Journal of Multivariate Analysis 160 (2017) 169–184

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Optimal detection of weak positive latent dependence between two sequences of multiple tests

Sihai Dave Zhao [a,*], T. Tony Cai [b], Hongzhe Li [c]

[a] Department of Statistics, University of Illinois at Urbana–Champaign, IL, United States
[b] Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA, United States
[c] Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

ARTICLE INFO

Article history:
Received 8 September 2016
Available online 14 July 2017

Keywords:
Detection boundary
Higher criticism
Independence testing
Optimal adaptivity
Sparsity

ABSTRACT

It is frequently of interest to jointly analyze two paired sequences of multiple tests. This paper studies the problem of detecting whether there are more pairs of tests that are significant in both sequences than would be expected by chance. The asymptotic detection boundary is derived in terms of parameters such as the sparsity of non-null cases in each sequence, the effect sizes of the signals, and the magnitude of the dependence between the two sequences. A new test for detecting weak dependence is also proposed, shown to be asymptotically adaptively optimal, studied in simulations, and applied to study genetic pleiotropy in 10 pediatric autoimmune diseases.

1. Introduction

1.1. Overview

Joint analysis of two paired sequences of multiple tests, each arising from a separate independent study, arises in many applications. It has been particularly motivated by genomics research, where it is natural to investigate similarities in how genomic features, such as genes or genetic markers, behave across studies. For example, recent interest has focused on features that may be significant in both of two sequences of multiple tests. In differential gene expression experiments, enrichment analysis [32] is often used to test whether two experiments share more significantly differentially expressed genes than would be expected by chance. In the integration of an expression quantitative trait loci study and a genome-wide association study, the goal is frequently to detect and identify genetic variants that are associated with both gene expression and disease [26,44]. Replicability analysis [27,29,30] aims to discover significant findings that have been replicated across genomic studies. Finally, studies of genetic pleiotropy investigate whether the same genetic variants may be simultaneously associated with different traits [8,11–13,22,38].

These examples broadly fall into two categories of questions: the detection of whether there exist features that are significant in both of two studies, and the identification of those simultaneously significant features. This paper focuses on the detection problem; the identification problem is studied elsewhere [11,30,45,55]. Specifically, let I_{kj} be unobserved latent indicators of whether the jth test, $j \in \{1, \ldots, p\}$, is truly non-null in the kth study, $k \in \{1, 2\}$. Let T_{kj} be the corresponding test statistic such that, for $k \in \{1, 2\}$,

$$T_{kj} \mid I_{kj} = 0 \sim F_k^0, \quad T_{kj} \mid I_{kj} = 1 \sim F_k^1, \quad I_{kj} \sim Ber(\pi_k), \tag{1}$$

* Corresponding author.
E-mail addresses: sdzhao@illinois.edu (S.D. Zhao), tcai@wharton.upenn.edu (T.T. Cai), hongzhe@upenn.edu (H. Li).

http://dx.doi.org/10.1016/j.jmva.2017.06.009
0047-259X/

where the π_k quantify the proportion of non-null tests in each study. The F_k^0 and F_k^1 can be viewed as mixtures of possibly different null and non-null distributions for different j. For each k, model (1) corresponds to a two-group mixture model for T_{kj}, which is common in the literature [15,16,21,49,50]. It will be assumed that the T_{kj} are two-tailed test statistics and are thus stochastically larger when $I_{kj} = 1$. Because the two sequences of tests arise from different studies, which typically are conducted on independent samples, it is assumed that T_{1j} and T_{2j} are independent conditional on the latent indicators I_{1j} and I_{2j}.

The goal of this paper is to test whether there are more features j that are significant in both studies than would be expected by chance. Formally, if $\Pr(I_{1j} = 1, I_{2j} = 1) = \epsilon$, the goal is to test

$$\mathcal{H}_0 : \epsilon = \pi_1\pi_2 \quad vs. \quad \mathcal{H}_A : \epsilon > \pi_1\pi_2. \tag{2}$$

This is motivated by a study of genetic pleiotropy in 10 pediatric autoimmune diseases conducted by Hakonarson and colleagues at the Children's Hospital of Pennsylvania [41,42]. More details about the data can be found in Section 4.6. Testing (2) using genome-wide association study summary statistics from a pair of diseases can assess whether the two conditions have some degree of shared genetic architecture, which can lead to a better understanding of their etiologies.

Several features make testing (2) difficult for existing methods. First, the I_{kj} are not directly observed. Second, in genomics applications, non-null features are typically rare and have weak effect sizes. For example, only a relatively small proportion of the human genome is expected to be associated with a given phenotype, and then only weakly so. Finally, positive dependence between I_{1j} and I_{2j} can be very weak when it exists, because cross-study heterogeneity makes it unlikely that more than a handful of features will be simultaneously non-null in both of two independently conducted genomics studies, even if the studies are closely related.

This paper proposes a new test for (2) under these challenging conditions. The proposed test statistic is shown to be asymptotically adaptively optimal, so that it performs as well as the optimal likelihood ratio test statistic but without needing to specify parameter values under $\mathcal{H}_0$ and $\mathcal{H}_A$. In fact the proposed test is entirely nonparametric, so neither F_k^0 nor F_k^1 needs to be known. It is also computationally efficient to implement and can be computed for 10 million pairs of tests in under one minute. It is available in the `R` package `ssa`.

1.2. Related work

Because model (1) assumes that T_{1j} and T_{2j} are independent conditional on I_{1j} and I_{2j}, testing (2) is equivalent to testing for independence between T_{1j} and T_{2j}. Classical methods are based on goodness-of-fit tests comparing the empirical bivariate distribution of (T_{1j}, T_{2j}) to the product of the marginal empirical distributions. Variations include Cramér–von Mises, Anderson–Darling, and Kolmogorov–Smirnov type tests [20,31,48,52]. A number of methods for detecting positive quadrant dependence have also been studied in the actuarial sciences [37]. Independence testing has seen renewed interest in the statistical literature, where the focus is on detecting arbitrary types of dependence [19,46,51]; see in particular Heller et al. [28]. In contrast, this paper is concerned with detecting a particular form of dependence between T_{1j} and T_{2j}, induced by the weak positive latent dependence between I_{1j} and I_{2j}. It appears that this type of dependence has not yet been specifically considered, and existing methods may be suboptimal. Furthermore, the fundamental limits of detection have not been studied.

Testing (2) can also be seen as an extension of the single-sequence signal detection problem. There, given test statistics T_{kj} from a single study k, the goal is to determine whether there are any non-null signals: $\mathcal{H}_0 : \Pr(I_{kj} = 1) = 0$ vs. $\mathcal{H}_A : \Pr(I_{kj} = 1) > 0$. The fundamental limits of detection for this problem have been derived, and asymptotically adaptively optimal tests have also been developed [2,9,10,14,15,33–36]. Special attention has been paid to the setting where π_k is very close to zero and F_k^1 is not too different from F_k^0. As previously noted, this rare and weak signal setting is also the focus of this paper. However, results for the single sequence problem do not apply to testing (2).

Several additional methods for testing (2) have been developed in the genomics literature. A popular approach is to estimate the I_{kj}, by thresholding the T_{kj}, and then to test for dependence using the estimated I_{kj} [32,47]. However, it is unclear how the thresholds on T_{kj} should be chosen. Alternatively, the GPA method [11] fits the (T_{1j}, T_{2j}) to a four-group mixture model, each group corresponding to one of the four possible values of the tuple (I_{1j}, I_{2j}), and uses a generalized likelihood ratio test for (2). However, GPA imposes parametric assumptions on F_k^0 and F_k^1. In addition, theoretical results from the single-sequence detection problem suggests that generalized likelihood ratio tests will have poor asymptotic properties when non-null T_{kj} are rare and weak [6,25]. Recently, Zhao et al. [56] proposed a simple test for (2) and studied its asymptotic properties. However, their theoretical results require distributional assumptions on the T_{kj}, and their test is only asymptotically optimal under specialized conditions.

The rest of the paper is organized as follows. Section 2 introduces the proposed test statistic and Section 3 studies its asymptotic adaptive optimality. Section 4 presents the results of simulation studies and the pediatric autoimmune disease analysis. The paper concludes with a discussion in Section 5. Additional simulation and data analysis results, and all proofs, can be found in the Supplementary Material; see Appendix A.