Peter MacDonald

April 16, 2017

Powerful modern computers have introduced large-scale data sets to diverse fields of research, and testing of hundreds or even thousands of hypotheses simultaneously has become commonplace in statistical applications such as genetics, neural science, and astronomy. In the classical formulation, m independent null hypotheses $H_1, H_2, ..., H_m$, of which m_0 are true and $m_1 = m - m_0$ are false, are tested simultaneously. Denote the associated p-values by $p_1, p_2, ..., p_m$ and $p_{(1)} \leq \cdots \leq p_{(m)}$ in ascending order. For $t \in [0, 1]$, define the following empirical processes (Storey et al., 2004):

$$\begin{aligned} V(t) &= \#\{\text{null } p_i : p_i \leq t\}, \\ S(t) &= \#\{\text{alternative } p_i : p_i \leq t\}, \\ R(t) &= V(t) + S(t). \end{aligned}$$

Then the false discovery rate (FDR) at a p-value cut-off $t \in (0, 1]$ is defined as

$$\text{FDR}(t) = E\left[\frac{V(t)}{R(t) \vee 1}\right].$$

Since its inception in Benjamini and Hochberg (1995), the FDR, the expected proportion of false positives, has been adopted as an error measure that strikes a pleasing middle ground between liberal single-inference procedures and conservative measures based on the traditional familywise error rate (FWER). Much research effort has been made to improve Benjamini and Hochberg's initial method for control of FDR, in particular developing efficient estimators of the FDR that lead to powerful procedures with proper FDR control.

My research has provided the first proof of finite sample FDR control for a large class of powerful data-adaptive methods which are utilized in applications, including the right-boundary procedure of Liang and Nettleton (2012). Through an intricate approximation argument, I was able to further extend this result to encompass the popular slope-based procedure of Benjamini and Hochberg (2000). Although asymptotic control has been proven for these procedures, finite sample control is the gold standard in the multiple testing literature and is particularly important to theoretically justify their continued use.

My research has also touched on the related field of grouped multiple testing. The grouped testing problem is highly similar to the usual multiple testing problem described above, but in addition hypotheses are organized into finitely many known groups. It is known that by leveraging this group label information in a pooled analysis, it is possible to control the overall FDR with greater power than can be achieved by ignoring group information, or by analyzing each group individually (Cai and Sun, 2009). Much of the literature on this particular topic only gives asymptotic assurances of FDR control, but through the application of martingale theory, I have been able to construct the first procedure that assures finite sample FDR control, with power comparable to the best procedures currently available in the literature.