“Automated bias-variance trade-off: Intuitive inadmissibility or inadmissible intuition”

Nilanjan Chatterjee and Bhramar Mukherjee

We would like to thank Dr. Meng for attempting to provide more insight into the properties of a shrinkage estimator that we had earlier introduced for estimation of gene-environment interaction in case-control studies [M & C, 2008]. Unfortunately, some of Dr. Meng’s comments may generate misconception about gene-environment studies and the properties of our estimator in that particular context. To keep the discussion brief, we will only elaborate on two major issues.

First, it may appear from Dr. Meng’s guided insight (see Section 3.3.2.2) that the assumption of gene-environment independence is more natural for subjects with D=0 (i.e. the controls) and that we, as well as other authors, have made the assumption for the entire population just because it is more convenient to test the assumption in that way. Exact opposite is true. It is easy to imagine, for example, that inherited genetic susceptibility factors (G) and an external environmental exposure, like air pollution (E), are independently distributed in an underlying population. These two factors, however, can interact with each other to cause a future event like lung cancer. Here gene-environment interaction means effect modification, not to confuse with gene-environment independence. In general, there is no reason to believe that G and E will be independent conditional on disease status. If the disease can be assumed to be rare, then the natural assumption for the population should also hold approximately for the disease-free subjects since the diseased subjects form a rare fraction of the entire population. The assumption of rare disease is not very restrictive since case-control studies are only done for rare diseases. The assumption is also not necessary as we [Chatterjee and Carroll, 2005] and others [Lin and Zheng, 2006] have shown how to make inference in case-control studies under the gene-environment independence assumption for the population without making the assumption of rare disease. Along the same line, we would like to point out that Dr. Meng misrepresented the paper by Schmidt and Schaid by quoting that these authors used the rare disease assumption while they precisely showed that there is a valid interpretation of case-only odds-ratio of interaction without the rare disease assumption.

Second, Dr. Meng seems to have missed the reference by Chen, Chatterjee and Carroll (2009) which showed many of the same asymptotic properties of our proposed estimator that Dr. Meng outlined in his article. Based on this theory and extensive numerical studies, we already had noted that the proposed estimator can perform better than standard logistic regression estimator when the restricted model is close to being true, but otherwise they will perform similarly in large sample. It is, however, important to keep in mind that in our specific application the restricted model, i.e. gene-environment independence, often will be true in practice. In large-scale genome-wide association studies, for example, vast majority, if not all, of hundreds of thousands of genetic markers are expected to have no relationship with various environmental exposures in the underlying population. Thus exploiting the assumption of gene-environment independence to study how they may interact to cause a disease is reasonable. Using a data adaptive shrinkage estimator, as opposed to using a purely constrained estimator, can protect against bias in those few cases where gene-environment association may exist. In such high dimensional application what matters is how the different methods perform on average over the likely distribution of parameter
values. Our numerical investigations have shown that our proposed method can outperform both completely constrained and unconstrained estimators in terms integrated mean-square-error under likely distributions of gene-environment joint distribution. Further theoretical comparisons between the methods in terms of integrated loss functions may shed more insight into appropriate application of the method for different high-dimensional problems.

In conclusion, we would like to point out that theoretical results are always valuable, but their full implications cannot be appreciated without understanding the scientific context.

**Additional References beyond the ones provided by Dr. Meng:**

