Computational Statistics and Data Analysis 99 (2016) 115–130

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

On point estimation of the abnormality of a Mahalanobis index

Fadlalla G. Elfadaly [a,b,*], Paul H. Garthwaite [a], John R. Crawford [c]

[a] Department of Mathematics and Statistics, The Open University, UK

[b] Department of Statistics, Faculty of Economics and Political Science, Cairo University, Egypt

[c] School of Psychology, King's College, University of Aberdeen, UK

HIGHLIGHTS

- We seek an estimate of the population proportion with extreme Mahalanobis index.
- Nine point estimates are examined in an extensive simulation study.
- Maximum likelihood estimates have substantial bias.
- Methods based on polynomial approximations give low bias, but can be out-of-range.
- An adapted median estimator that always gives sensible estimates is proposed.

ARTICLE INFO

Article history:
Received 23 June 2015
Received in revised form 27 October 2015
Accepted 21 January 2016
Available online 29 January 2016

Keywords:
Bernstein polynomials
Mahalanobis distance
Median estimator
Plug-in maximum likelihood
Quadrature approximation
Unbiased estimation

ABSTRACT

Mahalanobis distance may be used as a measure of the disparity between an individual's profile of scores and the average profile of a population of controls. The degree to which the individual's profile is unusual can then be equated to the proportion of the population who would have a larger Mahalanobis distance than the individual. Several estimators of this proportion are examined. These include plug-in maximum likelihood estimators, medians, the posterior mean from a Bayesian probability matching prior, an estimator derived from a Taylor expansion, and two forms of polynomial approximation, one based on Bernstein polynomial and one on a quadrature method. Simulations show that some estimators, including the commonly-used plug-in maximum likelihood estimators, can have substantial bias for small or moderate sample sizes. The polynomial approximations yield estimators that have low bias, with the quadrature method marginally to be preferred over Bernstein polynomials. However, the polynomial estimators sometimes yield infeasible estimates that are outside the 0–1 range. While none of the estimators are perfectly unbiased, the median estimators match their definition; in simulations their estimates of the proportion have a median error close to zero. The standard median estimator can give unrealistically small estimates (including 0) and an adjustment is proposed that ensures estimates are always credible. This latter estimator has much to recommend it when unbiasedness is not of paramount importance, while the quadrature method is recommended when bias is the dominant issue.

* Correspondence to: Department of Mathematics and Statistics, The Open University, Milton Keynes, MK7 6AA, UK. Tel.: +44 7599946166; fax: +44 1908654355.

E-mail address: f.elfadaly@open.ac.uk (F.G. Elfadaly).

http://dx.doi.org/10.1016/j.csda.2016.01.014

1. Introduction

The Mahalanobis distance is frequently used in multivariate analysis as a statistical measure of distance between a vector of scores for a single case and the mean vector of the underlying population or a sample of data. It was developed by Mahalanobis (1936) as a distance measure that incorporates the correlation between different scores. See also DasGupta (1993). The Mahalanobis distance of a vector $\mathbf{x}$, of say ν_1 variables (scores), from a population mean $\boldsymbol{\mu}$ is defined as

$$\Delta = \sqrt{(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \tag{1}$$

where $\boldsymbol{\Sigma}$ is the population covariance matrix. The square of the Mahalanobis distance, Δ^2, is sometimes referred to as the Mahalanobis index (Huberty and Olejnik, 2006, p. 271). If the population follows a multivariate normal distribution (MVN) and $\mathbf{x}$ is an observation from this same distribution, then the Mahalanobis index follows a central chi-square distribution on ν_1 degrees of freedom. In this paper, interest focuses on estimating P, the proportion of the population that gives a more unusual Mahalanobis index than $(\mathbf{x}^*-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}^*-\boldsymbol{\mu})$ where $\mathbf{x}^*$ is a specified vector, under the assumption that the population distribution is a MVN distribution. That is

$$P = \Pr\{(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) > (\mathbf{x}^*-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}^*-\boldsymbol{\mu})\}, \tag{2}$$

where $\mathbf{x} \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For example, $\mathbf{x}^*$ might be a patient's profile from a set of medical tests, when P would be the proportion of the population with a profile that is more unusual than that of the patient.

The corresponding Mahalanobis distance in a sample, of say n observations, is defined as

$$\widetilde{D} = \sqrt{(\mathbf{x}-\bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}})}, \tag{3}$$

where $\bar{\mathbf{x}}$ and $\mathbf{S}$ are the sample mean vector and sample covariance matrix, respectively. Under the assumption that $\mathbf{x}$ and the sample data are from the same MVN distribution, the sample Mahalanobis index ($\widetilde{D}^2$) is proportional to a central F distribution with ν_1 and $\nu_2 \equiv n-\nu_1$ degrees of freedom. See, for example, Mardia et al. (1979).

We were initially motivated by the need to estimate the abnormality of a single patient's profile in neuropsychology. The problem arises, for example, when psychologists need to assess how a patient with some brain disorder or a head injury is different from the general population or some particular subpopulation. This assessment is usually based on the patient's scores in a set of tests that measure different traits or abilities. The abnormality of the case's profile of scores can then be expressed in terms of the Mahalanobis index between this profile and the mean of the normative population or normative sample. The degree of abnormality is measured by

$$\widehat{P} = \Pr\{(\mathbf{x}-\bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}}) > (\mathbf{x}^*-\bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}^*-\bar{\mathbf{x}})\}, \tag{4}$$

where $\mathbf{x}^*$ is the case's profile and is treated as a fixed quantity.

A Hotelling's T^2 significance test for testing whether the case could belong to the normative population is proposed in Huizenga et al. (2007). Their test is based on the central F distribution to which the Hotelling's test statistic is proportional. Crawford et al. (2016) give a confidence interval for the probability (P) of getting a more extreme profile than the case. The confidence interval is based on a non-central F distribution with a non-centrality parameter that is proportional to the case's Mahalanobis index. The confidence intervals are correct, in that their coverage levels equal the nominal confidence level exactly. In contrast, the p-value from the Hotelling's T^2 test provides an obvious point estimator of P, but it is biased. Indeed, the problem of finding an unbiased estimator of P has not been resolved.

Here we consider a number of obvious estimators of P and propose some new, less obvious estimators. The bias and mean square error of all the estimators are compared in extensive simulations. No estimator is uniformly better than all alternatives, but a small selection of the estimators is clearly to be preferred. As well as bias and mean square error, other criteria and desirable qualities in an estimator are also considered. In this paper, no distributional assumptions are made about the source of $\mathbf{x}^*$, other than when testing whether $\mathbf{x}^*$ could be the profile of a member of the normative population.

The need to estimate the value of P for Mahalanobis distances does not only arise in psychology. In the literature, the commonly used estimates of P are the p-value computed from the chi-square distribution of the sample Mahalanobis index, or the p-value from the central F distribution associated with Hotelling's T^2 test. For example, in remote sensing image analysis, Foody (2006) was interested in measuring the closeness of an image pixel to a single class centroid. For that, he used the Mahalanobis distance and converted the calculated Mahalanobis distance, of a particular image pixel from a specified class centroid, to its associated p-value from the chi-square distribution. He then interpreted the p-value as the probability of obtaining a Mahalanobis distance as extreme as that observed for a particular pixel with respect to a specified class, thus effectively equating the p-value to P.

In environmental and health science, Liu and Weng (2012) used Mahalanobis distance in public health studies to enhance the resolution of satellite imagery. They conducted a spatial–temporal analysis of West Nile Virus outbreak in Los Angeles in 2007 using sensing variables and infective mosquito surveillance records. Mahalanobis distance was used to identify and map the risk areas where habitat was suitable for infective mosquitoes. Liu and Weng (2012) calculated the distance between a vector of environmental variables and the mean vector of environmental factors at the closest locations of mosquito infections. Locations with smaller values of Mahalanobis distances indicated a more favorable habitat for the mosquitoes