data.adapt.multi.test: Data-Adaptive Statistics for High-Dimensional Testing

16 November 2016

Summary

This package contains an implementation of the data-adaptive statistical approach to estimating effect sizes in high-dimensional settings. To address the issue of multiple testing in situations where the dimensionality is high but sample size is comparatively small (*e.g.*, genomics problems), we provide here an implementation of data-adaptive multiple-testing procedures in the form of a package for the R language for statistical computing (R Core Team 2016).

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 Bonferroni method to control the Family-Wise Error Rate(FWER) (Dunn 1961). It has been well studied in literature that for a fixed targeted effect size and fixed sample size, power decreases as the number of tests and corresponding critical value increase (Lazzeroni and Ray 2010). Lazzeroni and Ray (2010) shows 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 doing multiple testing. This means that practitioners need to invest in prohibitively more resources to collect samples in order to get meaningful results under high-dimensional multiple testing.

Utilizing this recently developed data-adaptive statistical framework, information loss induced by standard multiple testing procedures can be avoided by reducing the dimensionality of problems via variable reduction. This newly developed methodology 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 provide a way to introduce more rigorous statistical inference into problems being increasingly addressed by clever 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 estimation 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 formulated in this framework are expected to both outperform pre-specified test statistics and provide improved power as well as Type I error control, all while simultaneously allowing for appropriate statistical inference to be performed.

We illustrate the use of data-adaptive test statistics for parameter discovery by considering a simulated data set with 100 observations in 1000 dimensions, with a "true" signal constrained to just 10 covariates/dimensions. By applying the approach discussed above, using cross-validation to rank features, we obtain a ranking of the most important covariates – that is, those dimensions most closely associated with the "true" signal. A ranking of features across folds of cross-validation is displayed below:

From the plot displayed above, it is clear to see that there is a rather sharp divide in the ranking of covariates associated with the "true" signal – that is, these are those covariates that consistently rank highly in the importance measure employed across the many rounds of cross-validation performed. The plot of p-values displayed below shows these same features with low p-values, with a clearly strong divide consistent with that displayed in the previous plot:

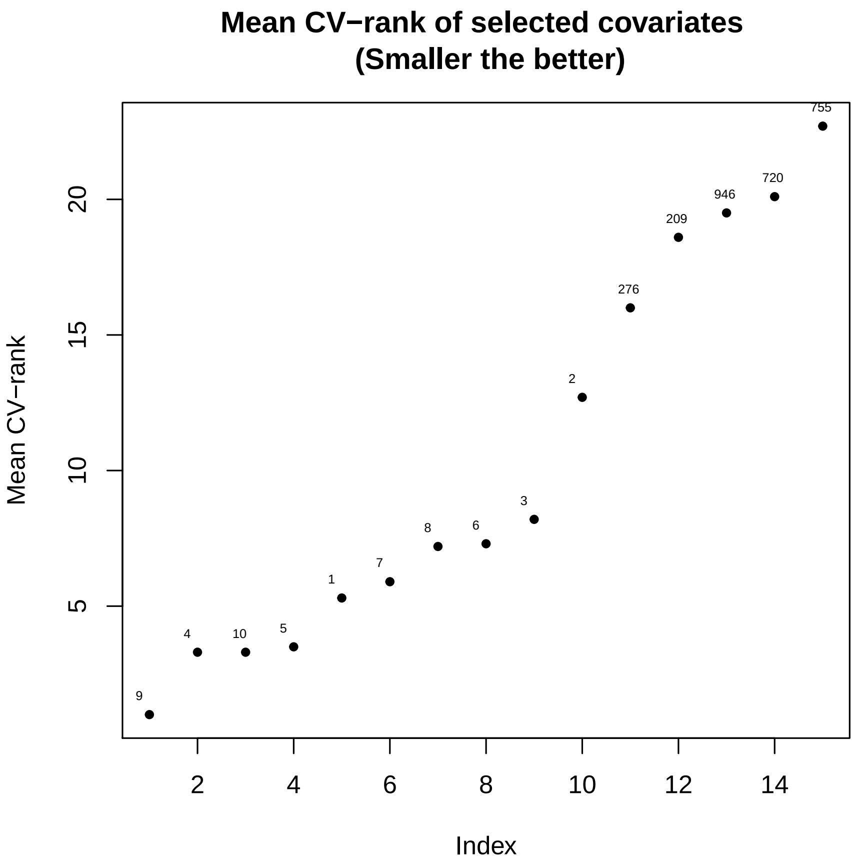

Figure 1: Average Rank of Top Covariates: here, the top ten covariates have CV-rank aligning linearly, indicating a stable ranking pattern.

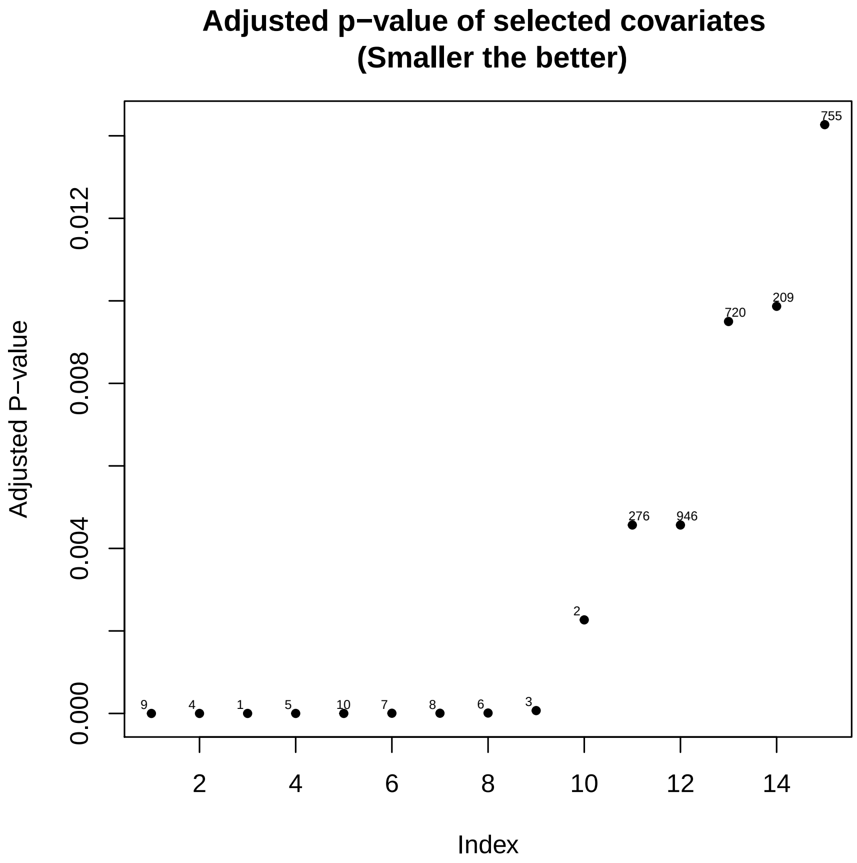

Figure 2: Adjusted P-values for the Reduced Set of Hypotheses

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.

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.

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. 2016. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.