Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Parametric methods for confidence interval estimation of overlap coefficients

Dan Wang [a,b], Lili Tian [a,*]

[a] Department of Biostatistics, University at Buffalo, Buffalo, NY 14214, USA

[b] Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN 46285, USA

ARTICLE INFO

Article history:
Received 5 January 2016
Received in revised form 12 August 2016
Accepted 19 August 2016
Available online xxxx

Keywords:
High-throughput platforms
Genomic biomarker
Overlap coefficient
Generalized inference
Parametric bootstrapping
Mixture Gaussian
EM algorithm

ABSTRACT

Overlap coefficient (*OVL*), the proportion of overlap area between two probability distributions, is a direct measure of similarity between two distributions. It is useful in microarray analysis for the purpose of identifying differentially expressed biomarkers, especially when data follow multimodal distribution which cannot be transformed to normal. However, the inference methods about *OVL* are quite sparse. This article proposes two methods, a generalized inference (*GI*) approach and a parametric bootstrapping (*PB*) method, are proposed to construct confidence intervals of *OVL* under the assumption of normality. In conjunction with the *EM* algorithms, these methods are extended to mixture Gaussian (*MG*) distributions. The performances of these methods are evaluated empirically under a variety of distributions including normal, gamma and mixture Gaussian. At last, the proposed approaches are applied to a published microarray dataset from a gene expression study of three most prevalent adult lymphoid malignancies.

1. Introduction

Let X_1 and X_2 denote the continuous response variables for two user-defined groups (e.g. case and control) respectively, and let f_{X_1} and f_{X_2} be the corresponding probability densities. The overlap area under the curves of f_{X_1} and f_{X_2} (denoted as *OVL*) is

$$OVL = \int_{-\infty}^{\infty} \min[f_{X_1}(x|\Theta_1), f_{X_2}(x|\Theta_2)]dx, \tag{1}$$

where Θ_1 and Θ_2 stand for parameter spaces for $f_{X_1}(X_1, \Theta_1)$ and $f_{X_2}(X_2, \Theta_2)$, respectively. If the distributions are discrete, *OVL* can be calculated by replacing the integral with a summation. The *OVL* is scaleless with value ranging from 0 (i.e. two distributions being completely distinct) to 1 (i.e. two distributions being identical). *OVL* directly measures the similarity (or difference) between two distributions. Hence, it can serve as a diagnostic measure which is sensitive to any differences between two distributions despite the structures of the underlying distributions.

The concept of *OVL* was first proposed by Weitzman (1970), and it was generalized to n dimensions by Bradley et al. (1982). Since then, *OVL* has been widely used in various practical applications, such as quantitative ecology (Gastwirth, 1975), cluster analysis in mathematical geology (Sneath, 1977), stress–strength models of reliability analysis (Ichikawa, 1993),

[*] Corresponding author.
E-mail address: ltian@buffalo.edu (L. Tian).

http://dx.doi.org/10.1016/j.csda.2016.08.013
0167-9473/© 2016 Elsevier B.V. All rights reserved.

electromyographic assessment of muscular asymmetry (Ferrario et al., 2000), and treatment assessment in clinical trials (Mizuno et al., 2005).

Recently, *OVL* was introduced to genomic study by Silva-Fortes et al. (2012). The resurgence of *OVL* in genomic studies is attributable to the fact that it is a more convenient and proper diagnostic measure compared to other traditional diagnostic measures, such as *AUC* (area under receiver operating curve). High-throughput technologies such as microarray have revolutionized genomic studies in the past decade and the massive amount of data generated by these high-throughput methods poses a variety of challenges to existing statistical methods. Since a major goal of genomics is to identify genes significantly differentially expressed in diseased versus healthy groups, it is of paramount significance to find a diagnostic index which is sensitive to any differences between diseased and healthy groups. However, traditional diagnostic indices such as *AUC* mainly focus on examining the difference of means between groups, and fail to capture other possible differences, e.g. shapes between two distributions. For instance, bimodal or multimodal gene expression data commonly exist in genomic studies due to differing molecular subtypes or unknown subclasses within a population of cells. For such data, the diseased and healthy groups can differ dramatically, while having the same mean (Silva-Fortes et al., 2012; Parodi et al., 2008). Regardless of the underlying distributions, *OVL* serves as a convenient and proper measure of the diagnostic ability of biomarkers while traditional diagnostic indices such as *AUC* might fail. More details about *OVL* versus *AUC* can be found in Appendix A.

The reason that *OVL* has not been widely used as a diagnostic measure is partially due to the lack of methods for confidence interval estimation of *OVL*. Currently, both existing parametric and nonparametric methods for *OVL* inference have certain limitations. Parametric methods (Al-Saidy et al., 2005; Al-Saleh and Samawi, 2007; Samawi and Al-Saleh, 2008; Chaubey et al., 2008; Helu and Samawi, 2011; Reiser and Faraggi, 1999; Mulekar and Mishra, 2000; Mizuno et al., 2005) have not yet been applied to general Gaussian (i.e. without equal mean or equal variance condition) or mixture Gaussian distributions, and non-parametric methods only focused on the point estimator for the cases with large sample sizes (Clemons and Bradley, 2000; Mizuno et al., 2005; Schmid and Schmidt, 2006; Anderson et al., 2012). To popularize *OVL* as a diagnostic index, it is important to develop methods for estimating the confidence intervals of *OVL*.

The goal of this paper is to propose methods for confidence interval estimation of *OVL* under a variety of distributions, including normal, normal transformed and multimodal distributions. In Section 2, we propose a generalized inference (*GI*) method and parametric bootstrapping (*PB*) method to construct the confidence interval estimation of *OVL* under normality for original and transformed data. Section 3 deals with mixture normal distributions by combining *EM* algorithms with the *GI* and *PB* methods. Section 4 presents the details of simulation study to check the performance of the proposed method. In Section 5, the proposed methods are applied to a published microarray dataset from a gene expression study of three most prevalent adult lymphoid malignancies. Section 6 concludes the paper with a discussion. Appendix A presents a brief review of *AUC* as well as a comparison between *OVL* and *AUC*.

2. Under normality: original and transformed data

This section presents two parametric approaches, i.e. a generalized inference approach (*GI*) and a parametric bootstrapping approach (*PB*) for confidence interval estimation based on normality. Let $X_{11}, X_{12}, \ldots, X_{1n_1}$ and $X_{21}, X_{22}, \ldots, X_{2n_2}$ denote the n_1 and n_2 observations for the control (X_1) and case (X_2) groups, respectively. Assume $X_{ij} (i = 1, 2; j = 1, 2, \ldots, n_i)$ follow normal distribution with mean μ_i and variance σ_i^2. The parameter space Θ_i in formula (1) is (μ_i, σ_i^2) where $i = 1, 2$. Hence *OVL* can be calculated as

$$OVL = \int \min[f_{X_1}(x|\mu_1, \sigma_1^2), f_{X_2}(x|\mu_2, \sigma_2^2)]dx. \tag{2}$$

When normality cannot be justified for original data but can be achieved via a monotonic transformation such as Box–Cox transformation (Box and Cox, 1964), the proposed methods can be applied to the transformed data due to the fact that *OVL* is invariant under monotonic transformation. This approach has been found to be useful in *ROC* analysis for a wide variety of scenarios (Molodianovitch et al., 2006; Fluss et al., 2005; Zou and Hall, 2000; Faraggi and Reiser, 2002; Schisterman et al., 2004). To be specific, a power transformation of the Box–Cox type is

$$X_i^{(\lambda)} = \begin{cases} \dfrac{X_i^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(X_i) & \lambda = 0, \end{cases}$$

where it is assumed that $X_i^{(\lambda)} \sim N(\mu_i, \sigma_i^2)$. For the transformed data, the appropriate likelihood function can be constructed as follows:

$$l(\mu_1, \mu_2, \sigma_1, \sigma_2, \lambda|X_1, X_2) = l(\mu_1, \sigma_1, \lambda|X_1) + l(\mu_2, \sigma_2, \lambda|X_2),$$

where $l(\mu_i, \sigma_i, \lambda|X_i) = -\frac{1}{2}\log(2\pi\sigma_i^2) - \sum_{j=1}^{n_i} \frac{X_{ij}^\lambda - \mu_i}{2\sigma_i^2} + (\lambda - 1)(\sum_{j=1}^{n_i} \log(X_{ij}))$. The parameters $(\lambda, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$ can be estimated using the maximum likelihood estimation procedure, and the transformed data will be used for confidence interval estimation of *OVL*.

Please cite this article in press as: Wang, D., Tian, L., Parametric methods for confidence interval estimation of overlap coefficients. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.08.013