Enliven Archive
Intensifying Thoughts

Research Article www.enlivenarchive.org Enliven: Biostatistics and Metrics

A Note on Change-Point Estimation in a Multinomial Sequence

James A. Koziol[1*]

[1]College of Health, Human Services and Science, Ashford University, San Diego, California

[*]**Corresponding author**: James A. Koziol, College of Health, Human Services and Science, Ashford University , San Diego, California 92123, Tel: (858) 776 - 1926; E-mail: james.koziol@faculty.ashford.edu

Received Date: 19th July 2014

Accepted Date: 03rd September 2014

Published Date: 06th September 2014

Citation: Koziol JA (2014) A Note on Change-Point Estimation in a Multinomial Sequence. Enliven:Biostat Metr 1(1):001.

Abstract

We revisit the multinomial change-point problem posed by Riba and Ginebra [1] relative to a putative change in authorship of *Tirant lo Blanc*. We utilize Lancaster partitions of chi-squared tests of homogeneity, as well as techniques suggested from cluster analysis, in our investigations. Our findings are largely congruent with those of Girón, Ginebra, and Riba, with strong evidence for two authors.

Keywords

Multinomial change-point; Lancaster decomposition; Cluster analysis

Introduction

Riba and Ginebra [1] have recently explored classical parametric methods for detecting a change-point in the distribution of a sequence of independent multinomial random variables. A standard technique, which they exploited, consists of maximization of the likelihood function, assuming a single change in the vector parameter of multinomial probabilities at an unknown point in the sequence.

In this note, we examine a different technique for assessing a single change-point in a sequence of independent multinomial observations. Our procedure is similar to that of Riba and Ginebra in that it also is based on underlying chi-squared tests of homogeneity, but is not novel: rather, it is derivative of the fundamental papers of Lancaster [2] and Irwin [3], in which are described decompositions or partitions of overall chi-squared statistics into independent components. One advantage of the Lancaster approach in the multinomial change-point setting is the transparency of underlying distribution theory, compared to that attending the maximum of a sequence of dependent chi-squared test statistics.

Riba and Ginebra illustrated their techniques with data derived from the book *Tirant lo Blanc.*

In the following section, we briefly characterize the change-point problem relative to Riba and Ginebra's data extraction of word lengths from the chapters of *Tirant lo Blanc*. We then investigate change-points in Riba and Ginebra's table of word lengths, using Lancaster decompositions. The individual components reveal that the word lengths in the various chapters of *Tirant lo Blanc* are rather heterogeneous, even with a presumptive change in authorship. We investigate this heterogeneity further with a cluster analytic technique, and summarize findings in the concluding section.

Tirant lo Blanc

Tirant lo Blanc is an epic romance, published in Valencia in 1490. It is one of the most important works of Catalan literature, and influenced the evolution of the Western novel in general and Miguel de Cervantes' classic work of fiction, *Don Quixote de la Mancha*, in particular. Authorship of *Tirant lo Blanc* is attributed to the Valencian knight Joanot Martorell, with the latter chapters added by Martí Joan De Galba. Riba and Ginebra [1] set out to determine whether they could detect a change-point in the literary style of *Tirant lo Blanc* that might indicate where in the text this putative change in authorship occurred.

For our purposes here, we utilize the data comprising Riba and Ginebra's Table 1. This is a 425 x 10 table of counts of words of lengths 1, 2, …, 10+ in the 487 chapters of *Tirant lo Blanc*, obtained after eliminating those chapters with fewer than 200 words. In both this table and an accompanying table of frequencies of common words in the successive chapters, Riba and Ginebra assume an independent multinomial sampling scheme: that is, the rows in the tables comprise sequences of conditionally independent multinomial observations, given respective total word counts $n_1, n_2, \ldots, n_{425}$ in the 425 rows, and respective cell probabilities $\pi_1, \pi_2, \ldots, \pi_{425}$, where $\pi_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{i10})$. Riba and Ginebra then formulate the change-point problem as a test of the null hypothesis

$$H_0 : \pi_1 = \pi_2 = \ldots = \pi_{425}$$

versus the alternative

$$H_1 : \pi_1 = \ldots = \pi_i \neq \pi_{i+1} = \ldots = \pi_{425},$$

where i is unknown. By examining chi-squared tests of homogeneity at the i^{th} row, $i = 2, 3, \ldots, 425$, Riba and Ginebra found persuasive evidence for a change-point in the multinomial probability vector between Chapters 371 and 382.

The Lancaster Chi-squared Decomposition

A common likelihood-ratio based test for homogeneity in multi-way contingency tables is the G^2 test, originally devised by Wilks [4,5] Let us assume multinomial sampling in an I x J contingency table, with cell probabilities $\{\pi_{ij}\}$, observed cell frequencies $\{n_{ij}\}$, and total sample size n. Wilks's likelihood ratio test for assessing the null hypothesis of independence (all $\pi_{ij} = \pi_{i+}\ \pi_{+j}$) is given by $G^2 = 2\sum_i\sum_j n_{ij} \log\left(\frac{n_{ij}}{\hat{n}_{ij}}\right)$

where $\left\{\hat{n}_{ij} = \frac{n_{i+}n_{+j}}{n}\right\}$ are the estimated expected frequencies under the null hypothesis of independence. When independence holds, G has an asymptotic chi-squared distribution with degrees of freedom (I-1)(J-1); and,as is well-known (Agresti [6]), G^2 and the conventional Pearson chi-squared statistic χ^2 are asymptotically equivalent. These statistics are valid tests of homogeneity under independent multinomial sampling schemes (Agresti [6]),

Lancaster [2] described partitions of the overall chi-squared statistic for independence χ^2 (with $(I\text{-}1)(J\text{-}1)$ degrees of freedom) into asymptotically independent components. Irwin [3] showed that the partitions follow directly from decompositions of the underlying quadratic forms of the chi-squared statistics with Helmert matrices. The most refined partition involves decomposition into $(I\text{-}1)$(J-1) independent components via double Helmert transformation, each component with 1 degree of freedom. As Agresti [6] has demonstrated, the partitioning technique extends immediately to G^2 formulations of the independence test.

An alternative procedure for detecting a change-point in a sequence of independent multinomials is immediately afforded from a Lancaster partition, as follows. The overall G^2 statistic for homogeneity in Table 1 [with value 8284.7, approximately chi-squared with $(I\text{-}1)$(J-1) = 424*9 = 3816 degrees of freedom, $p < 10^{-8}$] can be decomposed into I-1 = 424 asymptotically independent chi-squared components, each with J-1 = 9 degrees of freedom, from testing homogeneity of row 1 vs. row 2, row 1 + row 2 vs. row 3, … , row 1 + … + row 424 vs. row 425. [Indeed, in the general case, the algebraic equivalence of the overall G^2 statistic and the sum of the I-1 independent component statistics, each with I-1 degrees of freedom, is almost immediate].

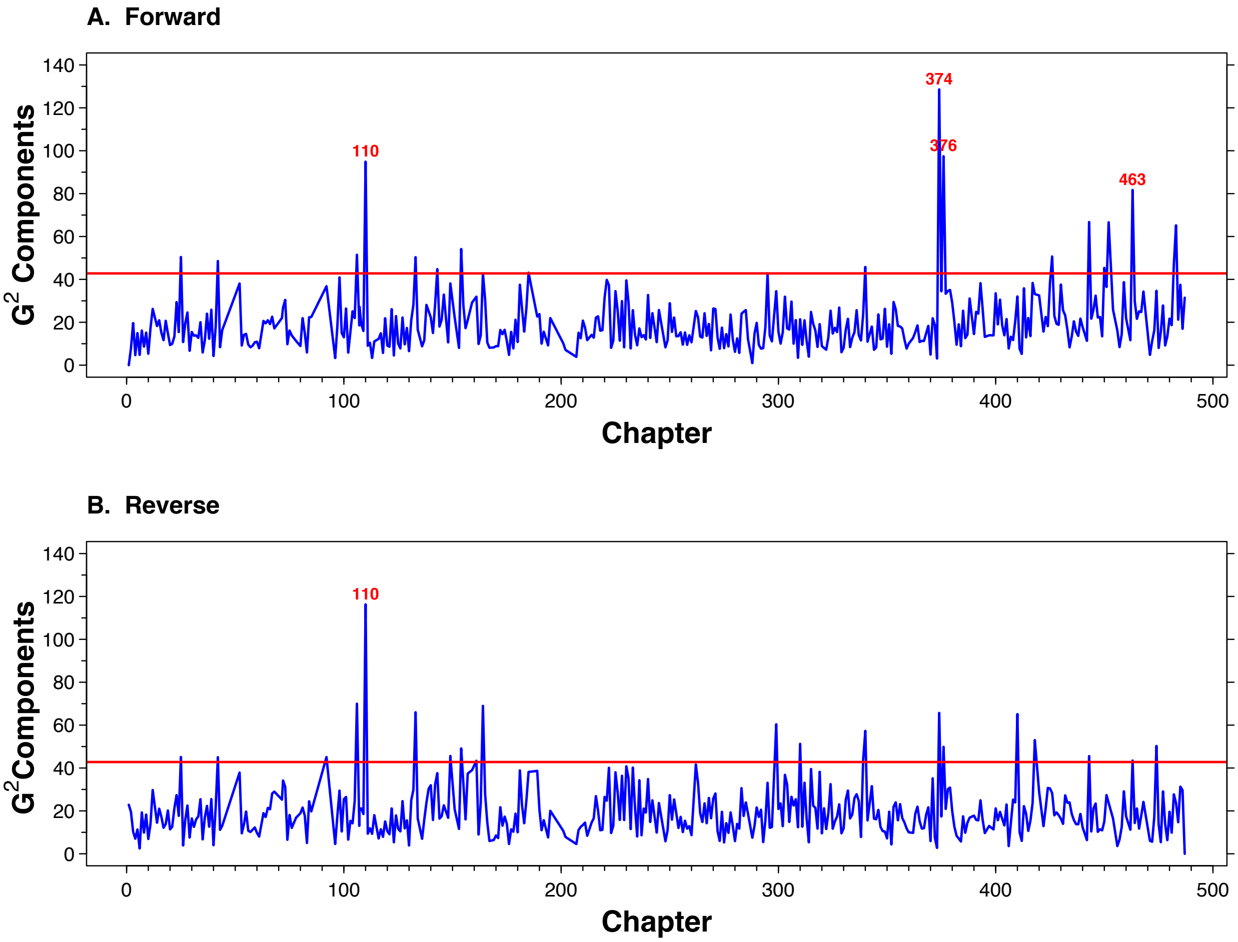
Note, however, that there is a second decomposition, obtained by sequentially testing row 425 vs. row 424, row 425 + row 424 vs. row 423, … , row 425 + … + row 2 vs. row 1. [There is distributional invariance with respect to reversal of the order of partitioning, a situation commonly found in change-point settings with a single transition (e.g., Koziol [7])]. This decomposition in the reverse order also provides a valid test of the underlying null hypothesis of homogeneity of the row probabilities, but there is no a priori reason to expect the two decompositions to yield identical inferences, especially if the null hypothesis of homogeneity is contraindicated.

In Figure 1 we plot the forward components and the reverse components from these decompositions as a function of row position. Perhaps the most striking finding in Figure 1 is the vast number of components that achieve conventional "statistical significance": in the forward direction, there are 44, 33, and 19 components exceeding the respective Bonferroni 0.05, 0.01, and 0.001 boundaries [33.31, 37.27, and 42.79 respectively]; in the reverse direction, the cardinalities are 42, 35, and 21 respectively. In the forward direction, the four largest components occur at Chapter 374 [vs Chapters 1+…+373,], Chapter 376, Chapter 110, and Chapter 463, with respective component values 128.6, 97.4, 94.9, and 81.7. All other components in the forward direction are less than 70. In the reverse direction, there is only one component exceeding 70: this occurs at Chapter 110 [vs Chapters 111+…+487], with the value 116.3. In terms of word length, at least, Chapter 110 appears to be an outlier relative to the surrounding chapters. As for location of a change-point, results are more equivocal. In the forward direction, there seems to be some support for a change-point at or around Chapter 374. There is no unique region comparable to this location with the reverse components.

Cluster Analysis

As noted above, inspection of the G^2 components clearly demonstrates that the word lengths in the various chapters of *Tirant lo Blanc* are quite heterogeneous. We therefore undertook a more exploratory analysis of word lengths, with roots in cluster analytic procedures.

We first performed a k-means cluster analysis of the word lengths, using Euclidean distances between observed cell frequencies in the various rows as the distance metric, and focusing on the case k = 2. This yielded a partition of the 425 chapters into subgroups of cardinality 266 and 159. Generally, early chapters are in the larger subgroup, later chapters in the smaller subgroup, but with much overlap. We attempted to minimize the overlap and thereby impute a change-point in group membership with the following procedure. We began with the two subgroups identified by the k-means clustering algorithm. Now, sequentially, we selected a chapter number from [1-487]. Then for each subgroup, we tabulated the cardinality of the chapters in that subgroup that were numbered less than or equal to the selected chapter, and the cardinality of chapters numbered greater than the selected chapter. To compare these counts between the two subgroups, we formed the 2x2 table of counts, and then calculated the usual 1 degree of freedom chi-squared statistic for this table. The resulting chi-squared statistics are plotted in Figure 2. The chi-squared statistics should be indicative of the degree of separation between the two subgroups, with the maxima of the chi-squared statistics of particular interest. [Again, this is not a novel notion: see Miller and Siegmund [8], Koziol [9] and Boulesteix [10] for example.] In Figure 2, there is a relatively flat global maximum in chi-squared statistics between Chapters 363 and 372 [all values slightly larger than 60.0.] Interestingly, there are two smaller peaks: the chi-squared statistics have local maxima at Chapter 411 and Chapter 437, though the absolute magnitudes of these statistics are not nearly so large as at the global maximum.


A. Forward

B. Reverse

Figure 1. Independent components of the overall G^2 statistic for assessing homogeneity of word lengths across the 487 chapters of *Tirant lo Blanc*. The components are obtained from a Lancaster decomposition of G^2, either in the forward direction (A) or the reverse direction (B); see text for further details. For comparison purposes, a Bonferroni-corrected 0.001 level critical value is indicated in each panel.

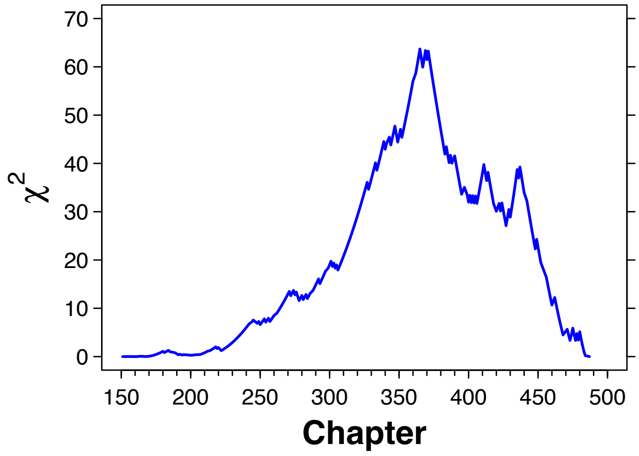


Figure 2. Chi-squared statistics for degree of separation between the subgroupings of chapters of *Tirant lo Blanc*, obtained from a k-means clustering algorithm with k = 2. The global maximum occurs between Chapters 363 and 372; there are local maxima at Chapters 411 and 437.

Discussion

We have provided complementary analyses to those of Ginebra and Riba concerning the existence of a change in authorship in *Tirant lo Blanc* based on word lengths. The components of G^2 provide strong evidence of the heterogeneity of word lengths across the chapters, but the evidence for change-points at particular locations is less decisive. The components of G^2 should be viewed as supplementary to the more fundamental likelihood maximization technique of Ginebra and Riba: in this setting, at least, the G^2 components seem quite sensitive to inter-chapter variability in word lengths, which is obscured in Ginebra and Riba's analysis.

The components flagged Chapter 110 of *Tirant lo Blanc* for its "anomalous" word lengths. Indeed, Chapter 110 does have proportionately fewer words of length 8, 9, or 10+, 3.82%, 1.27%, and 1.07% respectively, than the average of 4.88%, 3.15%, and 2.86%, but seems otherwise unexceptional. Nevertheless, the extent of heterogeneity uncovered by the components, rather than a deterrent to detection of a change-point, might instead by potentially exploited, as in the cluster analytic technique we used.

Our motivation with the cluster analysis was to obtain relatively homogeneous subgroups, by first ignoring the sequential ordering of the chapters. Then, we utilized a common statistical optimization procedure to repartition the subgroups sequentially, thereby cumulating evidence of a change-point. Though this procedure is somewhat ad hoc, it is reassuring that the location of a change-point determined thereby is largely in accord with that of Ginebra and Riba.

The change-point problem has been widely studied in the statistics literature, and there are a plethora of techniques available to the practitioner. In this particular setting, nonparametric techniques [e.g., Koziol and Wu, [11]] might well be less sensitive to inter-chapter heterogeneity than parametric procedures. Also, Girón, Ginebra, and Riba [12] have presented an elegant Bayesian analysis of the multinomial change-point problem, which provided strong evidence of a change-point near chapters 371 and 382. As Girón, Ginebra, and Riba remarked, a particular advantage of the hierarchical Bayesian approach would be robustness to the assumption of conditionally independent, multinomially distributed vectors of word length counts across the chapters. Indeed, this approach may well accommodate the inter-chapter variability that we have uncovered.

Lastly, we remark that although our initial motivation was to address the Riba and Ginebra application of a multinomial change-point model in a linguistics context, we note that multinomial change-point models are also prevalent in more biostatistical and bioinformatic contexts. Perhaps the canonical examples of the latter would be multinomial models for DNA nucleotide bases (cytosine (C), guanine (G), adenine (A), or thymine (T)) in genome sequences, and for amino acids (20 in all) in protein sequences. Changes in base or amino acid frequencies oftentimes would have physiological implications. For example, one might profitably adapt the techniques introduced by Riba and Ginebra to DNA sequences so as to detect changes in CG frequency, indicative of CG islands, that is, short stretches of DNA in which the frequency of the CG sequence is inflated relative to other regions. [The CG island is also sometimes referred to as a CpG island, the p denoting a phosphodiester bond between the cytosine and the guanine nucleotides.] CG islands are sometimes taken as markers for active genes; or, conversely, the deficiency of CG sequences is a characteristic of inactive genes or non-coding regions. Multinomial models can also be derived for higher order combinations (e.g., k-mer patterns) of the basic building blocks of nucleotides and amino acids.

Acknowledgments

We thank Professors Ginebra and Riba for kindly making available their data, and for generously explaining specific details of their analyses.

References

1 Riba A, Ginebra J (2005) Change-point estimation in a multinomial sequence and homogeneity of literary style. Journal of Applied Statistics 32: 61-74.

2 Lancaster HO (1949) The derivation and partition of $\chi2$ in certain discrete distributions. Biometrika 36: 117-129.

3 Irwin JO (1949) A note on the subdivision of $\chi2$ into components. Biometrika 36: 130-134.

4 Wilks SS (1935) The likelihood test of independence in contingency tables. The Annals of Mathematical Statistics 6: 190-196.

5 Wilks SS (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann Math Stat 9: 60-62.

6 Agresti A (1990) Categorical Data Analysis. John Wiley New York.

7 Koziol JA (1996) A note on signed rank statistics for the changepoint problem. Statistics 27: 325-338.

8 Miller R, Siegmund D (1982) Maximally selected chi-square statistics. Biometrics 48: 1011-1016.

9 Koziol JA (1991) On maximally selected chi-square statistics. Biometrics 47: 1557-1561.

10 Boulesteix AL (2006) Maximally selected chi-square statistics for an ordinal variable. Biometrical Journal 48: 451-462.

11 Koziol JA , Wu S-CH (1996) A review of nonparametric tests for changepoint problems, with application to a recombinant drug therapy clinical trial. Journal of Biopharmaceutical Statistics 6: 425-441.

12 Girón J, Ginebra, J , Riba A (2005) Bayesian analysis of a multinomial sequence and homogeneity of literary style. American Statistician, 59: 19-30.