Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2006) 570–586
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Interval estimation in a finite mixture model: Modeling P-values in multiple testing applications

Qinfang Xiang[a], Jode Edwards[b], Gary L. Gadbury[a,*]

[a] *Department of Mathematics and Statistics, University of Missouri – Rolla, Rolla, MO 65409, USA*
[b] *USDA ARS, Department of Agronomy, Iowa State University, Ames, IA, USA*

Received 20 February 2005; received in revised form 6 September 2005; accepted 16 November 2005
Available online 9 December 2005

Abstract

The performance of interval estimates in a uniform-beta mixture model is evaluated using three computational strategies. Such a model has found use when modeling a distribution of P-values from multiple testing applications. The number of P-values and the closeness of a parameter to the boundary of its space both play a role in the precision of parameter estimates as does the "nearness" of the beta-distribution component to the uniform distribution. Three computational strategies are compared for computing interval estimates with each one having advantages and disadvantages for cases considered here.
Published by Elsevier B.V.

Keywords: Bootstrap; Gene expression; Hessian; Interval estimation; MCMC; Microarray; MLE; Uniform beta mixture

1. Introduction

A finite mixture model (McLachlan and Peel, 2000) of the form

$$f(x;\theta) = \lambda f_0(x;\theta_0) + (1-\lambda) f_1(x;\theta_1) \tag{1}$$

has recently been used in multiple testing applications where f_0 models a statistic for which a null hypothesis is true, and f_1 is the model for which an alternative is true (Lee et al., 2000; Efron and Tibshirani, 2002; Allison et al., 2002; Pounds and Morris, 2003; Gannoun et al., 2004). If the statistic is transformed into a P-value using the appropriate reference distribution (Sackrowitz and Samuel-Cahn, 1999), f_0 is the uniform density on (0, 1). The distribution of a P-value when the alternative hypothesis is true, f_1, depends on the effect size, sample size, and the distribution of the test statistic used to define the P-value (Hung et al., 1997). When many P-values are observed from multiple tests, those for which the alternative hypothesis is true should be smaller than expected from a uniform distribution, a fact used by Schweder and Spjøtvoll (1982) and more recently by Delongchamp et al. (2004). Parker and Rothenberg (1988) found that a mixture of a uniform distribution and non-uniform beta distributions was sufficiently flexible to model the distribution of P-values from multiple hypothesis tests.

This paper considers a model of the form,

$$f(p_i) = \lambda + (1-\lambda)\beta(p_i; r, s), \quad p_i \in (0,1),\ i = 1,\ldots,k, \tag{2}$$

* Corresponding author. Tel.: +1 573 341 4648; fax: +1 573 341 4741.
E-mail address: gadburyg@umr.edu (G.L. Gadbury).

0167-9473/$ - see front matter Published by Elsevier B.V.
doi:10.1016/j.csda.2005.11.011

where p_i is a P-value for the ith hypothesis test (out of k tests), $\lambda \in (0, 1)$ is a weight on the uniform component, and $\beta(p; r, s)$ is a beta-probability density function (pdf) with parameters r and s. Such a model has been used by others to model the distribution of P-values from microarray experiments (e.g., Allison et al., 2002; Pounds and Morris, 2003; Gadbury et al., 2004). A typical microarray experiment tests for differential genetic expression across two or more treatment groups. Multiple testing issues arise due to the many thousands of genes that are simultaneously tested for differential expression (see Knudsen, 2002 or Speed, 2003, for more details on microarrays). The parameter λ is thus interpreted as the proportion of genes that have no differential expression due to treatment. An advantage of model (2) is its generality in that any test producing a valid P-value can use this approach. The usual t-test is common but the bootstrap and nonparametric methods have also been proposed (e.g., Tusher et al., 2001; Neuhauser and Lam, 2004). Gadbury et al. (2003), however, noted that the distribution of P-values from randomization tests can be too discrete to model using (2) when sample sizes (number of microarrays) are small.

This paper builds on prior work (Allison et al., 2002) by assessing the precision of maximum likelihood estimates (MLEs) and Bayesian estimates of the parameter $\theta = (\lambda, r, s)$ and functions of θ under varying values of θ and k, particularly when λ is near one and/or when r and s have values reflecting a "near uniform" beta distribution. The effect of the number of tests, k, on the precision of parameter estimates is an important consideration since microarray experiments may involve 1000 tests to over 60,000 tests. Moreover, the present genomic era (Wolfsberg et al., 2002) opens a new realm of high dimensional biology (HDB) where multiple testing situations become the norm rather than the exception. Investigations into genetic polymorphisms, gene expression levels, protein measurements, genetic sequences, or any combination of these and their interactions may lead to experiments with tests numbering in the hundreds or in the hundreds of thousands.

We evaluate the precision of estimates using 95% confidence intervals and a 3^4 complete factorial simulation design. The function of θ considered is a "true positive probability" (TP) evaluated at a threshold τ,

$$TP = \frac{(1-\lambda)B(\tau; r, s)}{\lambda\tau + (1-\lambda)B(\tau; r, s)}, \tag{3}$$

where $B(\tau; r, s)$ is the cumulative distribution function (CDF) of a beta distribution with parameters r and s, evaluated at τ, and τ is a user-selected threshold at which a P-value is declared "significant." TP can be interpreted as the proportion of genes differentially expressed among those so declared at a threshold τ. In microarray experiments, values of TP provide guidance for follow-on research by quantifying how likely certain genes are indeed affected by a treatment. The quantity, $1 - TP$ is similar in concept to the false discovery rate (FDR) reported by others (e.g., Benjamini and Hochberg, 1995, 2000; Storey, 2002). Previous work has produced point estimates for θ (e.g., Parker and Rothenberg, 1988; Allison et al., 2002; Pounds and Morris, 2003; Gadbury et al., 2004) and hence, for TP, but little was done to evaluate the precision of these estimates.

We also evaluate the precision of estimates for a quantity called dTP which is like the expression in (3) except using density functions rather than CDF's. This quantity can be interpreted as a Bayesian posterior probability of expression. Subtracting dTP from one is analogous to the local false discovery rate discussed in other work (Efron, 2004). Since TP can be thought of as an "average" of posterior probabilities over all genes declared differentially expressed, dTP values will be lower than TP when evaluated at the same τ. Hereafter we focus on estimation results for TP, but results for dTP are included in supplementary material (the internet link is given in Section 3.1).

We compare three different methods for interval estimation: one using asymptotic normality of MLE's, another using potential symmetry of sampling distributions but using the bootstrap for standard errors, and the third using a Bayesian model with Markov–Chain Monte–Carlo (MCMC) to estimate a posterior probability distribution. All three methods had their advantages and limitations. Chung et al. (2004) considered a mixture of two exponential distributions and found that MCMC, when appropriately implemented, can offer advantages over the bootstrap when multimodal likelihood surfaces arise due to the label switching problem (Celeux et al., 2000). Another recent study by Dias and Wedel (2004) compared computational methods for estimation of parameters in a Gaussian mixture. Since our initial mixture component was completely specified, we did not need to address the label switching problem, and the implementation of MCMC was more straightforward as a result. Issues did arise, however, when the beta component was near uniform.

In summary, this paper provides insight into answers to two questions: 1. For what values of parameters and number of tests can one obtain precise parameter estimates? and 2. What are the relative advantages and limitations of three computational strategies for computing interval estimates. It is important to note that we only consider the ability to