False discovery rate control and the dependence structure of test statistics

Roy Cerqueti, Mauro Costantini, and Claudio Lupi

Abstract The False Discovery Rate (FDR) was proposed in Benjamini and Hochberg (1995) as a powerful approach to the multiplicity problem that does not require strong control of the familywise error rate (FWER). The original approach was developed for independent test statistics and was later extended to dependent statistics in Benjamini and Yekutieli (2001) and Yekutieli (2008). In this paper we extend the existing results by showing that the assumptions on the dependence structure among classes of univariate statistics, that lead to the FDR control, may be represented by specific copulas.

Key words: False discovery rate, copulas, dependence

1 Introduction

When many hypotheses are tested simultaneously, the risk of falsely rejecting truly null hypotheses increases dramatically. On the other hand, one would like to identify as many "discoveries" as possible, while incurring in a small proportion of false positives. This is the motivation of the concept of *False Discovery Rate* (FDR) introduced by Benjamini and Hochberg (1995) (BH) and further developed by Benjamini and Yekutieli (2001). An even more general scheme, the *separate subsets*

Roy Cerqueti
University of Macerata, Dept. of Economic and Financial Institutions, Via Crescimbeni, 20, I-62100 Macerata (Italy), e-mail: roy.cerqueti@unimc.it

Mauro Costantini
Brunel University, Dept. of Economics and Finance, Kingston Lane, Uxbridge. Middlesex UB8 3PH (UK), e-mail: Mauro.Costantini@Brunel.ac.uk

Claudio Lupi
University of Molise, Dept. of Economics, Management, and Social Sciences, Via F. De Sanctis, I-86100 Campobasso (Italy), e-mail: lupi@unimol.it

Benjamini-Hochberg (ssBH) procedure, has been introduced recently in Yekutieli (2008) in order to control the FDR in the presence of rather general forms of dependence among the tests statistics.

Denote by $\boldsymbol{p} = (p_1, \ldots, p_m)'$ the vector of the m p values associated to the components of the collection of m statistics $\boldsymbol{t} = (t_1, \ldots, t_m)'$. Consistently with Yekutieli (2008), we assume that the p values in $\boldsymbol{p}$ are co-monotone transformations of the corresponding test statistics in $\boldsymbol{t}$. Divide $\boldsymbol{p}$ in S sub-vectors $\boldsymbol{p}^s$, for $s = 1, \ldots, S$. With a very intuitive notation, the statistics corresponding to $\boldsymbol{p}^s$ constitute a vector, that will be indicated with $\boldsymbol{t}^s$. Assume that the cardinality of $\boldsymbol{p}^s$ is m^s and denote as $\boldsymbol{p}_0^s$ the p values in $\boldsymbol{p}^s$ corresponding to the true null hypotheses. The level q ssBH procedure runs into two steps as follows:

i. For $s = 1, \ldots, S$, apply the BH procedure at level qm^s/m to test $\boldsymbol{p}^s$, and denote as $\boldsymbol{r}_{BH}^s$ the p values corresponding to the rejected hypotheses.
ii. Reject the null hypothesis corresponding to $\boldsymbol{r}_{ssBH} = \bigcup_{s=1}^{S} \boldsymbol{r}_{BH}^s$.

In the present paper we show that the assumptions on the dependence structure among classes of univariate statistics, that lead to the FDR control, may be captured by specific copulas.

2 Main results

In order to proceed, we start from a particular case, and then we extend the analysis to more general frameworks.

A condition on the sets $\boldsymbol{p}^s$ is now needed.

Condition 1 *One of the following assumptions holds:*

(i) if $p_i \in \boldsymbol{p}_0$, then there exists a unique $s_i \in \{1, \ldots, S\}$ such that $p_i \in \boldsymbol{p}^{s_i}$. Moreover, for each $s = 1, \ldots, S$, it must be:

$$m^s = \begin{cases} 2, & \text{if } \boldsymbol{p}_0^s \neq \emptyset; \\ \text{arbitrary}, & \text{otherwise.} \end{cases}$$

(ii) $\boldsymbol{p}^{s_i} \cap \boldsymbol{p}^{s_j} = \emptyset$, for $s_i \neq s_j$, and $m^s = 2$, for each $s = 1, \ldots, S$.

Condition 1 means that the division of the set $\boldsymbol{p}$ in the subsets $\boldsymbol{p}^s$ is such that each p value of a true null hypothesis is contained in one $\boldsymbol{p}^s$, and each $\boldsymbol{p}^s$ containing a p value of a true null hypothesis has cardinality equals to 2. This is not a restrictive hypothesis, since the decomposition of $\{\boldsymbol{p}^s\}_{s=1,\ldots,S}$ to be used for the ssBH procedure can be arbitrarily chosen. It is worth noting that when (ii) of Condition 1 holds, then $m^s = 2$, for each $s = 1, \ldots, S$; if (i) is true, then $\exists\ \tilde{S} \leq S$ such that $m^s = 2$, for each $s = 1, \ldots, \tilde{S}$.

We are now able to state our first main result:

Proposition 1. *Assume that Condition 1 holds and that the dependence between the statistics in $\boldsymbol{t}^s$ is described by a copula C_s such that:*

$$C_s(u,v) = uv + \theta\phi(u)\phi(v), \tag{1}$$

for each $s = 1,\dots,S$, with $\theta \in [-1,1]$ and $\phi : [0,1] \to [0,1]$ satisfying the following conditions:

(i) $\phi(0) = \phi(1) = 0$;
(ii)ϕ is Lipschitzian in $[0,1]$, i.e.: $|\phi(x) - \phi(y)| \le |x-y|$, for each $x,y \in [0,1]$;
(iii)ϕ is convex or concave in $[0,1]$.

Then the level q ssBH procedure controls the FDR at level qm_0/m.

The main limitation of this approach is that it refers to *couples*, i.e. subsets of cardinality 2. However, it is possible to generalize the result by using a s-variate approach, with $s > 2$. A copula will be introduced also in this case to ensure the FDR control, with the crucial difference that now we treat a s-variate framework, with $s > 2$.

We first recall a generalization of the monotonic property for functions, that will turn out to be useful below:

Definition 1. A function

$$\psi : [0,1] \to [0,+\infty) \tag{2}$$

is completely monotone in $[0,1]$ if and only if $\psi \in C^\infty(0,1) \cap C^0[0,1]$, and $(-1)^n \psi^{(n)}(x) \ge 0, \forall n = 0,1,2,\dots;\ \forall x \in (0,1)$.

We are now able to state our second main result:

Proposition 2. *Consider a continuous strictly decreasing convex function*

$$\psi : [0,1] \to [0,+\infty) \tag{3}$$

such that $\psi(1) = 0$ and $\lim_{x\to 0^+} \psi(x) = +\infty$. Assume that the dependence between the statistics in $\boldsymbol{t}^s$ is described by an Archimedean copula C_s^ψ with generator ψ, i.e.:

$$\begin{cases} C_s^\psi(u_1,\dots,u_s) = \psi^{-1}\left(\sum_{i=1}^s \psi(u_i)\right); \\ u_k = F_k(x_k), \qquad x_k \in \mathbb{R},\ \forall k = 1,\dots,s. \end{cases} \tag{4}$$

Furthermore, assume that ψ is completely monotone in [0,1].

Then the level q BH procedure controls the FDR at level qm_0/m.

Our third main result considers more general copulas, at the cost of some mildly stronger assumptions.

Condition 2 *Let us introduce a set of $s \times m$ functions*

$$h_{jk} : [0,1] \to [0,1], \qquad j = 1,\dots,m;\ k = 1,\dots,s \tag{5}$$

such that:

(i) h_{jk} is differentiable and strictly increasing, for each j,k;
$(ii)$$h_{jk}(0) = 0$ and $h_{jk}(1) = 1$;
$(iii)\frac{1}{m}\sum_{j=1}^m h_{jk}(x) = x$, for each $k = 1,\dots,s$ and $x \in [0,1]$.

Moreover, define

$$\psi : [0,1] \to [0,1] \tag{6}$$

such that:

(iv) ψ *is* $s+2$ *times differentiable in* $(0,1)$*;*
(v) $\psi^{(i)} > 0$*, for* $i = 1,\ldots,s$*;*
(vi) $\psi(0) = 0$ *and* $\psi(1) = 1$*;*
(vii) $\left(\psi^{-1}\right)^{(s+2)}(x)\left(\psi^{-1}\right)^{(s)}(x) - \left[\left(\psi^{-1}\right)^{(s+1)}(x)\right]^2 \geq 0$*, for each* $x \in (0,1)$*.*

Suppose that

$$\psi(u_k) = h_{jk}^{-1}\left(\frac{e^{u_k}-1}{e-1}\right), \qquad j = 1,\ldots,m;\ k = 1,\ldots,s. \tag{7}$$

Proposition 3. *Assume that Condition 2 holds and that the dependence between the statistics in* $\boldsymbol{t}^s$ *is described by the following asymmetric Archimedean copula:*

$$C_{AS}^{\psi}(u_1,\ldots,u_s) = \psi^{-1}\left(\frac{1}{m}\sum_{j=1}^{m}\prod_{k=1}^{s}h_{jk}(\psi(u_k))\right). \tag{8}$$

Then the level q BH procedure controls the FDR at level q_0/m*.*

3 Conclusion

Propositions 1, 2, and 3 provide *sufficient* conditions for the FDR to hold in the presence of fairly general dependence schemes. In particular, starting from the classical Sklar's Theorem (Sklar, 1959) and the relationship between copulas and correlation, our results allow to explicitly deal with a simple correlation structure among test statistics and the related FDR control.

References

1. Benjamini, Y., Hochberg, Y., (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1), 289-300.
2. Benjamini, Y., Yekutieli, D., (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4), 1165-1188.
3. Sklar, A., (1959). Fonctions de répartition á n dimensions et leurs marges, *Publications de l'Institut de Statistique de L'Université de Paris 8*, 229-231.
4. Yekutieli, D., (2008). False discovery rate control for non-positively regression dependent test statistics. *Journal of Statistical Planning and Inference*, 138(2), 405-415.