Statistical Methodology 32 (2016) 77–90

Contents lists available at ScienceDirect

Statistical Methodology

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

Homogeneity testing via weighted affinity in multiparameter exponential families

Alexander Katzur, Udo Kamps *

Institute of Statistics, RWTH Aachen University, D-52056 Aachen, Germany

ARTICLE INFO

Article history:
Received 25 November 2015
Received in revised form
7 April 2016
Accepted 8 April 2016
Available online 16 April 2016

Keywords:
Homogeneity test
Discrimination
Exponential family
Affinity
Sequential order statistics
Non-central chi-squared-distribution

ABSTRACT

Based on stochastically independent samples with underlying density functions from the same multiparameter exponential family, a weighted version of Matusita's affinity is applied as test statistic in a homogeneity test of identical densities as well as in a discrimination problem. Asymptotic distributions of the test statistics are stated, and the impact of weights on the deviation of actual and required type I error for finite sample sizes is examined in a simulation study.

1. Introduction

As a measure of similarity or dissimilarity of probability distributions, Matusita's affinity measure (cf., e.g., Matusita [13–15]) has been considered in the literature, such as in classification and discrimination problems for multivariate normal distributions. For two or more distributions on $\Omega \subset \mathbb{R}^n$ with distribution functions $F_1, \dots, F_J$ and respective density functions $f_1, \dots, f_J$ with respect to (w.r.t.) measure ν, their affinity ρ_J is defined by

$$\rho_J(F_1, \dots, F_J) = \int_\Omega \left(\prod_{j=1}^{J} f_j(x) \right)^{1/J} d\nu(x). \tag{1}$$

It is known that $0 \le \rho_J(F_1, \dots, F_J) \le 1$, where the upper bound is attained iff all the distribution functions coincide (almost everywhere (a.e.) w.r.t. ν). Matusita [13,14] derived a representation for

* Corresponding author.
E-mail address: Udo.Kamps@rwth-aachen.de (U. Kamps).

http://dx.doi.org/10.1016/j.stamet.2016.04.002
1572-3127/

the affinity of two multivariate normal distributions and analyzed its distribution when replacing unknown parameters by arithmetic means and (or) the sample covariance matrix. Moreover, he obtained a representation for the affinity of $J \geq 2$ multivariate normal distributions. He also suggested a homogeneity test on the basis of ρ_J and applied it to the multivariate normal case. Toussaint [19] replaced the geometric mean of density functions in (1) by a weighted geometric mean with vector $\omega = (\omega_1, \dots, \omega_J)'$ of weights, $\omega_j > 0,\ 1 \leq j \leq J$, and $\sum_{j=1}^{J} \omega_j = 1$, i.e.,

$$\rho_J(F_1, \dots, F_J \mid \omega) = \int_\Omega \prod_{j=1}^{J} \big(f_j(x)\big)^{\omega_j} \, d\nu(x), \qquad (2)$$

taking values in the closed unit interval.

Concerning asymptotic distributions of respective test statistics, Garren [9] considered Matusita's affinity (1) in exponential families, and derived its asymptotic distribution based on plug-in maximum likelihood estimators (MLEs), provided that the numbers of observations in the J samples grow equally fast. He then applied his results to propose a simple and a two-sided hypothesis test in the one-sample situation as well as a decision rule for discriminating between two unknown distributions based on three samples. Previously, in a very general setting, Zografos [20] considered parametric probability spaces and, under certain regularity conditions, he obtained an asymptotic distribution for f-dissimilarities, which include the affinity measure, with plug-in MLEs using asymptotic normality of the MLEs. f-dissimilarities have been used in several papers for testing statistical hypotheses (cf., e.g., Morales et al. [16]).

In the present paper, we aim at examining some tests for the parameters of exponential families by means of the weighted affinity (2), which, in this situation, allows for a closed form representation. This turns out to be helpful when dealing with small sample sizes. We present a homogeneity test, a simple and a two-sided test as well as a discrimination approach for a two class classification problem. For the distribution of the test statistic of the homogeneity test under the alternative, Zografos's results can be applied here, since our exponential family setting as well as the weighted affinity are contained in his general set-up. Under the null hypothesis of homogeneity, we find distributional convergence of the test statistic to a distribution free expression that only depends on known constants and multivariate normal random vectors with zero mean and identity covariance matrix. Moreover, as a by-product, it turns out that Garren's results for the case $F_1 = \cdots = F_J$ are actually distribution free, and this applies as well to the weighted affinity. Simulations indicate that Garren's assumption of equally fast growing samples can lead to a major deviation of the actual type I error rate from the required type I error rate based on the asymptotic distribution of the test statistic, if the finite sample sizes are actually unequal (see Section 4). We provide two modifications to overcome this problem, weights $\omega_1, \dots, \omega_J$ are introduced and a relaxed assumption on the growth behavior of the sample sizes is considered, i.e., numbers of observations in different samples do not need to grow equally fast. The impact of those modifications on the actual type I error is illustrated in a simulation study.

The present work is structured as follows. In Section 2, exponential families are introduced and a closed form expression for the weighted affinity of several distributions from the same exponential family of distributions is provided. In Section 3, the hypothesis tests and their test statistics that are based on the affinity (2) are examined. The asymptotic distribution of the homogeneity test statistic under both hypotheses is derived in Section 3.1. Furthermore, if the alternative is replaced by a contiguous alternative hypotheses, an asymptotic distribution can be obtained by means of the non-central χ^2-distribution. The asymptotic distributions of the test statistic of the simple and the two-sided test and for the statistic of the discrimination approach are provided in Sections 3.2–3.4, respectively. In Section 4, the homogeneity test is analyzed by means of a simulation study. Special attention is put on the impact of the modifications, i.e., the weights and the relaxed growth assumption, as well as on the usage of the asymptotic distribution with small and moderate sample sizes.

2. Exponential families and weighted affinity

We will use the following notation of an exponential family.