Simultaneous Testing of Grouped Hypotheses: Finding Needles in Multiple Haystacks

by
Tony Cai
University of Pennsylvania
Department of Statistics
The Wharton School
University of Pennsylvania
Philadelphia, PA 19104 USA
tcai@wharton.upenn.edu

Abstract

In large-scale multiple testing problems, data are often collected from heterogeneous sources and hypotheses form into groups that exhibit different characteristics. Conventional approaches, including the pooled and separate analyses, fail to efficiently utilize the external grouping information. We develop a compound decision theoretic framework for testing grouped hypotheses and introduce an oracle procedure that minimizes the false non-discovery rate subject to a constraint on the false discovery rate. It is shown that both the pooled and separate analyses can be uniformly improved by the oracle procedure. We then introduce a data-driven procedure that is shown to be asymptotically optimal. Both theoretical and numerical results demonstrate that exploiting external information of the sample can greatly improve the efficiency of a multiple testing procedure. The results also provide insights on how the grouping information is incorporated for optimal simultaneous inference.