Computational Statistics and Data Analysis 97 (2016) 33–46

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A unifying approach to the shape and change-point hypotheses in the discrete univariate exponential family

Chihiro Hirotsu [a,*], Shoichi Yamamoto [b], Harukazu Tsuruta [c]

[a] *Collaborative Research Center, Meisei University, Tokyo, Japan*
[b] *RPM Co. Ltd., Tokyo, Japan*
[c] *Medical Informatics, AHS, Kitasato University, Sagamihara, Kanagawa, Japan*

ARTICLE INFO

Article history:
Received 16 June 2015
Received in revised form 10 October 2015
Accepted 25 November 2015
Available online 30 November 2015

Keywords:
Cumulative sum statistic
Goodness-of-fit test
Markov property
Maximal contrast test
Recursion formula
Restricted alternative

ABSTRACT

A unifying approach to the shape and change-point hypotheses is extended generally to a discrete univariate exponential family. The maximal contrast type tests are newly proposed for the convexity and sigmoidicity hypotheses based on the complete class lemma of tests for the restricted alternatives. Those tests are also efficient score tests for the slope change-point and inflection point models, respectively. For each of those tests the successive component statistics are the doubly- and triply-accumulated statistics. They have nice Markov properties for the exact and efficient recursion formulae for calculating the p-value. Further the sum of squares of the component statistics are developed as the cumulative chi-squared statistics for the directional goodness-of-fit tests of the dose–response model. Therefore the interesting applications will be in monitoring of spontaneous reporting of the adverse drug events and the directional goodness-of-fit tests.

1. Introduction

A general relationship between the shape restrictions and change-point models in the normal means has been discussed in Hirotsu and Marumo (2002). In particular the monotone hypothesis corresponds to the simple step change-point model whose components are the special case of monotony. For this monotone case the maximal standardized cumulative sum statistic, max acc. t, has been constructed based on the complete class lemma in Hirotsu (1982), which has been proved to be well behaved compared with an isotonic regression approach by Bartholomew (1959a,b), see Hirotsu et al. (1992, 2011). In fact the restricted maximum likelihood approach has no obvious optimality for a restricted parameter space like this. It is also too complicated to extend to non-normal distributions, to interaction problems and to other shape constraints such as convex and sigmoid. Actually in the book by Miller (1998) a choice of maximin contrast test by Abelson and Tukey (1963) is recommended for the isotonic inference to escape from the complicated calculations of the isotonic regression. However, such one degree of freedom contrast test cannot keep high power against the wide range of the monotone hypothesis even by a careful choice of the contrast. Instead we propose a more robust approach against the wide range of the restricted alternatives, which can be extended in a systematic way to convex and sigmoid hypotheses.

It has been shown that the max acc. t test is also an efficient score test for the step change-point hypothesis (Hirotsu, 1997). Then the convex and sigmoid restrictions are closely related to the slope change-point and inflection point models, for which the doubly- and triply-accumulated statistics are newly developed in this paper. Since Page (1954, 1961) cumulative

* Corresponding author. Tel.: +81 42 591 5111.
E-mail address: hirotsu@ge.meisei-u.ac.jp (C. Hirotsu).

http://dx.doi.org/10.1016/j.csda.2015.11.012
0167-9473/

sum statistics based approach has been widely developed in the statistical process control, see also Montgomery (2009) for various applications. More recently it is also extended to the field of environmental statistics as in Manly and Mackenzie (2003), for example. However, as stated in the review paper by Amiri and Allahyari (2012) most papers are assuming the step, linear trend or monotonic change, and it seems that the slope change-point and inflection point models are not popular in those fields. One should refer also to Hawkins (1977) and Worsley (1986) for the likelihood ratio test approach to the change-point hypothesis as the previous work to the present paper. Thus the present paper gives a unified approach to the shape and change-point hypotheses which have been investigated in two different streams of statistics. The unifying approach is useful also in practice since in the applications it is usually of importance not only to detect a general tendency but also to detect a particular change-point as seen in the examples of Section 11.

More precisely, each corner vector of the convex polyhedral cone defined by the convexity hypothesis corresponds to a component of the slope change-point model. The convexity hypothesis

$$C:\ \boldsymbol{L}_K^{*\prime}\boldsymbol{\mu} \geq 0, \quad \text{with at least one inequality strong,} \tag{1}$$

is defined by the second order differential matrix

$$\boldsymbol{L}_K^{*\prime} = \begin{bmatrix} \frac{1}{x_2 - x_1} & \frac{1}{x_1 - x_2} + \frac{1}{x_2 - x_3} & \frac{1}{x_3 - x_2} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{x_3 - x_2} & \frac{1}{x_2 - x_3} + \frac{1}{x_3 - x_4} & \frac{1}{x_4 - x_3} & 0 & \cdots & 0 \\ & & & & & \cdots & \\ 0 & 0 & 0 & \cdots & \frac{1}{x_{K-1} - x_{K-2}} & \frac{1}{x_{K-2} - x_{K-1}} + \frac{1}{x_{K-1} - x_K} & \frac{1}{x_K - x_{K-1}} \end{bmatrix}_{(K-2)\times K},$$

where x_i denotes the time or location of the ith event, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)'$ a vector of unknown parameters and $\boldsymbol{A}'$ the transpose of a matrix $\boldsymbol{A}$. It forms a convex polyhedral cone with the corner vectors $\boldsymbol{L}_K^*(\boldsymbol{L}_K^{*\prime}\boldsymbol{L}_K^*)^{-1}$ or in other words each $\boldsymbol{\mu}$ satisfying (1) is expressed by a unique and positive linear combination of the columns of $\boldsymbol{L}_K^*(\boldsymbol{L}_K^{*\prime}\boldsymbol{L}_K^*)^{-1}$, that is, $\boldsymbol{\mu} = \boldsymbol{L}_K^*(\boldsymbol{L}_K^{*\prime}\boldsymbol{L}_K^*)^{-1}\boldsymbol{c}$ with $\boldsymbol{c} \geq 0$. Thus the matrix $\boldsymbol{L}_K^*(\boldsymbol{L}_K^{*\prime}\boldsymbol{L}_K^*)^{-1}$ is more directly concerned with the convexity hypothesis than the defining matrix $\boldsymbol{L}_K^{*\prime}$. On the other hand as shown in Section 2 each column of $\boldsymbol{L}_K^*(\boldsymbol{L}_K^{*\prime}\boldsymbol{L}_K^*)^{-1}$ represents a component of the slope change-point model M_k at $x = x_{k+1}$,

$$M_k:\ \begin{cases} \dfrac{\mu_2 - \mu_1}{x_2 - x_1} = \dfrac{\mu_3 - \mu_2}{x_3 - x_2} = \cdots = \dfrac{\mu_{k+1} - \mu_k}{x_{k+1} - x_k} = \beta_k, \\ \dfrac{\mu_{k+2} - \mu_{k+1}}{x_{k+2} - x_{k+1}} = \dfrac{\mu_{k+3} - \mu_{k+2}}{x_{k+3} - x_{k+2}} = \cdots = \dfrac{\mu_K - \mu_{K-1}}{x_K - x_{K-1}} = \beta_k^*, \end{cases} \quad k = 1, \ldots, K-2. \tag{2}$$

The null hypothesis $H_0:\ \boldsymbol{L}_K^{*\prime}\boldsymbol{\mu} = 0$ corresponds to $M_0:\ \beta_k = \beta_k^* \equiv \beta_1$ or equivalently to a linear regression model

$$M_0:\ \mu_i = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, K. \tag{3}$$

For the normal model, $\boldsymbol{y} \in N(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$, let $\boldsymbol{z}$ be $(\boldsymbol{L}_K^{*\prime}\boldsymbol{L}_K^*)^{-1}\boldsymbol{L}_K^{*\prime}\boldsymbol{y}$. Then the maximal standardized element z_m of $\boldsymbol{z}$ has been derived from a complete class lemma for the convexity hypothesis (1) and shown to be an appropriate statistic also for testing the slope change-point model (2) (Hirotsu, 1982; Hirotsu and Marumo, 2002). Further a very efficient algorithm for probability calculation of z_m for any K has been proposed based on the second order Markov property of the successive elements of $\boldsymbol{z}$. Those properties have been derived only by the covariance structure in case of the normal model since the key vector $\boldsymbol{z}$ is a linear function of $\boldsymbol{y}$. For a general exponential family the key vector $\boldsymbol{z}$ becomes $\boldsymbol{z} = (\boldsymbol{L}_K^{*\prime}\boldsymbol{L}_K^*)^{-1}\boldsymbol{L}_K^{*\prime}\widehat{\boldsymbol{v}}_0$ with $\widehat{\boldsymbol{v}}_0$ an efficient score vector evaluated at the null hypothesis as shown in Section 2, and it looks formidable to develop an exact analysis since an explicit expression of $\boldsymbol{z}$ is so complicated. In this paper, however, the ideas for the normal model are extended generally to the discrete univariate exponential family and an exact analysis is developed. Firstly in Section 3 it is shown that the component statistics forming $\boldsymbol{z}$ are essentially the doubly-accumulated statistics and the maximal standardized component of them is proposed as the test statistic. This is a very important finding for extending the normal theory to a general discrete univariate exponential family. In Section 4 an approach to constructing the joint conditional distribution of the component statistics is developed given the complete sufficient statistics under the null model for developing a similar test. Such a conditional distribution given plural conditional variables is quite unfamiliar for the discrete random variables. The second order Markov property of the successive component statistics and the factorization of their joint distribution into the products of conditional probabilities are shown. In Section 5 a bottom up procedure is developed for determining the probability function definitely and some inequalities are introduced for executing the recursion formula efficiently. Then Section 6 is devoted for calculating the tail probability and moments of the maximal contrast test statistic. As a promising statistic for a directional goodness-of-fit test of a dose–response model the cumulative chi-squared is introduced in Section 7, which is defined as the sum of squares of the standardized elements of the key vector. In Section 8 some power comparisons are made to show the robustness of the proposed methods. In Section 9 these ideas are extended to testing the sigmoidicity and inflection point model. Then the basic statistics are triply-accumulated statistics and the related statistical property is the third order Markov. In Section 10 a more direct method for calculating p-value is given for the non-explosive sequence for both of the maximal and cumulative chi-squared statistics.