International Journal of Pure and Applied Mathematics

Volume 42 No. 3 2008, 399-404

ON THE UNBIASEDNESS OF OPTIMAL DISCOVERY PROCEDURE

Igor Melnykov

Department of Mathematics
Colorado State University at Pueblo
2200 Bonforte Blvd., Pueblo, CO 81001, USA
e-mail: igor.melnykov@colostate-pueblo.edu

Abstract: The optimal discovery procedure (ODP) is a new approach to multiple tests of significance that was recently introduced by Storey, see [3]. Similarly to the Neyman-Pearson Lemma in the single hypothesis setting, the ODP was shown to possess optimality when testing multiple hypotheses. We investigate the unbiasedness of this procedure. In particular, we show that it is not unbiased when the tests are performed on the exponential distribution and show the conditions for asymptotic unbiasedness.

AMS Subject Classification: 62H15, 62F03
Key Words: multiple hypothesis testing, optimal discovery procedure, unbiased tests

1. Introduction

Consider the task of testing statistical hypotheses H_i based on the observed data sets $\mathbf{x}_i, i = 1, 2, \ldots, m$. Most traditional testing techniques require that each test statistic is evaluated separately in regards to the evidence that it shows against the null hypothesis. This evidence may be expressed in the form of p-values and the set of hypotheses to be rejected is determined according to the rule that maintains the desired error rate. The key feature of this approach is that all test statistics are processed separately, see [5].

Received: August 17, 2007

Storey (see [3], [4]) proposed a new approach to multiple testing, called the optimal discovery procedure (ODP), that takes into account all observations when evaluating the significance of each test statistic, and proved that this technique is optimal for each fixed number of expected false positives. The ODP is somewhat similar to the Neyman – Pearson Lemma (see [1]) that determines the conditions of existence of the most powerful test.

The ODP function is given by the expression

$$S_{ODP}(\mathbf{x}) = \frac{g_{m_0+1}(\mathbf{x}, \theta_{m_0+1}) + g_{m_0+2}(\mathbf{x}, \theta_{m_0+2}) + \ldots + g_m(\mathbf{x}, \theta_m)}{f_1(\mathbf{x}, \theta_1) + f_2(\mathbf{x}, \theta_2) + \ldots + f_{m_0}(\mathbf{x}, \theta_{m_0})},$$

where $f_i, i = 1, 2, \ldots, m_0$, denote true null densities and g_i, $i = m_0 + 1, m_0 + 2, \ldots, m$, true alternative densities. Hypothesis i is rejected if $S_{ODP}(\mathbf{x}) \geq \lambda$ for some $\lambda \in [0, \infty)$.

In practice, an estimated thresholding function $\hat{S}^*_{ODP}(\mathbf{x})$ is used. Given that all null densities f_i are the same,

$$\hat{S}^*_{ODP}(\mathbf{x}) = \frac{\sum_{i=1}^{m_0} f(\mathbf{x}, \hat{\theta}_i) + \sum_{i=m_0+1}^{m} g_i(\mathbf{x}, \hat{\theta}_i)}{f(\mathbf{x}, \theta_0)},$$

where $f(\mathbf{x}, \hat{\theta}_i)$ and $g(\mathbf{x}, \hat{\theta}_i)$ denote the plug-in estimates of the densities. Similarly, hypothesis i is rejected if $\hat{S}^*_{ODP}(\mathbf{x}_i) \geq \lambda$ for some $\lambda \in [0, \infty)$. In the process of estimation, no distinction has to be made between the null and alternative densities in the numerator, as it contains all m densities.

Storey shows that the testing procedure based on the above estimated ODP statistic is more powerful than the procedure relying on the univariate UMP unbiased tests in a number of scenarios, see [3].

2. ODP and Unbiasedness. Example

Suppose that the i-th data set includes n_i observations. Thus, $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in_i})$, $i = 1, 2, \ldots, m$. For shortness, we will also write

$$\bar{f}(\mathbf{x}, \underline{\theta}) = \frac{1}{m_0} \sum_{i=1}^{m_0} f_i(\mathbf{x}, \theta_i) \text{ and } \bar{g}(\mathbf{x}, \underline{\theta}) = \frac{1}{m - m_0} \sum_{i=m_0+1}^{m} g_i(\mathbf{x}, \theta_i).$$

Note that the dimension of $\underline{\theta}$ depends on the context. Thus, it reflects m_0 entries in $\bar{f}(\mathbf{x}, \underline{\theta})$, but $(m - m_0)$ entries in $\bar{g}(\mathbf{x}, \underline{\theta})$.

Let the outcomes of m hypothesis tests be as shown in Table 1.

A desirable feature of a multiple testing procedure is its unbiasedness (see [2], [3]). In terms of the quantities of Table 1, it is defined as $E\,[V/R] \leq$

	Reject null	Not reject null	Total
True null	V	W	m_0
False null	S	T	$m-m_0$
Total	R	A	m

Table 1: The outcomes of m hypothesis tests

$1-E\,[T/A]$ meaning that the proportion of correct decisions regarding true null hypotheses is at least as large as the proportion of incorrect decisions, on average. The unbiasedness condition can also be rewritten as $E[V]/m_0 \leq E[S]/(m-m_0)$. It is easy to show that both conditions are equivalent.

Writing the second inequality in terms of densities f_i and g_i,

$$\frac{1}{m-m_0}\int_\Gamma \Big(g_{m_0+1}(\mathbf{x},\hat{\theta}_{m_0+1}) + g_{m_0+2}(\mathbf{x},\hat{\theta}_{m_0+2}) + \cdots + g_m(\mathbf{x},\hat{\theta}_m)\Big)\mathrm{d}\mathbf{x}$$
$$\geq \frac{1}{m_0}\int_\Gamma \Big(f_1(\mathbf{x},\hat{\theta}_1) + f_2(\mathbf{x},\hat{\theta}_2) + \cdots + f_{m_0}(\mathbf{x},\hat{\theta}_{m_0})\Big)\mathrm{d}\mathbf{x},$$

where $\Gamma = \{S_{ODP}(\mathbf{x}) \geq \lambda\}$. In the notation with $\bar{f}$ and $\bar{g}$,

$$\int_\Gamma \bar{g}(\mathbf{x},\underline{\hat{\theta}})\mathrm{d}\mathbf{x} \geq \int_\Gamma \bar{f}(\mathbf{x},\underline{\hat{\theta}})\mathrm{d}\mathbf{x}\,.$$

Storey (see [3]) considers the performance of the estimated ODP in the case of a normal distribution $N(\mu,1)$, where $H_i:\mu_i=0$ is tested against the two-sided alternative. In the supporting simulation, the ODP proves more efficient than the UMP unbiased test.

We note that the ODP approach is interpreted nicely for any distribution when the parameter of interest coincides with the mode of the density, such as μ in $N(\mu,1)$. The observations closer to the mode will contribute more to the numerator of $\hat{S}^*_{ODP}(\mathbf{x})$, especially in the situation where many data points from different tests H_i appear near the same value in the close proximity from each other.

The interpretation is not so clear if the density parameter in question is not directly related to the height of the density graph. In particular, this is the case when the density of a test statistic is considerably skewed. As it turns out, the estimated ODP is also not unbiased under these circumstances.

Consider $X_{ij} \sim f(x) = \frac{1}{\lambda}e^{-\frac{x}{\lambda}}, x>0$, as an example. Suppose $H_i:\lambda=1$ is tested against $\lambda \neq 1$. Here, $\hat{\lambda} = \bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$ has $\Gamma(n_i,\lambda/n_i)$ distribution. For simplicity, we assume $n_i=1$ for all i. We show through the simulation that the estimated ODP is not unbiased in the case of exponential distribution.

We consider 2000 hypotheses of which one half are true null and the other half are true alternative. The estimated ODP statistic is

$$\hat{S}^*_{ODP}(x) = \frac{\sum_{i=1}^m \frac{1}{\hat{\lambda}_i} e^{-\frac{x}{\hat{\lambda}_i}}}{e^{-x}} = \sum_{i=1}^m \frac{1}{\hat{\lambda}_i} e^{-x\left(\frac{1}{\hat{\lambda}}-1\right)}.$$

We use $\lambda_1 = 1.05$. $\alpha = 0.05$ leads to the expected number of false positives (EFP) of 50. The simulation was conducted 100,000 times showing the average number of true positives at 49.469. The distribution of the number of true positives (ETP) appears only slightly skewed right. Using a normal approximation, the 95% confidence interval for ETP is (49.409, 49.529). Nonparametric bootstrap for 200 ETP averaged over 5000 iterations each also shows significant results with 95% confidence interval of (49.2592, 49.7956).

λ_1	Average rejections	95% Confidence Interval for ETP
0.99	50.158	(50.098, 50.219)
1.00	49.945	(49.885, 50.005)
1.02	49.668	(49.608, 49.728)
1.05	49.469	(49.409, 49.528)
1.10	49.780	(49.720, 49.839)
1.20	52.381	(52.321, 52.441)

Table 2: Average number of rejections for some values of λ_1

The simulation yields similar results for other values of λ on the interval (1, 1.1) (see Table 2). For values of λ below 1, the power is higher than 5%. While determining the exact values of λ where the procedure is biased would require extensive simulation, it is clear that the estimated ODP is biased for a small range of values of λ just above 1.

In the light of the above example, it is interesting to see under what conditions estimated ODP is unbiased. As the difference between the exact and estimated ODP lies in the values of parameters $\theta_1, \theta_2, \cdots, \theta_m$, intuitively, the quality of the procedure should improve substantially as $\theta_1, \theta_2, \cdots, \theta_m$ are estimated more accurately.

Theorem 1. *Let the densities $f_i(\mathbf{x}, \theta_i), i = 1, 2, \ldots, m_0$ and $g_i(\mathbf{x}, \theta_i), i = m_0 + 1, \ldots, m$ be continuous in their respective parameters θ_i. Let $\hat{\theta}_i$ be consistent estimators of $\theta_i, i = 1, 2, \ldots, m$. Then, the estimated ODP is asymptotically unbiased.*

λ_1	Average rejections	95% Confidence Interval for ETP
0.95	56.942	(56.876, 57.008)
1.02	50.101	(50.041, 50.161)
1.05	53.444	(53.382, 53.506)
1.10	67.184	(67.113, 67.254)

Table 3: Average number of rejections with $n_i = 10$.

Proof. We would like to show that

$$\lim_{\min n_i \to \infty} \int_\Gamma (\bar{g}(\mathbf{x}, \underline{\hat{\theta}}) - \bar{f}(\mathbf{x}, \underline{\hat{\theta}}))\mathrm{d}\mathbf{x} \geq 0.$$

Rewriting the expression inside the integral, we obtain

$$\int_\Gamma (\bar{g}(\mathbf{x}, \underline{\hat{\theta}}) - \bar{f}(\mathbf{x}, \underline{\hat{\theta}}))\mathrm{d}\mathbf{x} = \int_\Gamma (\bar{g}(\mathbf{x}, \underline{\hat{\theta}}) - \bar{g}(\mathbf{x}, \underline{\theta}))\mathrm{d}\mathbf{x}$$

$$+ \int_\Gamma (\bar{g}(\mathbf{x}, \underline{\theta}) - \bar{f}(\mathbf{x}, \underline{\theta}))\mathrm{d}\mathbf{x} + \int_\Gamma (\bar{f}(\mathbf{x}, \underline{\theta}) - \bar{f}(\mathbf{x}, \underline{\hat{\theta}}))\mathrm{d}\mathbf{x}.$$

Let $\epsilon > 0$ and consider the first term in the above sum to write:

$$\int_\Gamma \left| (\bar{g}(\mathbf{x}, \underline{\hat{\theta}}) - \bar{g}(\mathbf{x}, \underline{\theta})) \right| \mathrm{d}\mathbf{x} \leq 2 \cdot P\left(\left| \bar{g}(\mathbf{x}, \underline{\hat{\theta}}) - \bar{g}(\mathbf{x}, \underline{\theta}) \right| \geq \epsilon \right) + \epsilon.$$

Then, by the consistency of $\hat{\theta}_i$ and continuity of g_i,

$$\lim_{\min n_i \to \infty} \int_\Gamma \left| \bar{g}(\mathbf{x}, \underline{\hat{\theta}}) - \bar{g}(\mathbf{x}, \underline{\theta}) \right| \mathrm{d}\mathbf{x} = 0.$$

Similarly,

$$\lim_{\min n_i \to \infty} \int_\Gamma \left(\bar{f}(\mathbf{x}, \underline{\hat{\theta}}) - \bar{f}(\mathbf{x}, \underline{\theta}) \right) \mathrm{d}\mathbf{x} = 0.$$

We note that the exact ODP is unbiased by its optimality. Hence,

$$\lim_{\min n_i \to \infty} \int_\Gamma (\bar{g}(\mathbf{x}, \underline{\hat{\theta}}) - \bar{f}(\mathbf{x}, \underline{\hat{\theta}}))\mathrm{d}\mathbf{x} = \lim_{\min n_i \to \infty} \int_\Gamma (\bar{g}(\mathbf{x}, \underline{\theta}) - \bar{f}(\mathbf{x}, \underline{\theta}))\mathrm{d}\mathbf{x} \geq 0.$$

Thus, the estimated ODP is asymptotically unbiased. □

It can be verified that the procedure is no longer biased in the case of exponential distribution with $n_i = 10$ for all i (see Table 3).

3. Conclusion

Thus, we have shown through the simulation that the optimal discovery procedure is not unbiased when multiple tests on the parameter λ of an exponential distribution are performed. The lack of unbiasedness is due to the estimation of the densities necessary to implement the procedure. However, this can be overcome through the more precise estimation of the pdfs. In fact, we have shown that the ODP is asymptotically unbiased under certain assumptions.

References

[1] J. Neyman, E.S. Pearson, On the problem of the most efficient tests of statistical hypotheses, *Philosophical Transactions of the Royal Society*, **81** (1933), 289-337.

[2] S. Sarkar, FDR-controlling stepwise procedures and their false negative rates, *Journal of Statistical Planning and Inference*, **125** (2004), 119-137.

[3] J. Storey, The optimal discovery procedure: a new approach to simultaneous significance testing, *Journal of the Royal Statistical Society: Series B*, **69**, No. 3 (2007), 347-368.

[4] J. Storey, J.Y. Dai, J. Leek, The optimal discovery procedure for large-scale significance testing with applications to comparative microarray experiments, *Biostat.*, **8** (2007), 414-432.

[5] P.H. Westfall, S. Young, *Resampling-Based Multiple Testing*, Wiley (1993).