は困難度パラメタ（difficulty parameter）と呼ばれ，項目の難しさを表すパラメタである。$a_i > 0$ は，通常，正の値を取り，識別力パラメタ（discrimination parameter）と呼ばれる。この項目の意味は後述する。

式1における能力値パラメタθ_jと困難度パラメタd_iに着目しよう。$\theta_j = d_i$であるとき，式1は，ちょうど0.5の値を取る。すなわち，能力値パラメタと困難度パラメタが等しいときは，被験者が項目に正答するか誤答するかわからない，と判定される。$a_i > 0$であるから，$\theta_j > d_i$であるとき，被験者が項目に正答する確率が，被験者が項目に誤答する確率より高いと判定される。また，$\theta_j < d_i$であるとき，被験者が項目に誤答する確率が，被験者が項目に正答する確率より高いと判定される。

ここで，識別力パラメタ（discrimination parameter）a_iを説明する。前述のように，$\theta_j - d_i$が大きいほど，被験者jが項目iに正答する確率が高くなる。$\theta_j - d_i > 0$である場合を考えよう。識別力パラメタは通常，正の値を取る。$\theta_j - d_i$が同じ値であっても，識別力パラメタの値が大きければ，被験者jは項目iに正答する確率が高くなる。逆に，識別力パラメタの値が小さければ，$\theta_j - d_i$が同じ値であっても，被験者jが項目iに正答する確率は（誤答する確率よりは高いものの）低くなり，被験者jが項目iに正答するとははっきり言いづらくなる。このように，識別力パラメタは，その項目が，能力値が高い被験者と能力値が低い被験者をどの程度はっきり分けるか（識別するか），という度合いを表している。直感的には，識別力パラメタの高さは，項目が良問である度合いを，正答確率の観点から表しているとも考えられる。識別力パラメタが低ければ，能力値の高い被験者でも項目に正答できなくなる確率が上がり，また，能力値の低い被験者が項目に正答してしまう確率も上がってしまう。そのような問題は，通常，良問であるとは言いづらいので，直感的には良問である度合いを正答確率の観点から表しているととらえられる。

式1が2PLと呼ばれるのは，各項目に対して困難度と識別力の2つのパラメタが存在するためである。すべての項目に対して，識別力を考慮せず，$a_i = 1$を仮定したモデルを1PLと呼ぶ。1PLは，Raschモデル[4]と等価である。式1では，a_iとθ_jの積の項，すなわち，項目パラメタと被験者パラメタの積の項が出てくることが見て取れるが，$a_i = 1$である1PLでは，こうした項は出てこない。

式1では，θ_j, a_i, d_iの3種類のパラメタがある。データに対して，この尤度関数を最大化するパラメタを探索する。

$$L(\mathbf{w}) = \prod_{i=1}^{I}\prod_{j=1}^{J} P(y_{ij}|i, j) \tag{3}$$

ただし，$\mathbf{w}$はすべてのパラメタをまとめたベクトル，すなわち，$\{\theta_j | j \in \{1, \ldots, J\}\} \cup \{a_i | i \in \{1, \ldots, I\}\} \cup \{d_i | i \in \{1, \ldots, I\}\}$を並べたベクトルである。確率値は非常に小さい値になりうるため，通常は，$L(\mathbf{w})$を直接最大化するのではなく，$\mathcal{L}(\mathbf{w}) = -\log(L(\mathbf{w}))$として$\mathcal{L}(\mathbf{w})$を最小化する。

2PLではなく，1PLについては，ロジスティック回帰モデルと等価である。この場合，$\mathcal{L}(\mathbf{w})$は$\mathbf{w}$について凸関数になり，大域的最適解を求められることが知られている[2]。

３．推定手法

３．１　MMLE

前述のように，2PLモデルのパラメタ推定においては，尤度関数から能力パラメタを周辺化して，項目パラメタのみに依存する形にすることで，同時推定に比べて一致性などいくつかの利点があることが知られている[2]。

Marginalized Maximum Likelihood Estimation (MMLE)法は，Python言語を用いて2PLモデルを行おうとしたときに，有力な選択肢の1つとなる**pyirt**パッケージ[5]内で使われている方法である。この方法は，“Marginalized”（周辺化）と名前についているものの，周辺化の仕方が，R言語の**ltm**パッケージ[6]など他の手法とは異なる。後者では，通常，ガウス - エルミート求積などの数値積分の手法を使って能力パラメタの周辺化を行うのに対し，この手法では，能力パラメタの取りうる値を量子化し，能力パラメタを連続変数ではなく離散変数として表し，Expectation Maximization (EM)アルゴリズムを用いて，数値積分を用いずに周辺化を近似する操作を行う[7]。

今回は，Python言語で項目反応理論を実行するときに選択肢に挙がりやすいこと，この手法が数値積分の設定に依存しておらず，比較が容易であることから，この手法を選択した。

３．２　Adam

ADAptive Moment estimation (Adam)[3]は，深層学習モデルのパラメタ推定器（optimizer）として標準的に用いられるアルゴリズムである。この手法は，いわゆる最急降下法（Gradient Descent）を，ランダムにサンプルした限られた訓練データのみを用いて行うことにより，$\mathcal{L}(\mathbf{w})$の最小値を探そうとする確率的勾配降下法（Stochastic Gradient Descent）がもとになっている。最急降下法の式を次に記す。

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \alpha \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}) \tag{4}$$

ここで，$\mathbf{w}_t$ は時刻 t での全パラメタ $\mathbf{w}$ を表す。

確率勾配降下法では，$\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w})$ の計算の際に，式３で表されるように，すべてのデータ（式３の場合は IJ 件の y_{ij}）を参照するのではなく，このうち，限られた一部をサンプルして $\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w})$ を計算する（確率的勾配）。これにより，計算を高速化できるうえ，勾配方向にデータのサンプルによる確率的な変動が生まれるため，局所解に停滞することをある程度避けられる効果がある。データをサンプルして更新，を繰り返す中で，全データを１回使用することをエポック（epoch）という。α は学習率（learning rate）と呼ばれ，この値が大きければ大きいほど，一度に更新される度合いが大きくなる。

Adam[3] の式は複雑であるため，紙面の都合上，割愛する。直感的には，確率的勾配の計算の際に，確率的勾配が振動をすることを防ぐ「モーメンタム」と呼ばれる技法と，確率的勾配の大きさ（学習率）を自動調節するRMSPropと呼ばれる技法を組み合わせたものである。

４．評価

４．１　データセット

試験データには，公開されている語彙テストの結果データセットを用いた[8]。このデータセットには，応用言語学分野で英語学習者の語彙量を測定するなどの目的で用いられるVocabulary Size Test (VST)[9] を，主に日本語を母語とする英語学習者100人に記述させたデータが格納されている。VSTは，100問からなるテストである。各設問では，文中に埋め込まれ，下線で明示された単語の意味と最も近い意味を持つ選択肢を，４つの選択肢から選択することが求められる。

各設問の各選択肢は，文法的な手がかりなどから単語の意味を推測できてしまわないように，どの選択肢も下線部と文字通り置き換えても，文として成立ように設計されている。さらに，選択肢は，英語母語話者によって問題として成立することが確認されている。全問に回答するためには，約30分程度の時間が必要である。

このデータセットの被験者は，クラウドソーシングサービスLancers[10] で募集した。募集時に，データセットが研究などの目的で公開される可能性があることを明示し，同意した被験者のみが回答している。英語を全く知らない者が被験者にならないように，募集時に，過去にTest of English for International Communication (TOEIC)を受けたことがある者だけを募集対象とした。

全体として100人が100件の設問に回答しているため，10,000件の反応データがある。
単語を90件選び，この単語を訓練データ（パラメタ推定に使用するデータ）として用いた。

４．２　比較する推定手法

前述のMMLEとAdamを比較した。MMLEによる2PLモデルの実装には，前述のようにpyirt[5] を用いた。一方，Adamについては，PyTorch[11] を用いて2PLモデルを自作し，これを，PyTorchに内包されているAdamの実装を用いてパラメタ推定した。

確率的勾配は，１度にサンプルするサイズ（バッチサイズ）によって推定に影響が出て，サンプルの選び方によってはうまく学習が進まないことがある。
通常は，バッチサイズを300程度に設定するが，今回は，バッチサイズによる性能の違いを見ることが主眼ではないため，一度に全パラメタを更新する手法で学習した。

エポック数は100とした。通常，深層学習モデルでは，Adamの学習率は0.01や0.001といった小さな値が用いられる。しかし，一度に全パラメタを見る設定にしたためか，100エポックでは，十分に負の対数尤度関数 $\mathcal{L}(\mathbf{w})$ が下がらなかったので，学習率を0.1に設定した。

pyirt[5] では，パラメタが動く範囲がデフォルト値で設定されている。具体的には，下記のように設定されている。

能力値パラメタ $\theta_j \in [-4, 4]$

困難度パラメタ $d_i \in [-2, 2]$

識別力パラメタ $a_i \in [0.25, 2]$

このうち，能力値パラメタ，困難度パラメタについては，PyTorchの **max_norm** 機能を用いてPyTorch実装でも同じ範囲を用いた。ただし，これはパラメタのノルムを設定する機能であるので，識別力パラメタについては，$[-2.0, 2.0]$ の範囲を設定した。

PyTorch実装では，パラメタの初期化を **nn.init.constant_** を用いて行った。具体的には，a_i については，1.0，d_i と θ_j については0.0に初期化した。

４．３　相関の指標

各パラメタについて，MMLEで推定した場合とAdamで推定した場合を比較する。この相関の指標としては，後述の理由で，ピアソンの積率相関係数と，スピアマンの順位相関係数を用いた[12)]。２つの確率変数X_1, X_2の間のピアソンの積率相関係数は次のように定義される。ただし，$cov(X_1, X_2)$はX_1とX_2の共分散，s_{X_1}，s_{X_2}は，それぞれX_1, X_2の標準偏差を表す。

$$\rho_{X_1,X_2} = \frac{cov(X_1, X_2)}{s_{X_1} s_{X_2}} \tag{5}$$

一般に多用されるピアソンの積率相関係数は，２変数の間に線形の相関があるかどうかを見る指標であるが，テストの分析においては，単純に，被験者の能力値や，項目の困難度などについては，順位の相関に興味があることも多い。このため，ピアソンの積率相関係数だけでなく，スピアマンの順位相関係数も併記した。

なお，スピアマンの順位相関係数は，両変数を順位に変換した変数間のピアソンの積率相関係数として定義される。ただし，同順については，同順間の順位の平均値で埋められる。例えば，２位，３位が同順であれば，両方とも2.5を順位の値として変換する。

４．４　実験結果と考察

図１にMMLEとAdamによる能力パラメタの推定値を示す。横軸がMMLEによる推定値，縦軸がAdamによる推定値である。能力パラメタに階段状の構造が見て取れる。これは，MMLEでは同じ能力値と判定された被験者が，Adamでは複数の値を取っていることを表す。この階段状の構造は，MMLEが能力パラメタを連続値ではなく，離散値として扱っていることに起因すると思われる。

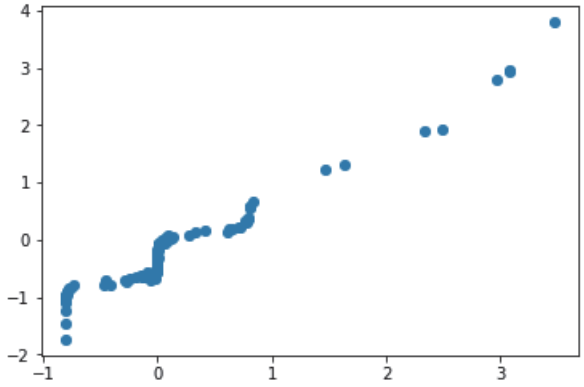

図１　MMLEによる能力パラメタの推定値（横軸）と，Adamによる能力パラメタの推定値（縦軸）の比較

階段状の構造があるものの，MMLEで推定された能力パラメタと，Adamによる能力パラメタは，比較的高い相関を示した。具体的には，ピアソンの積率相関係数で0.978，スピアマンの順位相関係数で0.997という高い値を示した。

図２に，MMLEによる困難度パラメタの推定値と，Adamによる困難度パラメタの推定値を示す。まず，全体的には相関があることが分かる。次に，MMLEではほぼ2.0と推定されている項目が，Adamだと複数の値に推定されていることが分かる。また，困難度の低い領域においては，うまく相関していないことが分かる。全体としては，ピアソンの積率相関係数で0.910，スピアマンの順位相関係数で0.948という高い相関を示した。

図３に，MMLEによる識別力パラメタの推定値と，Adamによる識別力パラメタの推定値を示す。やはり，全体的には相関があることが見て取れる。Adamでは一部の識別力パラメタの推定値が，設定範囲最大になってしまっていることが分かる。全体としては，ピアソンの積率相関係数で0.806，スピアマンの順位

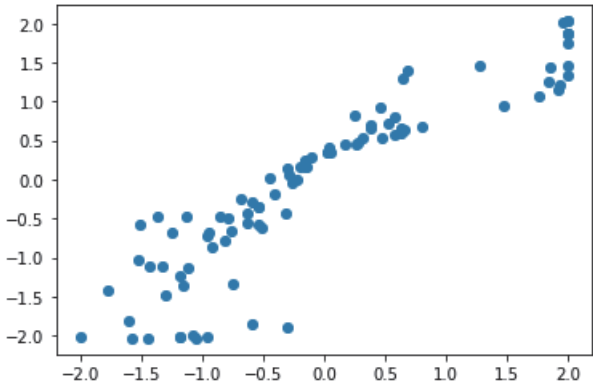

図２　MMLEによる困難度パラメタの推定値（横軸）と，Adamによる困難度パラメタの推定値（縦軸）の比較

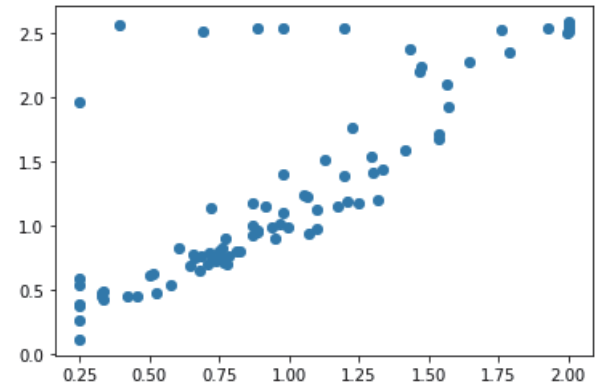

図３　MMLEによる識別力パラメタの推定値（横軸）と，Adamによる識別力パラメタの推定値（縦軸）の比較

相関係数で0.836という高い相関を示した。

以上，すべての相関について，無相関を対立仮説にした場合，相関は統計的に有意であった（$p < 0.01$）。全体として，Adamによるパラメタ推定値は，MMLEによるパラメタ推定値とよく相関することが見て取れた。

5．まとめ

深層学習モデルのパラメタの解釈性が近年，注目されているが，深層学習モデルのパラメタ推定法は，パラメタ解釈を直接の目的とした手法ではないため，従来の確率モデルの厳密なパラメタ推定法とは異なる。本研究では，試験など能力にかかわるパラメタの解釈の枠組みとしてよく利用される項目反応理論の2PLモデルについて，深層学習モデルの典型的なパラメタ推定法によって求められた推定値が，項目反応理論の従来の厳密なパラメタ推定法と一致するかどうか確認した。実験の結果，相関の高い順に，能力パラメタ，困難度パラメタ，識別力パラメタであることが分かった。また，全体として高い相関がみられることを示した。これは，2PLの考え方を用いた深層学習モデルのパラメタが，深層学習独自のパラメタ推定法に影響されず，解釈可能であることを示唆する。

今後の課題としては，他のデータセットを用いたさらなる検証が挙げられる。

謝辞

本研究は，科学技術振興機構ACT-X研究費（JPMJAX2006），ならびに日本学術振興会科学技術研究費補助金（18K18118）の支援を受けた。

参考文献

1） Emiko Tsutsumi, Ryo Kinoshita, and Maomi Ueno. Deep-IRT with independent student and item networks. In *Proc. of the 14th International Conference on Educational Data Mining (EDM),* Vol. 29, pp. 510–517, 2021.

2） Frank B. Baker. *Item Response Theory : Parameter Estimation Techniques, Second Edition.* CRC Press, July 2004.

3） Diederik P. Kingma and Jimmy Ba. Adam: Amethod for stochastic optimization. In *Proc. of the 3rd International Conference on Learning Representations* (ICLR), pp. 1–15, 2015.

4） Georg Rasch. *Probabilistic models for some intelligence and attainment tests. Danmarks Paedagogiske Institut,* 1960.

5） pyirt: A python library of irt algorithm designed to cope with sparse data structure. https://github.com/17zuoye/pyirt/, 2018. (August 16, 2021)

6） ltm: Latent Trait Models under IRT. https://cran.r-project.org/web/packages/ltm/, 2018. (August 16, 2021)

7） Bradley A. Hanson. IRT parameter estimation using the EM algorithm. Unpublished Manuscript. Available from: http://www.openirt.com/b-a-h/papers/note9801.pdf, 2000.

8） Yo Ehara. Building an English Vocabulary Knowledge Dataset of Japanese English-as-a-Second-Language Learners Using Crowdsourcing. In *Proc. of the 11th Edition of the Language Resources and Evaluation Conference* (LREC), pp. 484–488, 2018.

9） David Beglar and Paul Nation. A vocabulary size test. *The Language Teacher*, Vol. 31, No. 7, pp. 9–13, 2007.

10） Lancers. https://www.lancers.jp/. (August 16, 2021)

11） Pytorch. https://pytorch.org/. (August 16, 2021)

12） 武藤眞介. 統計解析ハンドブック（普及版）. 朝倉書店, 2010.