Chemometrics and Intelligent Laboratory Systems 182 (2018) 149–157

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Controlling two-dimensional false discovery rates by combining two univariate multiple testing results with an application to mass spectral data

Youngrae Kim [a], Johan Lim [a], Jong Soo Lee [b], Jaesik Jeong [c,*]

[a] Department of Statistics, Seoul National University, Seoul, South Korea
[b] Department of Mathematical Sciences, University of Massachusetts Lowell, MA, USA
[c] Department of Statistics, Chonnam National University, Gwangju, South Korea

ARTICLE INFO

Keywords:
Bonferroni rule
Composite hypothesis
Intersection/union null
Mass spectral data
Two-dimensional false discovery rate
Two-stage procedure

ABSTRACT

Mass spectral data exhibit a small number of signals (true peaks) among many noisy observations (signals or true peaks) in a high-dimensional space. This unique aspect of mass spectral data necessitates solving the problem of testing for many composite null hypotheses simultaneously. In this study, we develop a new procedure to control the false discovery rate of simultaneous multiple hypothesis tests, consisting of many "bivariate" composite null hypotheses. Two types of composite null hypothesis, the intersection-type and the union-type null, are considered separately. The proposed procedure comprises two stages. In the first stage, we simultaneously test each "univariate" simple hypothesis of "bivariate" composite hypotheses at the pre-decided false discovery rate. In the second stage, we combine the marginal univariate test results so that the two-dimensional false discovery rate for the "bivariate" composite null hypotheses is less than the desired significance level α. The new procedure provides a closed-form decision rule on the bivariate test statistics, unlike existing methods for controlling the two-dimensional local false discovery rate (2d-fdr). We numerically compare the performance of our procedure to existing 2d-fdr control methods in different settings. We then apply the procedure to the problem of differentiating the origins of herbal medicine using gas chromatography-mass spectrometry.

1. Introduction

High-dimensional spectral data are widely used in various biological and medical disciplines. Examples include near infrared spectral (NIR) data, nuclear magnetic resonance (NMR) data, liquid chromatography mass spectral (LC/MS) data, and gas chromatography mass spectral (GC/MS) data. NMR is used to observe the magnetic properties of the energy absorbed and re-emitted from an atomic nucleus, which is used to identify compounds in a given sample mixture [2–4]. Mass spectrometry (MS) ionizes chemical compounds and measures the mass-to-charge ratios of charged particles (ion fragments), and is popular in many bio-analytical sectors [5–7].

A mass spectrometry raw data set consists of true meaningful peaks (of interest) and noisy peaks. Clearly, the number of true and noisy peaks depend on the study sample. To find significant true peaks, various methodologies have been developed to pre-process the raw data (e.g., peak detection, peak normalization, and peak merging) [8,9]. Furthermore, the raw data usually contain unwanted local or global peak shifts due to instrumental instability or small differences in experimental conditions. Because a mis-alignment over samples weakens the strength of the signals, the spectrum must be aligned prior to an analysis [10–20].

Following the advances in pre-processing technologies, the focus in the MS field has shifted to statistical issues such as biomarker metabolite discovery and metabolite-metabolite network construction. Even though multiple testing has attracted much attention in terms of identifying significant metabolites or biomarkers, statistical analyses of MS data are relatively rare compared with those of other high-throughput data. One reason for this is that MS data require many pre-processing steps. A second, more important reason is that the conventional multiple testing approach is not able to find significant metabolites with a small variance, producing too many false positives. To overcome these disadvantages of the conventional *t*-test, many variations have been developed. The variations typically address the small variance in two different ways, namely, from a frequentist and an empirical Bayes perspective [21–25]. For example, one approach adds a constant to the small standard error of the sample mean difference [21,25]. In other instances, the posterior mean of the variance obtained using a χ^2 prior is considered as the standard error [22,24].

* Corresponding author.
E-mail address: jjs3098@gmail.com (J. Jeong).

https://doi.org/10.1016/j.chemolab.2018.09.006
Received 28 December 2017; Received in revised form 10 September 2018; Accepted 19 September 2018
Available online 26 September 2018
0169-7439/

Because there is a limitation on possible improvements to the single test statistic (modified *t*-statistic), a different solution to the small variance problem was suggested by Ploner et al. [1]. They suggested using a two-dimensional statistic to improve the performance of multiple testing, by applying a different cutoff rule to the modified *t*-statistic depending on the size of the variance. In other words, a different amount of fudge factor is added to the observed standard error, reducing the number of false positives. They also provide a corresponding 2d-fdr approach, which can be easily extended to a multi-dimensional case. Estimated fdr isolines (log se vs. *t*-statistic) are smoothed and cropped to the convex hull of the observed statistics. Tornado and volcano plots, which use a different *x*-axis (mean difference instead of the *t*-statistic), are employed to graphically show that significance depends on the magnitude of both the mean difference and its standard error. However, both plots have singularity problems that make it difficult to smooth the lines. Furthermore, the estimated fdr isolines are not explicitly represented in mathematical form.

In this paper, we propose new procedures to control the two-dimensional false discovery rate in the simultaneous testing of many bivariate composite hypotheses. We consider two types of (bivariate) composite null hypothesis, intersection-type null hypotheses and union-type null hypotheses, and develop procedures appropriate for each type of composite hypothesis. The proposed procedures (for both types of composite null hypotheses) comprise two stages. In the first stage, we simultaneously test each "univariate" simple hypothesis of bivariate composite hypotheses at the pre-decided FDR level. Of the many procedures available for multiple testing with univariate test statistics, we adopt the local false discovery rate procedure of Efron et al. [21] in the first stage of our procedure. In the second stage, we combine the marginal univariate test results so that the two-dimensional false discovery rate for the bivariate composite null hypotheses is less than the desired significance level, α. The combining rule for the intersection-type composite hypothesis is the same as the Bonferroni correction. Specifically, we apply a univariate FDR procedure to test each simple hypothesis at the levels α_1 and α_2, with $\alpha_1 + \alpha_2 = \alpha$. We then reject an individual hypothesis if it is rejected by either of the two univariate FDR procedures. The combining procedure for the union-type composite hypothesis is rather complex, and will be explained later. Compared with the existing two-dimensional FDR procedures [1,26], our proposed procedure differentiates between the types of composite null hypotheses and provides a closed-form decision rule on the bivariate test statistics. Accordingly, the decision on significance for a new object is easily made using the explicit rejection region.

In the next section, we introduce our two-stage procedures for the two types of composite null hypotheses. In Section 3, we numerically compare the performance of our procedure to that of two other 2d-fdr procedures [1,26]. In the numerical comparison, we first consider the union-type composite null hypotheses, which is more appropriate for the comparison with the method of Ploner et al. [1]. We then consider the intersection composite null for the comparison with the approach of Alishahi et al. [26]. In Section 4, we apply our procedure to the problem of identifying the origins of a herbal medicine using gas chromatography-mass spectral data. In Section 5, we conclude with a brief summary of the paper.

2. Two-stage procedure to control the FDR in two dimensions

2.1. Procedure for the intersection composite null

We first introduce a procedure for testing intersection-type composite null hypotheses simultaneously. Suppose, for $j = 1, 2, \ldots, J$, the j-th individual hypothesis is of the form

$$\mathscr{H}_{j.0} = \mathscr{H}_{j.01} \cap \mathscr{H}_{j.02},$$

where hypotheses $\mathscr{H}_{j.01}$ and $\mathscr{H}_{j.02}$ are tested using statistics T_1 and T_2, respectively. The FDR control procedure we propose is similar to the Bonferroni correction for the multiple testing error:

1. Set positive levels α_1 and α_2 as $\alpha_1 + \alpha_2 = \alpha$.
2. For T_1 and T_2, marginally test the hypotheses at the FDR levels α_1 and α_2, respectively, and find the index where the tests reject either hypothesis. Let the set of indices of such hypotheses be $\mathscr{K}$.
3. Reject the hypothesis $\mathscr{H}_{j.0}, j \in \mathscr{K}$.

Table 1
All possible combinations of truths (TR) and actions (A).

TR_1	TR_2	A_1	A_2	count	TR_1	TR_2	A_1	A_2	count
0	0	0	0	n_{0000}	1	0	0	0	n_{1000}
0	0	0	1	n_{0001}	1	0	0	1	n_{1001}
0	0	1	0	n_{0010}	1	0	1	0	n_{1010}
0	0	1	1	n_{0011}	1	0	1	1	n_{1011}
0	1	0	0	n_{0100}	1	1	0	0	n_{1100}
0	1	0	1	n_{0101}	1	1	0	1	n_{1101}
0	1	1	0	n_{0110}	1	1	1	0	n_{1110}
0	1	1	1	n_{0111}	1	1	1	1	n_{1111}

Below, we show how the above procedure controls the FDR, the expected proportion of false rejections among total rejections, such that it is less than α. Table 1 shows the number of hypotheses for all four combinations of the testing results of the $J = n_{++++} (= \sum_{i=0}^{1} \sum_{j=0}^{1} \sum_{k=0}^{1} \sum_{l=0}^{1} n_{ijkl})$ hypotheses. We make an independent Poisson assumption for each number (our count) n_{ijkl}s, with rates $J\lambda_{ijkl}$. Then, given J, the conditional distribution of the vector of n_{ijkl}s is the multinomial distribution with success probabilities $\pi_{ijkl} = \lambda_{ijkl} / \sum_{ijkl} \lambda_{ijkl}$s. The four combinations consist of the truth (TR) and action (A) corresponding to the first and second hypotheses $\mathscr{H}_{j.01}$ and $\mathscr{H}_{j.02}$; "0" or "1" indicate the null or alternative hypothesis on TR and A, respectively.

In the procedure above, we separately control the FDR for one-dimensional multiple testing for $\mathscr{H}_{j.01}$ and $\mathscr{H}_{j.02}$. The control of the FDR for testing the hypotheses $\{\mathscr{H}_{j.01}, j = 1, 2, \ldots, J\}$ implies

$$\mathrm{E}\left(\frac{n_{0+1+}}{n_{++1+}}\right) \le \alpha_1. \tag{1}$$

Similarly, the FDR control for the hypotheses $\{\mathscr{H}_{j.02}, j = 1, 2, \ldots, J\}$ implies

$$\mathrm{E}\left(\frac{n_{+0+1}}{n_{+++1}}\right) \le \alpha_2. \tag{2}$$

If the above two are satisfied, then we have

$$\mathrm{E}\left(\frac{\#\text{ of false rejections}}{\#\text{ of rejections}}\right) = \mathrm{E}\left(\frac{n_{0011} + n_{0010} + n_{0001}}{n_{++11} + n_{++10} + n_{++01}}\right) \le \alpha_1 + \alpha_2. \tag{3}$$

We show that Equation (3) holds; that is, combining (1) and (2) implies (3):

$$\frac{n_{0011} + n_{0010} + n_{0001}}{n_{++11} + n_{++10} + n_{++01}} \le \frac{2n_{0011} + n_{0010} + n_{0001}}{n_{++11} + n_{++10} + n_{++01}}$$

$$= \frac{n_{0011} + n_{0010}}{n_{++11} + n_{++10} + n_{++01}} + \frac{n_{0011} + n_{0001}}{n_{++11} + n_{++10} + n_{++01}}$$

$$\le \frac{n_{0+1+}}{n_{++1+} + n_{++01}} + \frac{n_{+0+1}}{n_{+++1} + n_{++10}}$$

$$= \frac{n_{++1+}}{n_{++1+} + n_{++01}} \cdot \frac{n_{0+1+}}{n_{++1+}} + \frac{n_{+++1}}{n_{+++1} + n_{++10}} \cdot \frac{n_{+0+1}}{n_{+++1}}$$

$$\le \frac{n_{0+1+}}{n_{++1+}} + \frac{n_{+0+1}}{n_{+++1}}.$$