人文科学とコンピュータ　44－4
（１９９９．１０．１５）

LSAによる共観福音書問題の解析

三宅　真紀（東京工業大学社会理工研究科）
mmiyake@dp.hum.titech.ac.jp
佐藤　研（立教大学コミュニティ福祉学部教授）
赤間　啓之（東京工業大学社会理工研究科助教授）

本論文では、聖書学の分野に、コーパス言語学的な統計的解析を用いた方法論を導入をすることを目的としている。そこで、新約聖書学において、福音書の成立上の相互関係を整合的に説明しようとする「共観福音書問題」に着目し、その最も説得的な解決法と考えられている「二資料説」について、数量化モデルを立てた。統計処理については、特異値分解（Singular Value Decomposition, SVD）を基盤とするLSA（Latent Semantic Analysis）を用いて仮説を検証し、数量化モデルを用いて「二資料説」を実証した。LSAは、膨大な量のテクストを扱うのに非常に適している。また、解析ソフトウェアの開発も同時に行い、将来的に聖書学研究者の統計的研究をサポートすることを目的としている。

Approaching to the synoptic problem by Latent Semantic Analysis

Maki MIYAKE（Tokyo Institute Technology）
mmiyake@dp.hum.titech.ac.jp
Migaku SATO（Rikkyo University）
Hiroyuki AKAMA（Tokyo Institute Technology）

In this paper, it is our aim to use a statistical analysis for the study of Bible.We deal with the " synoptic problem " in New Testament Studies. For the first step,a statistical model is created for the "two sources theory" which plays a important part in this problem.Then, the hypothesis is explored by a mathematical technique called Latent Semantic Analysis (LSA) . This thechnique uses Singular Value Decomposition（SVD）, a mathematical generalisation of factor analysis. And also it is applied to a large corpus of text. Finally the hepothesis is proved by the satistical model. In addition we develop a application applied to the statistical analysis for the study of Bible.

1. はじめに

近年、人文科学の分野においてもコンピュータの重要性が認められ、コンピュータを導入した研究が盛んである。聖書学においても同様な傾向が見られ、コンピュータを駆使して制作したコンコルダンスや聖書ソフトウェアの開発などと、コンピュータ技術を大いに取り入れた研究が行われている。しかしながら、コンピュータを用いた統計的研究については、数少ないといえる。この分野に、コーパス言語学的な統計解析を用いた方法論の導入は、新たな視点からの議論が可能にするためにも必要不可欠であると考える。本論文では、新約聖書学において、福音書の成立上の相互関係を整合的に説明しようとする「共観福音書問題」に着目し、その最も説得的な解決法と考えられている「二資料説」について数量化モデルを立てた。統計処理については、比較的最近開発された統計解析である LSA を用いて、仮説を検証し、数量化モデルを用いて「二資料説」を実証した。また、将来的に聖書学研究者の統計的研究をサポートすることを目的とした、解析ソフトウェアの開発も行った。

2. Latent Semantic Analysis

LSA (Latent Semantic Analysis) は、テクスト中に出現する単語の文脈上の意味的関係を、統計的な手法を用いて機械的に抽出することが可能であり、膨大な量のテクストを扱うのに適している。LSA は、ニューラルネットモデルと非常に関係しているが、因子分析を数学的に一般化した、特異値分解（Singular Value Decomposition, SVD）を基盤としたものである。LSA では、この SVD を用いて意味的空間（semantic space）の次元を大幅に落とすことが特徴的である。元来は、Deerwester, Dumais, Furnas, Landauer, Harshman らが、情報検索方法を発展させたものであった。[1] そして、Foltz や Landauer, Dumais によって、談話分析や言語習得に関する一般的問題論へと拡張されていった [2], [3]。

線形代数の定理において、あらゆる正方行列 M は、３つの行列の積の形で表すことができる。

$$M = A D A'$$

ここで、A, A' は行列の固有ベクトルを構成している行列であり、また、D は固有値（または特異値）の対角行列である。

LSA では、この定理を一般の (m, n) 型行列 M に拡張して用いる。

$$M = U D V^{-1}$$

ここで、U, V はそれぞれ m, n 次のユニタリ行列であり、固有ベクトルを構成している。また、D は特異値の対角行列である。

このとき、行列の特異値は大きさの順番に並べられており、また、分解した３つの行列の積をとると元の行列 M に等しい。ここで、LSA では、この特異値のほとんどを捨ててしまい、大きなものだけを残すのが特徴的である。従って、その３つの行列の積は、M を近似した値になる。

$$M \approx \hat{M}$$

この近似は、膨大な情報量を圧縮するのに非常に有利である。[4]

LSA では、テクストを行列として表示したものを生データとして扱う。このとき、各行は単語を表し、各列はテクストの段落（あるいは、他のテクスト）を表す。そして、各成分には、列の段落（あるいは、他のテクスト）で出現される、行の単語の頻度数を入れる。これらの成分は、SVD によって変換される。また、ある特定の段落における単語の重要度や一般の場合において単語類型が情報を持っている程度を表す関数によって、各成分の値の重みが付けられる。[5]

3. 共観福音書問題

新約聖書の文学類型の一つに福音書がある。この文学類型は、キリスト教会において新しく作り

出されたもので、宣教的意味を持つ。福音書には、マルコ、マタイ、ルカ、ヨハネ福音書の四文書がある。これらの福音書は、それぞれ別の著者によって書かれたものである。そして、様々な口伝伝承、文献資料を用いて叙述されており、イエスの登場・活動を描き、受難と復活で終わる。

この福音書の、マルコ、マタイ、ルカ福音書の三福音書については、イエスを叙述する観点や全体構成枠が共通していること、さらには、一字一句すら一致している重複記事が多いことから、「共観福音書」と呼ばれている。実際、共観福音書を比べてみると、マルコ福音書全体の約９５パーセントが、マタイ・ルカ福音書のいずれかと共通している。その共通部分は、マタイの約５８パーセント、ルカの約４１パーセントに相当している。また、マタイ、ルカ福音書において、マタイ・ルカにのみ共通している部分については、それぞれ約２０パーセントの割合である。このような共観福音書の類似性から、その要因を探り、文書間の成立上の相互関係を整合的に説明しようとするのが、「共観福音書問題」である。[6]

この問題に対して、様々な仮説が提唱されてきた。その中で、１９世紀の半ばに立てられた仮説で、現在では、ほぼ定説化しているのが「二資料説」である。この仮説は、C.H.Weiss によって立てられ、H.J.Holtzmann によって聖書学的・総合的に叙述された。[7]

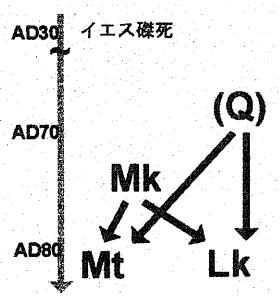

図 1　二資料説

この「二資料説」を構成する重要な一仮説に、K.Lachmann によって立てられた「マルコ優先説」がある。これは、マルコの物語の順序がマタイ、ルカによって継承されていることから、マルコが最も古く書かれたもので、マタイとルカがマルコを資料として用いたと考える説である。各福音書の成立時期については、マルコが紀元後７０年代半ば、マタイ、ルカ福音書はほぼ同時期で、８０年代と考えられている [8]。この「マルコ優先説」に加えて、「二資料説」は、マタイ・ルカ福音書のみに現れる箇所が頻出することから、マタイとルカは、マルコとは別の資料を用いていたことを想定した。一般に、この資料を「Q資料」と呼んでいる。「Q」は、「資料」を意味するドイツ語 Quelle の頭文字に由来する。つまり、マタイ・ルカ福音書は、共通の資料としてマルコ福音書と「Q」をそれぞれ用いたと考える説である。（図１）

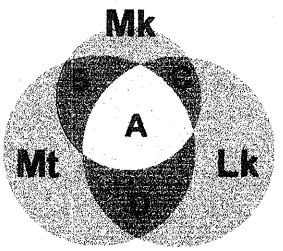

図 2　共通部分

ここで、共観福音書の各共通部分について図２に示す。マルコ（Mk）、マタイ（Mt）、ルカ（Lk）福音書を示す円の大きさは、３福音書の文書量の比率を表している。このとき、三文書共通部分をA、マルコ・マタイ共通部分を B、マルコ・ルカ共通部分を C、マタイ・ルカ共通部分を D とする。この共通部分を用いて、「二資料説」を検証するモデルを試みた。

4.　仮説検証モデル

「二資料説」を検証するにモデルを考えるに当

たって、ルカ福音書を基準にする。

ルカ福音書が「二資料説」に従い、マルコ福音書とQ文書の二つを資料として用いていたのならば、ルカ文書の出現単語は、マルコ・ルカ共通部分（C）の出現単語とQの部分と考えられているルカ・マタイ共通部分（D）の出現単語のどちらかと一致するはずである。すなわち、単語の出現頻度数を用いて、ルカ文書と図2で示された共通部分C,Dとの相関関係を調べたとき、ルカ福音書は、マルコから受容しいる部分であるCに強い相関を持つ部分とQから受容している部分であるDに強い相関を持つ部分の2つに分かれると考えられる。（図3）

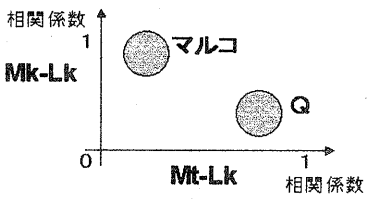

図 3　仮説検証モデル

このとき、三福音書共通部分のAについては、C,Dのどちらに含まれるのかは、この段階においては特定できないため、ここでは考慮の対象としない。さらに、マルコ福音書は、Q文書の存在を知らずに成立されたと考えられているので、C,Dは独立していることを前提とする。

このモデルに対して、実際にデータを取得し、LSAを用いて仮説の検証を行った。

5.　検証方法

次に、実際の共観福音書データの処理方法について示す。

テキストは『ネストレ＝アーラントの新約聖書(Novum Testamentum Grace) 第27版』を使用した。ルカ文書を、ルカに準じた平行箇所（166箇所）で区切り、それぞれの平行箇所で分析を行った。一つの平行箇所について出現する単語の頻度数をルカ文書と共通部分C,Dの4パターンについてカウントし、それを行列の型として表す。この単語については、基本的に出現した全ての単語を使用した。しかし、例外処理として、定冠詞と一部の接続詞・前置詞（and, but, inにあたるギリシャ語）については、頻度数が他の単語に対して過度に多く、どのパターンにおいても必ず出現されるため削除した。単語は、文章の一致度を見る目的とするために、同じ意味であっても形態が違う場合はそれぞれ別の単語と認識して扱った。このようにして出現単語の一部を除いた単語の頻度数行列を、LSAに用いる生データとした。この行列をLSAにかけて出力された行列の値を用いて、ルカ文書と共通部分C,Dとの相関をとる。この一連の作業を、共観福音書ソフトウェアとLSAソフトウェアの両方を用いて行った。

6.　結果・考察

平行箇所166個の中で、「共通部分C,Dが独立であるという二資料説」モデルの前提条件を満たしたのは、114箇所であった。この144個のデータについて、LSA処理後のルカ文書に対する、マタイ・ルカ共通とマルコ・ルカ共通とのそれぞれの相関係数をプロットしたものを図4に示す。

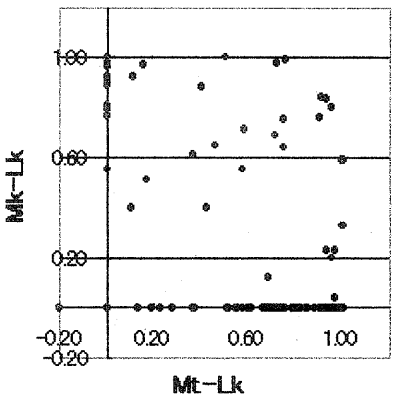

図 4　ルカ文書との相関図

図４において、横軸はルカとマタイ・ルカ共通部分（D）との相関係数、縦軸はルカとマルコ・ルカ共通部分（C）との相関係数である。

図４で示された、ルカとそれぞれの共通部分との相関は、共通部分Cのみに強い相関をもつ共通箇所と、共通部分Dのみに強い相関をもつ共通箇所の２つに大きく分かれていることがいえる。すなわち、ルカ文書においては、共通部分C（マルコ部分）を受容している箇所と、共通部分D（Q部分）を受容している箇所の２つに分かれていることが確認できた。したがって、「二資料説」のモデルにあてはまり、ルカ文書は、資料としてマルコとQの二資料を用いていたことが数量的に証明できた。

しかしながら、共通部分C,Dの両方に強い相関を持つ部分が若干存在する。これについては、今回のモデルにおいては、三福音書共通部分Aを考慮の対象としていないため、モデルの設定が不十分であると考えられる。

7. 解析ソフトウェア開発

次に、データを処理する際に用いたアプリケーションについて述べる。

開発中のため、福音書の平行箇所から出現単語頻度数を行列に表すためのアプリケーション（共観福音書ソフトウェア）と実際にLSAをするアプリケーション（LSAソフトウェア）を分離して、それぞれ別の言語を使って作成している。これは、最終的には、一つのアプリケーションとして連結させる予定である。

共観福音書ソフトウェア（図５）については、共観福音書（マタイ、マルコ、ルカ）を平行表示させ、平行箇所を任意に設定できるようにした。また、設定された平行箇所について、三文書共通、マタイ・マルコ共通、マタイ・ルカ共通、マルコ・ルカ共通、マタイ、マルコ、ルカの７つのパターンにおける出現単語頻度数をマトリックス表示で出力する。このとき、出現単語を行ごとに表し、各パターンを列表示している。このマトリックスが、LSA解析にかけるデータとなる。さらに、これはまだ試験段階であるが、共通部分について色分けをして表示することも可能である。これは、共通箇所が分かりやすく表示する目的である。

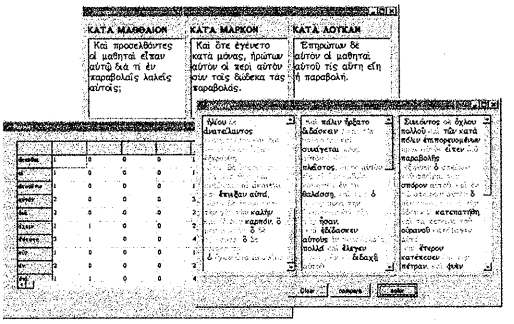

図 5　共観福音書ソフトウェア

LSAソフトウェア（図６）は、入力されたマトリックスに対して、LSAによって近似されたマトリックスに関する、列どうしの相関が出力されるようになっている。また、その途中経過である、特異値、固有値ベクトル、近似されたマトリックスの値も記録されている。さらには、LSA処理を行うことなく、生データのままで相関をとることができ、LSAとの値の比較も可能である。

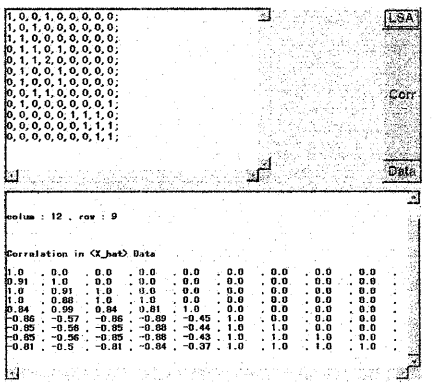

図 6　LSA ソフトウェア

実際、データの処理の過程は、まず共観福音書ソフトウェアを使って、各平行箇所の出現頻度数

のデータを求めた。そして、LSA ソフトウェアを使って、求めたデータを入力し、相関を求めた。

8. まとめ

本論文では、新約聖書学の「共観福音書問題」を取り上げ、その「二資料説」について数量化モデルを立て、LSA を用いて仮説を検証し、実証した。今回の実験結果から、モデルの設定には、改善の余地があることが判明した。今後は、改善したモデルを用いて実験をさらに行うことを予定としている。

参考文献

[1] Deerwester *et.al.*, (1990), Indexsing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41, 391-407.

[2] Foltz, P.W.(1996), Latent Semantic Analysis for text-based research. *Behavior Research Methods, Instruments, Computers*, 28, 197-202.

[3] Landauer *et al.*, (1997), A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, indunction and representation of knowledge. *Psychological Review*, 104, 211-40.

[4] Kintsch, W. (1998), Comprehension: A paradigm for cognition. *New York: Cambridge University Press.*

[5] Landauer et al., (1998), Introduction to Latent Semantic Analisis, *Discource Processes.* 25, 259-284.

[6] 荒井献他『総説　新約聖書』、日本基督教団出版局、1981

[7] 木幡藤子他編『現代聖書講座　第２巻』、日本基督教団出版局、1996

[8] 『新約聖書　福音書』佐藤　研他訳、岩波書店、1996