adaptest: Data-Adaptive Statistics for High-Dimensional Testing in R

15 October 2018

Summary

The `adaptest` R package contains an implementation of a methodology based on using *data-adaptive statistics* for estimating effect sizes, complete with appropriate inference, in high-dimensional settings while avoiding the inferential burdens of multiple testing corrections. To address the issue of multiple testing in situations where the dimensionality is high but sample size comparatively small (*e.g.*, analysis of RNA-seq data), we expose an implementation of a method for statistical inference on data-adaptive target parameters (Hubbard, Kherad-Pajouh, and van der Laan 2016) in the form of a software package for the R language and environment for statistical computing (R Core Team 2018).

Data-adaptive test statistics for multiple testing are motivated by efforts to address the limitations of existing multiple testing methods such as the popular Benjamini-Hochberg procedure to control the False Discovery Rate (FDR) (Benjamini and Hochberg 1995) or the Bonferroni method to control the Family-Wise Error Rate (FWER) (Dunn 1961). Such methods are well studied in the literature on multiple testing, and it is well established that, for a fixed targeted effect size and fixed sample size, power decreases as the number of tests and corresponding critical values increase (Lazzeroni and Ray 2010). Further, Lazzeroni and Ray (2010) show that if the power for a single test is 80%, the power is approximately 50% for 10; 10% for 1000; and 1% for 100,000 Bonferroni-adjusted tests, a classic method to correct for Type-I error when facing multiple testing issues. This simple example demonstrates that data analysts and other practitioners must invest, at a prohibitively high rate, additional resources to collect samples in order to obtain meaningful results under high-dimensional multiple testing constraints.

Utilizing this recently developed data-adaptive statistical framework, our method reduces information loss induced by standard multiple testing procedures through data-adaptive dimensionality reduction. This recent methodological advance, a data-adaptive multiple testing technique (Cai, Hejazi, and Hubbard), is a

natural extension of the data-adaptive target parameter framework introduced in Hubbard, Kherad-Pajouh, and van der Laan (2016) and Hubbard and van der Laan (2016), which present a new class of inference procedures that introduce more rigorous statistical inference into problems being increasingly addressed by smart yet *ad hoc* algorithms for data mining.

The approach of data-adaptive test statistics improves on current approaches to multiple testing by applying a set of data-mining algorithms (specified by the user) across splits of a particular sample of data, allowing for parameters of interest to be discovered from the data. Such methods uncover associations that are stable across the full sample and restrict multiple testing to a smaller subset of covariates by allowing for variable importance to be measured via the data-adaptive procedure. Test statistics are formulated on a separately held-out subset of data and are expected to both outperform pre-specified test statistics and provide improved power, all while simultaneously allowing for appropriate statistical inference to be performed.

We illustrate how to apply the `data-adaptive test statistics` for multiple testing by considering a simulated randomized trial with binary treatment and 1000 outcomes (e.g., biomarkers in the microarray analysis). The dataset size is 100 observations. Of the 1000 outcomes (biomarkers), outcome 1 - 10 have effect sizes equal to 0.6, while the treatment has no effect on outcomes 11 - 1000. After applying our `data-adaptive test statistics` method (using the `adaptest` function in the R package), we obtain a rank order (regarding effect size) for all outcomes across multiple cross-validation folds. We then average the rank order across folds, sort in ascending order, which gives us Figure 1. By looking at the top 15 outcomes in Figure 1, we observe that there are two large jumps in average rank order of the top 15 outcomes: between outcome 9 and 4, and between outcome 3 and 2. These jumps naturally divide the outcomes into tiers regarding importance. Outcome 9 consistently ranks highly in the importance measure employed across the many rounds of cross-validation performed. In this example, we recommend practitioner first to analyze outcome 9, and if data size allows, extend the analysis to the group of outcome from 4 to 3, and so on. Figure 2 displays adjusted p-values of the same set of outcomes as in Figure 1, with a group of outcomes (outcome 9 to outcome 3) with very significant effect.

The `adaptest` R package provides utilities for performing the estimation and hypothesis testing procedures discussed above, and detailed in Cai, Hejazi, and Hubbard, alongside utilities for easily producing data visualizations based on the results. The software introduces new classes and methods, based on R's S4 class system, to facilitate its integration into the Bioconductor ecosystem (Huber et al. 2015), making it well-suited for applications in computational biology, where high-dimensional data structures very often arise. The R package includes documentation and detailed vignettes that will allow for both (bio)statisticians and computational biologists to efficiently make use of this new tool in such data analytic problem settings.

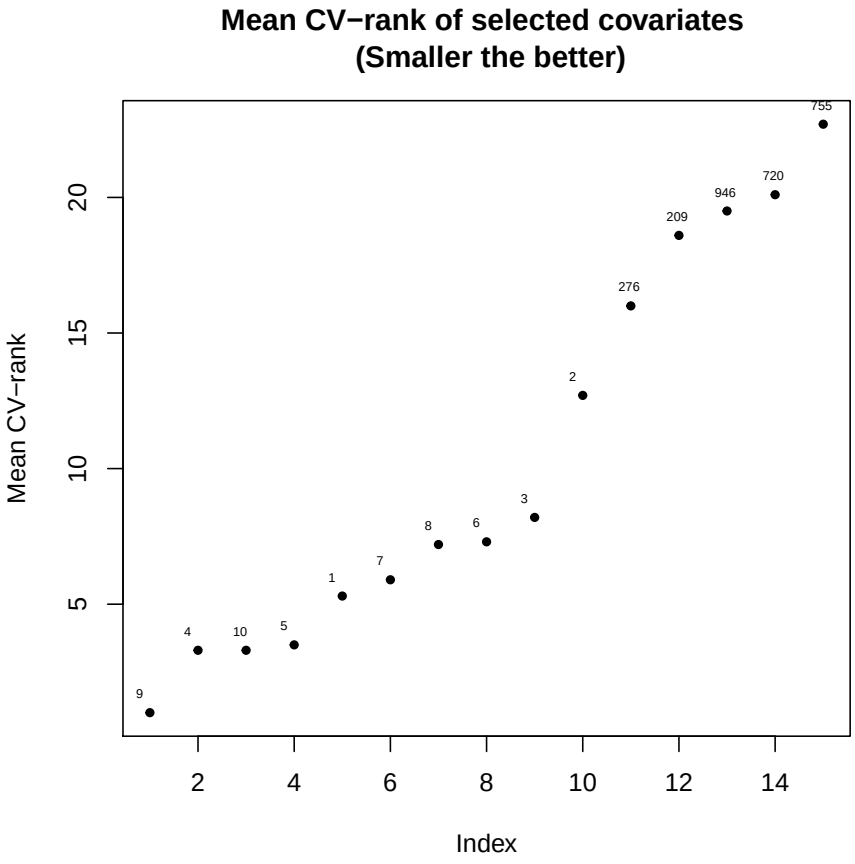

Figure 1: Average rank order of outcomes regarding absolute estimated effect size across cross-validation folds (simulated data). The top outcomes are displayed after being sorted in ascending order.

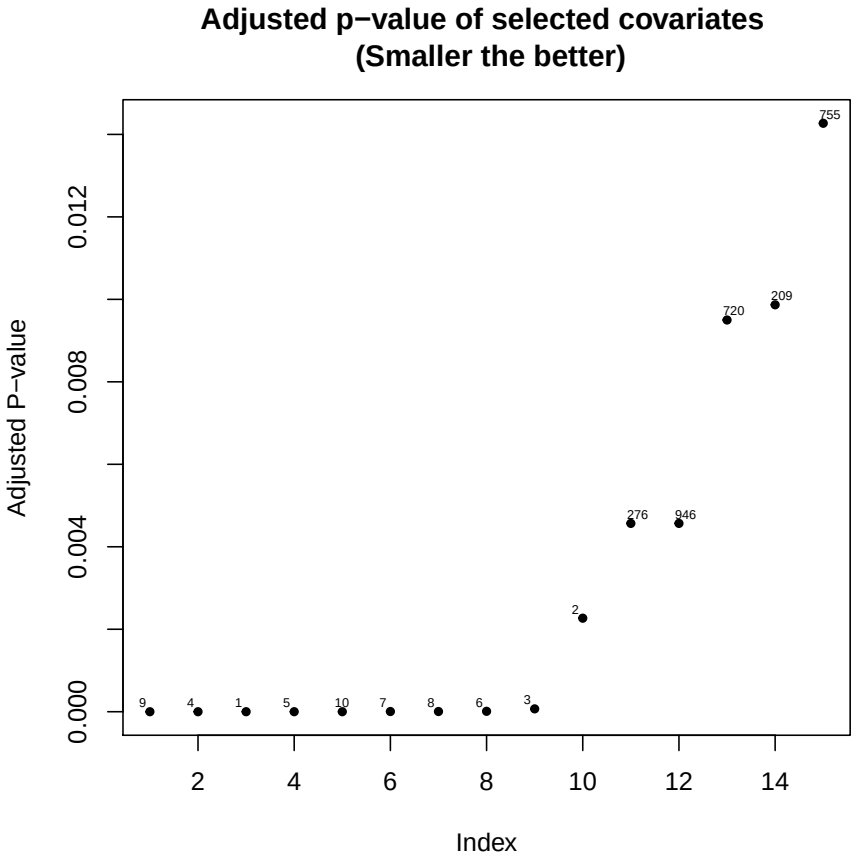

Figure 2: Adjusted p-values (using the Benjamini-Hochberg procedure) of the same set of candidate outcomes, computed on a validation set that is mutually exclusive from the data used to compute the rank order in Figure 1. The top outcomes are displayed after being sorted in ascending order.

References

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)*. JSTOR, 289–300. doi:10.2307/2346101.

Cai, Weixin, Nima S Hejazi, and Alan E Hubbard. "Data-Adaptive Statistics for Multiple Hypothesis Testing in High-Dimensional Settings." https://arxiv.org/abs/1704.07008.

Dunn, Olive Jean. 1961. "Multiple Comparisons Among Means." *Journal of the American Statistical Association* 56 (293). Taylor & Francis Group: 52–64. doi:10.2307/2282330.

Hubbard, Alan E, and Mark J van der Laan. 2016. "Mining with Inference: Data-Adaptive Target Parameters." In *Handbook of Big Data*, edited by Peter Buhlmann, Petros Drineas, Michael Kane, and Mark J van der Laan. CRC Press, Taylor & Francis Group, LLC: Boca Raton, FL.

Hubbard, Alan E, Sara Kherad-Pajouh, and Mark J van der Laan. 2016. "Statistical Inference for Data Adaptive Target Parameters." *The International Journal of Biostatistics* 12 (1): 3–19. doi:10.1515/ijb-2015-0013.

Huber, Wolfgang, Vincent J Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S Carvalho, Hector Corrada Bravo, et al. 2015. "Orchestrating High-Throughput Genomic Analysis with Bioconductor." *Nature Methods* 12 (2). Nature Research: 115–21. doi:10.1038/nmeth.3252.

Lazzeroni, LC, and A Ray. 2010. "The Cost of Large Numbers of Hypothesis Tests on Power, Effect Size and Sample Size." *Molecular Psychiatry*. Nature Publishing Group. doi:10.1038/mp.2010.117.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.