Asymptotic behavior of a risk estimate for the FDR-method in the problem of multiple hypothesis testing

S. I. Palionnaya [1], ***O. V. Shestakov*** [2]

[1]Moscow State University, Moscow, Russia, sofiapalionnaya@gmail.com
[2]Moscow State University, Moscow, Russia; Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia, oshestakov@cs.msu.su

The tasks of multiple hypothesis testing about the significance of observations occupy an important place in applied statistics and are used in a wide variety of fields such as genetics, biology, astronomy, computer graphics, etc. In particular, these problems arise when processing multidimensional data in order to identify significant features and remove insignificant (noise) ones. This economical representation of data is extremely important in the processing of audio and video data, genetic chains, encephalograms, spectrograms, etc. In current research on this topic, various filtering methods based on the sparse representation of the obtained experimental data are developed.

There are many statistical procedures that offer different ways to solve the problem of multiple hypothesis testing. In Benjamini and Hochberg [1], a measure called the false discovery rate (FDR) was proposed. This measure assumes to control the expected proportion of false rejections of the null hypothesis. The FDR has become widely used in cases where the number of hypotheses being tested is so large that it is preferable to allow a certain number of errors of the first kind in order to increase the statistical power.

In this report, we study the asymptotic properties of the mean-square risk estimate for the FDR method in the problem of multiple hypothesis testing for the mathematical expectation of a Gaussian vector:

$$X_i = \mu_i + W_i, \quad i = 1, \ldots, n,$$

where W_i are independent normally distributed random variables with zero expectation and known variance σ^2, and $\mu = (\mu_1, ..., \mu_n)$ is an unknown vector belonging to some "sparse" class, i.e. it is assumed that only a relatively small number of its components are significantly large.

We consider the following definitions of sparsity. Let $||\mu||_0$ denote the number of nonzero components of μ. Fixing η_n, define the class

$$L_0(\eta_n) = \{\mu \in R^n : ||\mu||_0 \leq \eta_n n\}.$$

For small values of η_n, only a small number of vector components are nonzero.

Another possible way to define sparsity is to limit the absolute values of μ_i. To do this, consider the sorted absolute values

$$|\mu|_{(1)} \geq ... \geq |\mu|_{(n)}$$

and for $0 < p < 2$ define the class

$$L_p(\eta_n) = \{\mu \in R^n : |\mu|_{(k)} \leq \eta_n n^{1/p} k^{-1/p} \text{ for all } k = 1, ..., n\}.$$

In the considered problem, one of the widespread and well-proven methods for constructing an estimate of μ is the method of (hard) thresholding of each vector component:

$$\hat{\mu}_i = \rho_H(X_i, T) = \begin{cases} X_i, & |X_i| > T, \\ 0, & |X_i| \leqslant T. \end{cases} \tag{1}$$

This procedure is equivalent to testing the hypothesis of zero mathematical expectation for each component of the vector. The penalty method is also widely used, in which the target loss function is minimized with the addition of a penalty term. In a particular case, this method leads to the so-called soft thresholding. The estimates of the vector components are calculated as

$$\hat{\mu}_i = \rho_S(X_i, T) = \begin{cases} X_i - T, & X_i > T, \\ X_i + T, & X_i < -T, \\ 0, & |X_i| \leqslant T. \end{cases} \tag{2}$$

The mean-square error (or risk) of the considered procedures is determined as

$$R(T) = \sum_{i=1}^{n} \mathbb{E}\left(\hat{\mu}_i - \mu_i\right)^2.$$

Note that this expression explicitly depends on the unknown values of μ_i, and it cannot be calculated in practice. However, it is possible to construct its estimate, which is calculated using only the observed data. This estimate is determined by the expression

$$\hat{R}(T) = \sum_{i=1}^{n} F[X_i, T],$$

where $F[X_i, T] = (X_i^2 - \sigma^2)\mathbf{1}(|X_i| \leqslant T) + \sigma^2\mathbf{1}(|X_i| > T)$ for the hard thresholding and $F[X_i, T] = (X_i^2 - \sigma^2)\mathbf{1}(|X_i| \leqslant T) + (\sigma^2 + T^2)\mathbf{1}(|X_i| > T)$ for the soft thresholding.

The report discusses statistical properties of this estimate in the case when the threshold value is selected according to the FDR method of Benjamini and Hochberg. Its strong consistency was proved in Palionnaya [2], and in Palionnaya and Shestakov [3] it was proved that it is also asymptotically normal. Estimates of the convergence rate are also obtained.

Acknowledgements. This research is supported by Russian Foundation for Basic Research, project number 19–07–00352. The research was conducted in accordance with the program of Moscow Center for Fundamental and Applied Mathematics.

References

1. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal Of The Royal Statistical Society Series* **57**:1 (1995) 28–300.

2. S. I. Palionnaya, Strong Consistency of the Risk Estimator in Multiple Hypothesis Testing with the FDR Threshold, *Vestnik Moskovskogo Universiteta, Seriya 15: Vychislitel'naya Matematika i Kibernetika* **4** (2020) 34–39.

3. S. I. Palionnaya, O. V. Shestakov, Asymptotic Properties of MSE Estimate for the False Discovery Rate Controlling Procedures in Multiple Hypothesis Testing, *Mathematics* **8**:11 (2020) 1913.