A Statistical Analysis of the Synoptic Gospels

V. Choulakian[1], S. Kasparian[1], M. Miyake[2], H. Akama[2],
N. Makoshi[2], M. Nakagawa[2]

[1]Université de Moncton, Moncton, N.B., E1A 3E9, Canada.

2Tokyo Institute of Technology, Japan

Abstract

A statistical analysis of two contingency tables calculated from the synoptic gospels is done by correspondence analysis (CA) and taxicab correspondence analysis (TCA). We deduce a variant of two gospel hypothesis from the results of TCA.

Keywords : Textual data analysis ; parallel texts ; synoptic gospels ; correspondence analysis ; taxicab correspondence analysis ; two source hypothesis ; two gospel hypothesis ; bootstrapping.

1. Introduction

The Synoptic Problem has been one of the controversial subjects in the studies of the New Testament ; only a few studies so far have attempted to give an objective statistical explanation of the mutual relationships among the synoptic gospels, Matthew, Mark and Luke, see for instance, Conzelmann and Lindemann (1988 : 45-53). Furthermore, a large number of studies have made various assumptions of their genealogical interdependence based on subjective comparisons of the texts. The website http://www.mindspring.com/~scarlson/synopt/ presents more than twenty hypotheses about the original sources of the three synoptic gospels. However, there are two main hypotheses : Two source hypothesis (2SH) and two gospel hypothesis (2GH). To describe these two hypotheses we introduce some notation. Let the three synoptic gospels Matthew, Mark and Luke be denoted by Mt, Mk and Lk, respectively. We consider a partition of the union of the three synoptic gospels into seven disjoint categories :

union (Mt, Mk, Lk) = union (a, b, c, d, e, f, g),

where

a = Mt ∩Mk ∩Lk,
b = Mt ∩Mk ∩Lk′,
c = Mt′ ∩Mk ∩Lk,
d = Mt ∩Mk′ ∩Lk,
e = Mt ∩Mk′ ∩Lk′,
f = Mt′ ∩Mk ∩Lk′,
g = Mt′ ∩Mk′ ∩Lk,

and s′ is the complement of s.

The 2SH is based on the assumption that the three synoptic gospels have two original sources, Mk and Q :

(1) Mk = union (a, b, c, f) and Q = union (d, e, g).

The 2GH is based on the assumption that Matthew was first, and was used by Luke, and that Mark is a confluence of Matthew and Luke. There are many variants or modifications of the 2SH and 2GH.

To tackle the study of the synoptic gospels in an objective way, two steps are required. The first step is to construct the contingency table, where the complete data of the lexical usage patterns in the three synoptic gospels are distributed into the seven categories. Recently, this was accomplished by Miyake et al. (2004) by creating the web-based biblical software named Tele-Synopsis ; further details can be found at (http://nerva.dp.hum.titech.ac.jp/tele-synopsis/parallel). A brief description of the first step is done in section 2. The second step is to do a statistical analysis of the collected data to discern valid and stable structures. The main aim of this paper is to present the second step ; and this will be done in section 3 by correspondence analysis (CA) and taxicab correspondence analysis (TCA). A brief mathematical description of the TCA, which was recently proposed by Choulakian (2006) is provided in the appendix. Finally, we conclude with some remarks in section 4.

2. Datasets from Tele-Synopsis

The construction of the dataset for a statistical analysis is done by the software Tele-Synopsis. Tele-Synopsis allows us to manipulate lexical data of parallel and variant texts, and uses the NA26th version of the texts by Nestle and Aland (1979) and for the parallels the Synopsis for the Four Gospels by Aland (1989) ; the latter is recognized as the most reliable parallel synoptic table to date in the biblical studies. This system has a merit to make it possible for users to independently add and remove each sentence so as to customize their own synoptic table by changing the temporary segmentation of the pericope.

Table 1: Construction of the contingency tables by types.

Distributive type (SD)

	raw data			contingency table						
	Mt	*Mk*	*Lk*	*a*	*b*	*c*	*d*	*e*	*f*	*g*
word A	2	1	3	1	0	0	1	0	0	1
word B	0	3	2	0	0	2	0	0	1	0

Commonality type (SC)

	raw data			contingency table						
	Mt	*Mk*	*Lk*	*a*	*b*	*c*	*d*	*e*	*f*	*g*
word A	2	1	3	1	0	0	0	0	0	0
word B	0	3	2	0	0	2	0	0	0	0

We shall consider two types of distributing the words of the synoptic gospels into the 7 categories : Distributive and Commonality. The Distributive type is to distribute a word occurrence into the 7 categories ; the contingency table thus constructed will be designated by SD. The Commonality type is to distribute a word occurrence into the attributed category ; the arising contingency table is named SC. Table 1 displays both types of construction of the contingency tables. It is interesting to see the underlying latent factors in these two contingency tables. Each contingency table consisted of 7276 segments. By eliminating noise

words, such as articles, prepositions, pronouns and conjunctions, the number of segments was reduced to 7099. Each contingency table of 7099 rows and 7 columns is designated by T_1.

3. A statistical analysis

First, we shall analyze the contingency table SC, then SD.

3.1. SC

The contingency table T_1 is submitted to CA and TCA. The left part of Table 2 displays the dispersion measures and the associated cumulative proportion of the variance explained in %. There is a clear difference between TCA and CA dispersion measures. In CA it is not evident how many dimensions to choose, 2 or 4? The cloud of points seems to be spherical. While in TCA, it is evident that the first three dimensions are significant and they explain 92.17 % of the total dispersion.

Table 2 : Dispersion measures and cumulative explained variation in TCA and CA of Synoptic data SC.

T_1					T_{10}			
	TCA		CA		TCA		CA	
α	λ_α^2	$CEV(\alpha)$	λ_α^2	$CEV(\alpha)$	λ_α^2	$CEV(\alpha)$	λ_α^2	$CEV(\alpha)$
1	0.346	36.67	0.522	21.01	0.115	37.37	0.210	30.93
2	0.278	66.12	0.499	41.11	0.087	65.7	0.180	57.43
3	0.246	92.17	0.407	57.47	0.074	89.72	0.103	72.59
4	0.030	95.31	0.394	73.34	0.012	93.80	0.070	82.95
5	0.025	97.91	0.341	87.06	0.010	97.19	0.061	91.90
6	0.020	100	0.321	100	0.008	100	0.055	100

The left part of Tables 3 and 4 show the first four factor scores of the 7 categories obtained by TCA and CA respectively. TCA scores on the first three significant dimensions show that the 7 categories can be grouped into four distinct classes : {a, b, d}, {c, f}, {e} and {g}. From Table 3 we also obtain the following tree representation.

(2) $\overbrace{\overbrace{a\ b\ d,}\ \ e}\quad \overbrace{\overbrace{c\ f,}\ \ g}$

This tree shows that the first factor separates the 7 categories of T_1 into two large groups {a, b, d, e} and {c, f, g}. And, the second and third axes separate each large group into two subgroups.

Now, we want to address the following question : Is the tree representation (2) stable? Two different approaches were used to check the stability of the results of TCA. In the first approach, we applied bootstrapping of the contingency table T_1, see for instance Greenacre (1984 : ch.8), Alvarez et al. (2004) and Lebart (2004). 1000 bootstrapped samples of the contingency table T_1 were drawn, and TCA applied to each of them. The bootstrapped samples were chosen by randomly choosing each row with the marginal row weight affected to it. And each time the same tree representation (2) was obtained. In the second approach,

TCA and CA was applied to subtables T_k, for k = 1,...,10 of the original table T_1. The subtable T_k is characterized by the fact that the frequency of each row is greater or equal to k. Table 5 displays the number of rows of the subtables T_k. From Table 5 we observe that there are 4433

rows with marginal frequency of 1. These 4433 rows constitute a cluster of seven points with heavy marginal weights. Usually, rows with small frequencies disturb the results of CA, in the sense that they dominate the solution : Note that in Table 4, s_1 of T_1 (the opposition is between Mt and the rest) does not have the same interpretation as the s_1 of T_{10} (the opposition is between Lk and the rest). But this phenomenon did not happen in the case of TCA. TCA of each subtable T_k produced the same tree representation (2). For a comparison of results obtained from the original table T_1, the right part of Tables 2, 3 and 4 display the corresponding results obtained from T_{10}.

Table 3: TCA factor scores of Synoptic data SC.

T_1					T_{10}			
	s_1	s_2	s_3	s_4	s_1	s_2	s_3	s_4
a	-0.54	-0.73	-0.77	-0.50	-0.23	-0.38	-0.46	-0.28
b	-0.56	-0.79	-0.69	0.81	-0.21	-0.48	-0.21	0.48
c	0.45	-0.74	0.51	-1.11	0.39	-0.35	0.15	-0.63
d	-0.58	-0.61	-0.79	-0.64	-0.33	-0.17	-0.41	-0.33
e	-0.66	0.40	0.41	0	-0.39	0.22	0.22	0
f	0.57	-0.90	0.80	0.22	0.38	-0.51	0.48	0.15
g	0.56	0.39	-0.34	0	0.32	0.24	-0.21	0

Table 4: CA factor scores of Synoptic data SC.

T_1					T_{10}			
	s_1	s_2	s_3	s_4	s_1	s_2	s_3	s_4
a	0.07	0.13	-0.93	0.01	0.03	0.25	-0.78	0.36
b	-0.43	0.22	-1.70	-1.28	-0.29	-0.33	-0.71	-0.60
c	-0.18	-0.24	-0.53	-0.06	-0.39	0.22	-0.18	0.26
d	0.17	0.30	-1.23	2.31	0.21	0.17	-0.32	0.65
e	0.14	0.96	0.40	-0.08	0.50	-0.37	0.12	-0.01
f	-1.66	-0.48	0.40	0.17	-0.93	-0.47	0.24	0.15
g	0.58	-0.70	0.17	-0.07	0.04	0.53	0.16	-0.12

In both methods dispersion measures of table T_{10} are smaller than the corresponding dispersion measures of table T_1 : This is a well known theoretical result that the elimination of points reduces the dispersion measures. However, in CA the cumulative explained variations of T_1 are smaller than the corresponding cumulative explained variations of T_{10} ; and in the latter case the first three dimensions are quite clearly separated from the rest and they explain 72.59 % of the total inertia. The first principal axis opposes parts of Mk = union (b, c, f) to union (e, d). The second principal axis opposes Lk = union (a, c, d, g) to union (b, e, f) ; and

the third principal axis opposes the common parts = union (a, b, c, d) to the unique parts = union (e, f, g).

Table 5: Number of rows of the subtable $\mathbf{T}_k$.			
$\mathbf{T}_k$	rows	$\mathbf{T}_k$	rows
$\mathbf{T}_1$	7099	$\mathbf{T}_6$	611
$\mathbf{T}_2$	2666	$\mathbf{T}_7$	506
$\mathbf{T}_3$	1528	$\mathbf{T}_8$	424
$\mathbf{T}_4$	1035	$\mathbf{T}_9$	373
$\mathbf{T}_5$	776	$\mathbf{T}_{10}$	313

As stated above TCA results of tables T_1 and T_{10} are very similar and have the same interpretation : The dispersion measures of the first three principal axes are clearly separated from the rest and they explain 92.17 % of the total dispersion in T_1 or 89.72 % of the total dispersion in T_{10}. This makes 17 % = 89.72 % - 72.59 % more than the corresponding value in CA. The first TCA axis opposes Mt = union (a, b, d, e) to the rest. The second TCA axis opposes Mk = union (a, b, c, f) and d to the rest. The third TCA axis opposes union (a, b, d, g) to the rest : This shows that Lk is at the conflation of Mt and Mk. So, Matthew seems to be predominant, followed by Mark. Based on the tree representation (2) and the interpretation of the principal axes, we define

(3) proto - Matthew = union (a, b, d).

We see that

Mt = union (proto - Matthew, e),

(4) Mk = union (c, f, parts of proto - Matthew),

Lk = union (g, parts of proto - Matthew, parts of Mk).

From these considerations we deduce the following genealogical tree, which represents a modified 2GH :

Proto-Matthew

↓ ↓ ↓

MT MK → LK

3.2. SD

The approach used in the analysis of SC was applied for the analysis of the contingency table SD. The last rows of the Tables 6 and 7 show that the first three dimensions explained 92.19 % of the dispersion in TCA, and 58.60 % of the total inertia in CA. Note that the results of T_1 in both Tables 6 and 3 are almost identical. That is, TCA of T_1 in Table 6 produces the tree representation (2). To check the stability of the results, TCA was repeated on the subtables T_i, for i = 1,..., 25, and the first three principal axes, which are composed of ±1 (see the appendix), were compared. We noticed that the first three principal axes of the subtables T_{16}

to T_{25} were identical. T_{16} has 160 rows ; this means that 7099-160=6939 rows were bad influential points and were eliminated from the analysis. The first three dimensions of T_{16} in Tables 6 and 7 have almost the same interpretation as the first three dimensions in Tables 3 and 4. However, T_{16} in Table 6 does not produce the tree representation (2).

Table 6 : TCA factor scores of Synoptic data SD.

	T_1				T_{16}			
	s_1	s_2	s_3	s_4	s_1	s_2	s_3	s_4
a	-0.52	-0.72	-0.76	-0.48	-0.22	-0.34	-0.53	-0.15
b	-0.54	-0.79	-0.66	0.79	-0.20	-0.50	-0.19	-0.19
c	0.44	-0.73	0.49	-1.08	0.38	-0.36	-0.20	0.26
d	-0.57	-0.54	-0.76	-0.66	-0.29	0.09	-0.24	0.56
e	-0.65	0.37	0.38	0	-0.35	0.16	0.20	-0.05
f	0.54	-0.87	0.73	0.21	0.35	-0.48	0.37	0.10
g	0.55	0.40	-0.33	0	0.30	0.25	-0.14	-0.07
λ_α^2	0.331	0.266	0.222	0.028	0.098	0.074	0.050	0.014
$CEV(\alpha)$	37.24	67.17	92.19	95.40	38.50	67.70	87.35	92.89

Table 7 : CA factor scores of Synoptic data SD.

	T_1				T_{16}			
	s_1	s_2	s_3	s_4	s_1	s_2	s_3	s_4
a	0.06	0.10	-0.86	-0.40	0.00	0.16	-0.77	0.34
b	0.56	0.12	-1.97	0.49	0.31	-0.37	-0.72	-0.48
c	0.21	-0.32	-0.50	-0.29	0.40	0.25	-0.22	0.38
d	-0.13	0.38	-0.17	-2.48	-0.29	0.11	-0.17	0.72
e	0.02	0.94	0.31	0.24	-0.47	-0.36	0.13	-0.04
f	1.45	-0.74	0.49	-0.01	0.90	-0.37	0.28	0.11
g	-0.70	-0.58	0.09	0.13	0.01	0.54	0.11	-0.13
λ_α^2	0.502	0.484	0.374	0.359	0.188	0.158	0.072	0.047
$CEV(\alpha)$	21.62	42.46	58.60	74.07	34.74	63.89	77.24	85.94

4. Conclusion

Usually textual contingency tables are sparse ; for instance, 78.6 % of the entries of the contingency table SC have zero frequencies. Sparseness implies the existence of outliers or influential observations. Influential points can be classified as good or bad. Bad influential points are named outliers. For instance in SC, 4433 rows with marginal frequencies of 1 make 7 cluster points with large weights. Generally, in the analysis of contingency tables of textual data by CA, rows of feeble marginal frequencies are considered outliers and deleted from analysis. The analysis of the dataset SC showed that this was not necessary in the case of TCA ; and the reason is that the influential points were positioned in the same direction as the first three principal axes. So these 7 cluster points are considered good influential points, while they are considered bad influential points in the case of CA. In the contingency table SD, 6939 rows were considered bad influential points and deleted from both TCA and CA. The procedure used in this paper shows that, TCA aids us to distinguish good influential points from bad influential points very easily : It suffices to compare principal axis weights,

composed of ±1, of T_is. Our aim in data analysis is to delete the least number of rows and to have maximum stable results.

CA, similar to principal component analysis and factor analysis, has rotational indeterminacy, because it is based on the L_2 norm. TCA does not have the rotational indeterminacy problem, because it is based on the L_1 norm. To compare bootstrapped results in TCA, all we have to do is to compare the principal axis weights which are composed of ±1.

We conclude by the fact that both T_1 contingency tables represent the population of the three synoptic gospels in two different ways, and the tree representation (2) was discovered by TCA in both T_1 contingency tables.

Appendix

Let P=T/n be a correspondence matrix, where T of dimension r x c is a contingency table, $n = \sum_{j=1}^{c} \sum_{i=1}^{r} T_{ij}$, the grand total of T. We define $p_{i.} = \sum_{j=1}^{c} p_{ij}$, $p_{.j} = \sum_{i=1}^{r} p_{ij}$, $D_r = Diag(p_{i.})$ a diagonal matrix having diagonal elements $p_{i.}$, and similarly $D_c = Diag(p_{.j})$. The q-th vector norm of a vector $v=(v_1,...,v_m)'$ is defined to be $\|v\|_q = (\sum_{i=1}^{m} |v_i|^q)^{1/q}$ for $q \geq 1$ and $\|v\|_\infty = \max_i |v_i|$. Let k = rank(P)-1.

In TCA the calculation of the dispersion measures λ_α, principal axes and principal factor scores g_α and s_α, for $\alpha = 0,1,...,k$, is done in an stepwise manner. We put $P_0 = P$. Let P_α be the residual correspondence matrix at the α-th iteration.

The variational definitions of the TCA at the α-th iteration are

(5) $$\lambda_\alpha = \max_v \frac{\| P_\alpha v \|_1}{\| v\|_\infty} = \max_u \frac{\| P_\alpha' u \|_1}{\| u\|_\infty} = \max_{u,v} \frac{u' P_\alpha v}{\| u\|_\infty \| v\|_\infty},$$

(6) $$= \max \| P_\alpha v\|_1 \text{ subject to } v_j = \pm 1 \text{ for } j=1,...,c,$$

(7) $$= \max \| P_\alpha' u\|_1 \text{ subject to } u_i = \pm 1 \text{ for } i=1,...,r.$$

Let

(8) $$v_\alpha = \arg\max_{v_{j=\pm 1}} \|P_\alpha v\|_1,$$

(9) $$u_\alpha = \arg\max_{u_{j=\pm 1}} \|P_\alpha u\|_1.$$

Then the transition formulas are

(10) $$g_\alpha = D_r^{-1} P_\alpha v_\alpha,$$

(11) $$s_\alpha = D_c^{-1} P_\alpha' u_\alpha,$$

(12) $$u_\alpha = \text{sgn}(g_\alpha),$$

(13) $$v_\alpha = \text{sgn}(s_\alpha),$$

where sgn (.) is the coordinatewise sign function, sgn (x) = 1 if x > 0, and sgn (x) = -1 if x ≤ 0.

The α-th taxicab dispersion measure can be represented in many different ways

(14) $$\lambda_\alpha = \| P_\alpha v_\alpha \|_l = \| D_r g_\alpha \|_l = u_\alpha' D_r g_\alpha ,$$

(15) $$= \| P_\alpha' u_\alpha \|_l = \| D_c s_\alpha \|_l = v_\alpha' D_c s_\alpha .$$

The (α+1)-th residual correspondence matrix is

(16) $$P_{\alpha+1} = P_\alpha - D_r g_\alpha s_\alpha' D_c / \lambda_\alpha .$$

Similar to the ordinary CA, the total dispersion is defined to be $\sum_{\alpha=1}^{k} \lambda_\alpha^2$, and the proportion of the explained variation by the α-th principal axis is $\lambda_\alpha^2 / \sum_{\alpha=1}^{k} \lambda_\alpha^2$, and the cumulative explained variation in % is $100 \sum_{\beta=1}^{\alpha} \lambda_\beta^2 / \sum_{\gamma=1}^{k} \lambda_\gamma^2$, for α = 1,...,k.

We note that

(17) $$P_1 = P - p_r p_c' \ ;$$

that is, the best rank one approximation of P is given by $(p_i p_j)$, which is the correspondence matrix obtained under the independence assumption between the row and column variables. This solution is considered trivial both here and in CA. The reconstitution formula in TCA and CA is

(18) $$p_{ij} = p_i p_j \left[1 + \sum_{\alpha=1}^{k-1} g_\alpha(i) s_\alpha(j) / \lambda_\alpha \right].$$

The calculation of the principal scores and the principal component weights of TCA can be accomplished by two algorithms. The first one is based on complete enumeration using equations (6) or (7). The second one is based on iterating the transition formulae (10, 11, 12, 13). This is an ascent algorithm. The iterative algorithm could converge to a local maximum ; so it should be restarted from multiple initial points. The rows or the columns of the data can be used as initial values.

More technical details about TCA and a deeper comparison between TCA and CA is done in Choulakian (2006).

References

Aland, K. (1989). *Synopsis of the Four Gospels*. 9th ed. Stuttgart, German Bible Society.

Alvarez, R., Bécue, M., Valencia, O. (2004). Étude de la stabilité des valeurs propres de l'AFC d'un tableau lexical au moyen de procédures de reéchantillonage. *JADT 2004* : 42-51.

Choulakian, V. (2006). Taxicab correspondence analysis. *Psychometrika*, 71, 2 : 1-13.

Conzelmann, H., and Lindemann, A. (1988). *Interpreting The New Testament*. Trans. Siegfried S. Schatzmann. Peabody, Mass., Hendrickson Publishers.

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press.

Lebart, L. (2004). Validité des visualisations de données textuelles. *JADT 2004* : 708-715.

Miyake, M., Akama, H., Sato, M. and Nakagawa, M. (2004). Tele-Synopsis for biblical research. *Proceedings of the IEEE ICALT* : 931-935.

Nestle, E., and Aland, K. (1979). *Nestle-Aland Novum Testamentum Graece*. 26th ed. Stuttgart, Deutsche Bibelstiftung.