Test (2008) 17: 461–471
DOI 10.1007/s11749-008-0134-6

DISCUSSION

Rejoinder on: Control of the false discovery rate under dependence using the bootstrap and subsampling

Joseph P. Romano · Azeem M. Shaikh · Michael Wolf

Published online: 9 December 2008

We are extremely appreciative of the insightful comments made by all the responders. The goal of constructing useful multiple testing methods which control the false discovery rate and other measures of error is currently a thriving and important area of research. On the one hand, the bootstrap method presented in the present work seems to work quite well and is supported by some theoretical analysis. On the other hand, many more important practical, computational, and mathematical questions remain, some of which are addressed by the responders and which we touch upon below.

We also appreciate the added references, which help to provide a more thorough discussion of the available methods. Our paper was the development of a particular methodology and was by no means a comprehensive account of the burgeoning FDR literature.

This rejoinder is discussed in the comments available at:
http://dx.doi.org/10.1007/s11749-008-0127-5, http://dx.doi.org/10.1007/s11749-008-0128-4,
http://dx.doi.org/10.1007/s11749-008-0129-3, http://dx.doi.org/10.1007/s11749-008-0130-x,
http://dx.doi.org/10.1007/s11749-008-0131-9.

J.P. Romano
Departments of Economics and Statistics, Stanford University, Stanford, USA
e-mail: romano@stanford.edu

A.M. Shaikh
Department of Economics, University of Chicago, Chicago, USA
e-mail: amshaikh@uchicago.edu

M. Wolf (✉)
Institute for Empirical Research in Economics, University of Zurich, Bluemlisalpstrasse 10,
8006 Zurich, Switzerland
e-mail: mwolf@iew.uzh.ch

1 Reply to José Ferreira and Mark A. van de Wiel

The non-null values $\theta_j = 0.2$ were chosen as an intermediate case between two non-interesting extremes: (i) if θ_j is very large, the corresponding H_j will be rejected with probability (almost) equal to one for all methods, and so there is little distinction in terms of power; (ii) if θ_j is very close to zero, H_j will be rejected with very small probability for all methods, and so, again, there is little distinction in terms of power. By trial and error, the value $\theta_j = 0.2$ was found to be an interesting middle ground. On the other hand, we understand the concern about the performance of our method for a sequence of alternatives which approach the null in a continuous fashion. To shed some light on this issue, we repeated the simulations, restricting attention to the scenario of common correlation, for the values $\theta_j = 0.1$ and 0.01. The results can be found in Tables 1 and 2. The average number of rejections naturally declines with θ_j, but qualitatively the results do not really change very much.

Concerning the empirical distribution of the p-values generated under the null: these p-values were computed using the t_{n-1} distribution for the studentized test statistics. Since under the null, $\theta_j = 0$, as opposed to $\theta_j < 0$, in our simulation set-up, the null test statistics have exactly this t_{n-1} distribution, and so the null p-values have exactly a uniform [0, 1] distribution. We therefore did not feel the need to give some information about the empirical distribution of the null p-values.

We were also quite surprised by how badly the STS version of the BH method does when the data are dependent. The choice of $\lambda = 0.5$ (or what the discussants call $x = 0.5$) may well be partly responsible. However, we would like to point out that we simply used the "default" value of Storey et al. (2004) rather than deliberately choosing a value of λ which makes the STS version look bad. The question of whether a different choice of λ might lead to a better performance is a very good one. This issue is also addressed by S. Sarkar and R. Heller who argue that the choice $\lambda = \alpha/(1+\alpha)$ results in more reliable FDR control under dependence. We redid Table 1 of the paper, replacing STS by STS*, where the latter uses $\lambda = 0.1/1.1$; see

Table 1 Empirical FDRs expressed as percentages (in the rows "Control") and average number of false hypotheses rejected (in the rows "Rejected") for various methods, with $n = 100$ and $s = 50$. The nominal level is $\alpha = 10\%$. The number of repetitions is 5,000 per scenario, and the number of bootstrap resamples is $B = 500$

	$\sigma_{i,j} = 0.0$				$\sigma_{i,j} = 0.5$				$\sigma_{i,j} = 0.9$			
	BH	STS	BKY	Boot	BH	STS	BKY	Boot	BH	STS	BKY	Boot
Ten $\theta_j = 0.1$												
Control	8.1	9.7	7.4	7.5	5.6	15.8	5.7	7.7	4.8	27.3	5.3	9.7
Rejected	0.4	0.5	0.4	0.4	0.8	2.1	0.8	1.0	0.9	3.4	0.9	2.2
Twenty five $\theta_j = 0.1$												
Control	5.1	7.6	4.8	4.3	4.7	7.9	4.6	4.4	4.3	11.1	4.8	6.1
Rejected	1.6	2.8	1.5	1.5	1.7	3.5	1.7	1.7	2.6	6.3	2.8	3.5
All $\theta_j = 0.1$												
Control	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Rejected	5.5	23.8	5.6	5.4	6.0	24.2	6.5	6.4	8.0	27.5	9.8	11.9

Table 2 Empirical FDRs expressed as percentages (in the rows "Control") and average number of false hypotheses rejected (in the rows "Rejected") for various methods, with $n = 100$ and $s = 50$. The nominal level is $\alpha = 10\%$. The number of repetitions is 5,000 per scenario, and the number of bootstrap resamples is $B = 500$

	$\sigma_{i,j} = 0.0$				$\sigma_{i,j} = 0.5$				$\sigma_{i,j} = 0.9$			
	BH	STS	BKY	Boot	BH	STS	BKY	Boot	BH	STS	BKY	Boot
Ten $\theta_j = 0.01$												
Control	8.1	8.3	7.3	8.1	5.3	13.4	4.9	7.8	4.0	26.6	3.6	7.8
Rejected	0.03	0.03	0.03	0.03	0.19	1.26	0.23	0.30	0.47	3.3	0.45	0.76
Twenty five $\theta_j = 0.01$												
Control	4.7	4.9	4.3	4.7	4.8	5.4	4.4	5.0	3.5	6.8	3.3	5.1
Rejected	0.08	0.09	0.08	0.08	0.10	0.14	0.09	0.10	0.37	1.88	0.39	0.54
All $\theta_j = 0.01$												
Control	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Rejected	0.17	0.20	0.16	0.17	0.20	0.33	0.19	0.20	0.63	3.84	0.69	0.95

Table 3 Empirical FDRs expressed as percentages (in the rows "Control") and average number of false hypotheses rejected (in the rows "Rejected") for various methods, with $n = 100$ and $s = 50$. The nominal level is $\alpha = 10\%$. The number of repetitions is 5,000 per scenario, and the number of bootstrap resamples is $B = 500$

	$\sigma_{i,j} = 0.0$				$\sigma_{i,j} = 0.5$				$\sigma_{i,j} = 0.9$			
	BH	STS*	BKY	Boot	BH	STS*	BKY	Boot	BH	STS*	BKY	Boot
All $\theta_j = 0$												
Control	10.0	10.0	9.1	10.0	6.4	8.3	6.0	9.9	4.8	8.5	4.4	9.8
Rejected	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ten $\theta_j = 0.2$												
Control	7.6	9.3	7.3	7.3	6.4	9.3	7.5	9.3	5.0	8.1	5.8	10.0
Rejected	3.4	3.7	3.4	3.4	3.5	3.7	3.5	4.1	3.7	3.8	3.7	6.0
Twenty five $\theta_j = 0.2$												
Control	5.0	7.8	6.2	6.7	4.3	8.6	7.4	8.9	3.9	8.0	7.1	9.5
Rejected	13.2	16.2	14.5	14.9	12.3	14.3	13.1	14.0	12.6	15.1	12.7	16.6
All $\theta_j = 0.2$												
Control	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Rejected	34.8	46.2	44.9	48.2	31.9	39.9	36.4	39.1	32.1	37.9	32.1	36.4

Table 3. Similar to the simulations carried out by Sarkar and Heller, STS* successfully controls the FDR in all scenarios considered and dominates both BH and BKY in terms of power. Compared to Boot, it is a bit more powerful for $\rho = 0$. Under positive dependence, there is no clear ranking. Depending on the value of $\rho > 0$ and the number of false hypotheses, either method can be more powerful than the other. Of course, SKS* is computationally much less expensive than Boot, which is an im-

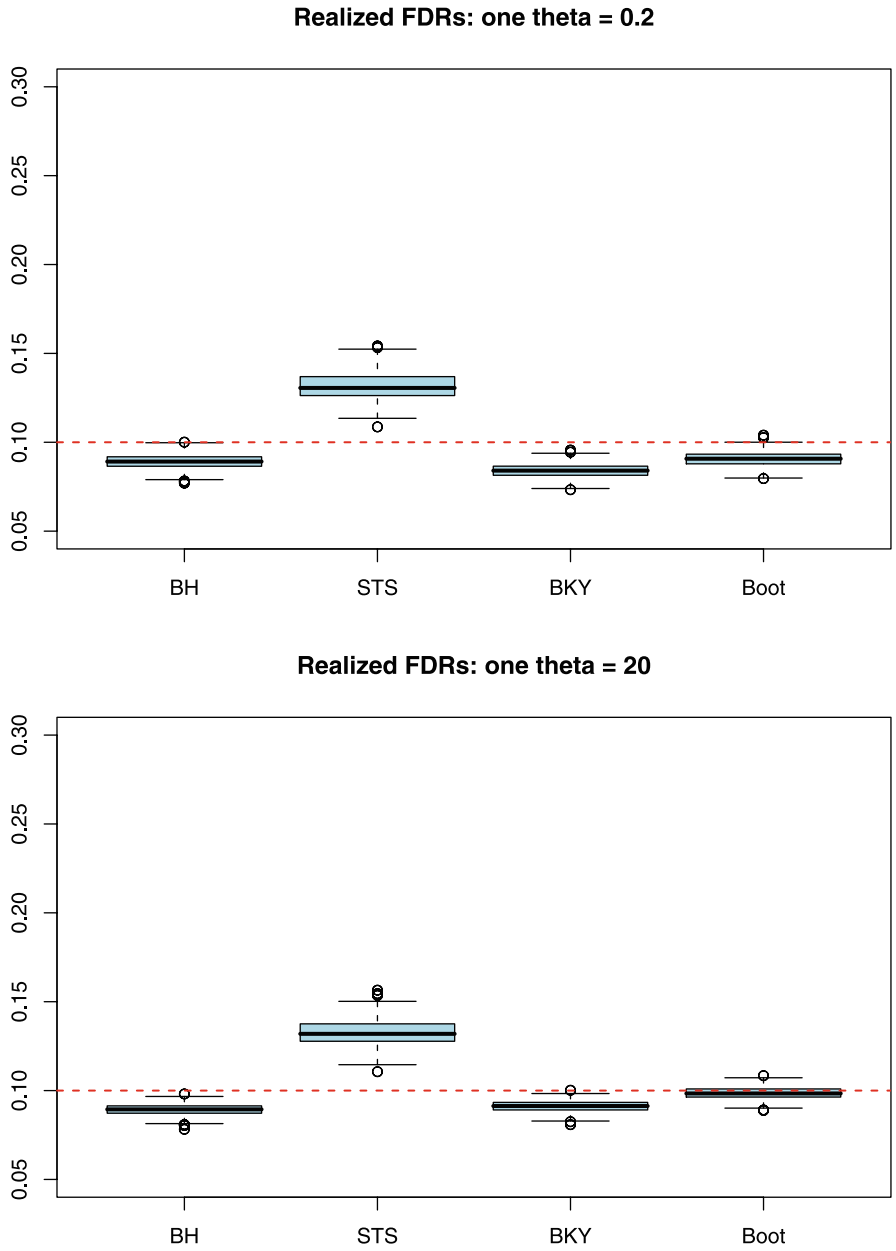

Fig. 1 Boxplots of the simulated FDRs similar to those described in Sect. 7.2, except that we use $s = 10$ instead of $s = 4$ hypotheses now. The horizontal dashed lines indicate the nominal level $\alpha = 0.1$

portant practical advantage, especially when s is very large. There may well be other methods to come up with estimates of s_0 that take the dependence structure into account, say via resampling, but this is beyond the scope of this reply.

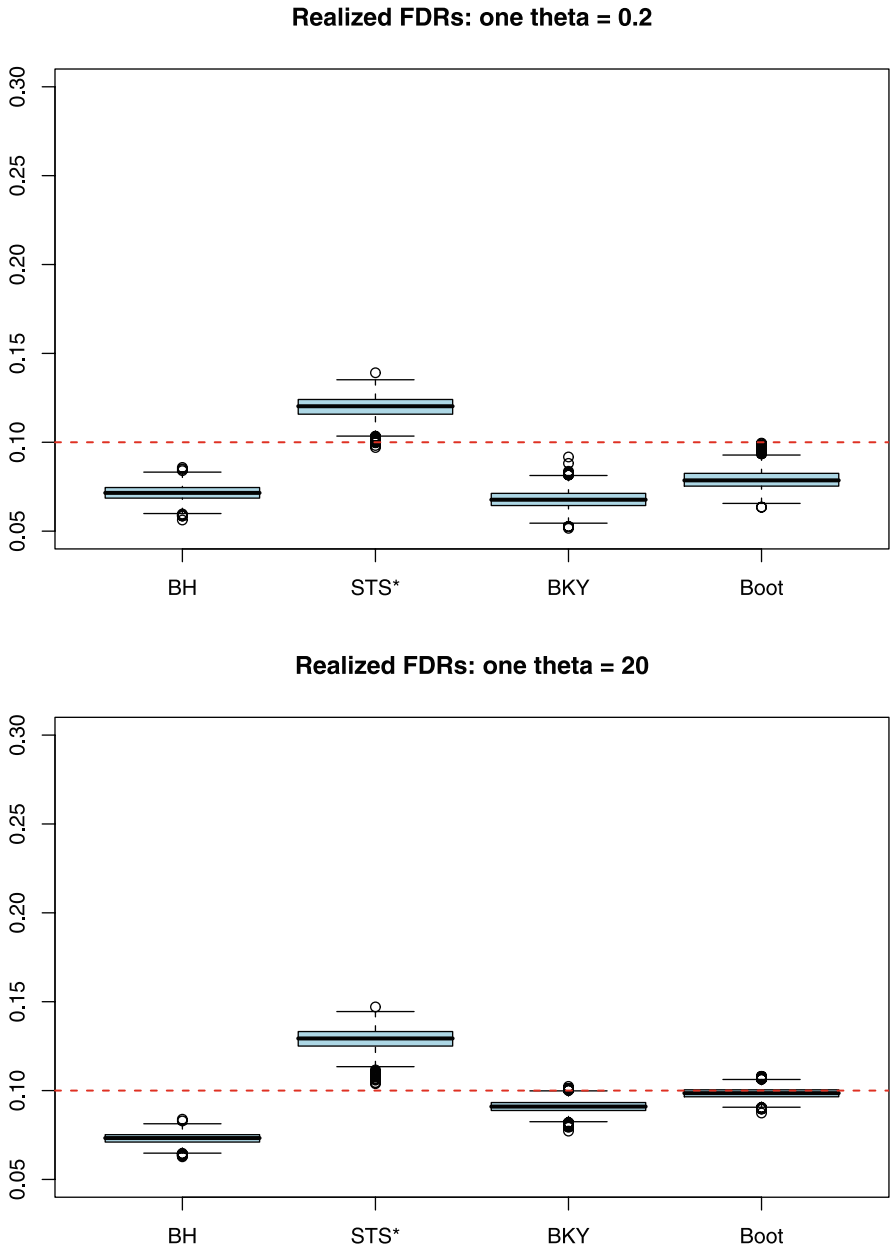

Fig. 2 Boxplots of the simulated FDRs similar to those described in Sect. 7.2, except that STS is replaced by STS* now. The horizontal dashed lines indicate the nominal level $\alpha = 0.1$

Concerning the simulations in Sect. 7.2, there were actually two reasons for the choice $s = 4$. On the one hand, we wanted to cover the space of random correlation matrices "more thoroughly." On the other hand, something like $s = 50$ is computationally infeasible. Unfortunately, the computational burden of our method is a draw-

back. While simulating a single scenario with $s = 50$ is no problem, doing it 1,000 times over (for 1,000 different correlation matrices) would take weeks. However, we were able to at least redo the exercise for the larger value $s = 10$; see Fig. 1. In terms of the FDR control of our bootstrap method, the results do not change qualitatively compared to $s = 4$. So there is reason to hope that they would continue to hold for $s = 50$, say. However, note that there is generally a reduced variation in the boxplots, especially for STS. This indicates that indeed, we cover the space of random correlation matrices "less thoroughly" for $s = 10$. For example, while all the realizations for STS lie above 0.1, we know that for some correlation matrices, the FDR is actually successfully controlled (e.g., for the identity matrix). On the other hand, while all the realizations lie below 0.2, we know that for some correlation matrices, the FDR is actually higher than that (e.g., for the constant correlation matrix with correlation close to one). So in some sense the plot for $s = 4$ is indeed more informative.

In view of the above discussion, we repeated the exercise, keeping $s = 4$ but replacing STS by STS*; see Fig. 2. It is seen that STS* is more conservative than STS but still fails to generally control the FDR. Therefore, if one wishes to use a method based on the marginal p-values and is ignorant about the underlying dependence structure, it might be safer to use BKY rather than STS*.

Finally, we agree with you that Remark 1 could be clearer, and we wish we had the possibility of reviews before the final version. In any case, for the benefit of new readers, the values of both $T_{n,r:t}$ and $T^*_{n,r:t}$ are always meant to be ordered so that they are nondecreasing as r increases.

2 Reply to Wenge Guo

2.1 High-dimensional, low sample size data analysis

We agree that for many applications, these days the number of hypotheses, s, is very large, while the number of data points, n, is very small (at least in comparison). Our bootstrap procedure was not designed for such applications. At this point, the justification of our methods is based on the assumption that $n \to \infty$ while s remains fixed. Also, mild assumptions are imposed on the data generating mechanism P, from which it follows that all false null hypotheses will be rejected with probability tending to one. Arguably, such assumptions are problematic when $s = 2{,}000$ and $n = 10$, for example, which might be considered a "typical" combination for microarray data.

Contamination with outliers, which is quite common for microarray data, is a severe problem for our procedure, at least when non-robust test statistics are used, such as the usual t-statistic. However, the problem lies more with these outliers *not* appearing in bootstrap resamples. Take the case of a single small sample that is "well behaved," apart from a solitary, very large outlier. The t-statistic, for testing the null hypothesis that the population mean is zero, will be close to one in absolute value (as the outlier gets larger and larger in absolute value). Whenever the outlier does not appear in the bootstrap sample, the bootstrap t-statistic—centered by the sample mean of the original data rather than zero—will be large in absolute value, and this happens with probability $(1 - \frac{1}{n})^n \approx 1/e \approx 0.38$. So the bootstrap test, applied to this single

sample, will not reject the null at any customary significance level, just like the usual t-test. Now consider a multiple testing set-up. Our bootstrap method is a stepdown procedure and the "first" critical value (that is, the critical value used to compare the largest of the test statistics) is the $1-\alpha$ quantile of the sampling distribution of the largest bootstrap t-statistic $T^*_{n,(s)}$. Even a single data set with a very large outlier, out of all s individual data sets, can dominate the sampling distribution of this maximum, leading to large critical value. As a result, not even a single hypothesis might get rejected. It is plausible that stepup methods are more robust in this sense. Unfortunately, no bootstrap stepup methods have been suggested in the literature at all so far, not even for the more traditional FWER. This appears, therefore, an important topic for future research.

On the other hand, the fact that stepup procedures based on individual p-values are more robust, in their ability to make rejections at all, to very large outliers in individual samples, does not necessarily mean that they will lead to reliable inference, at least when based on non-robust individual test statistics such as the usual t-statistic. It might be worthwhile to explore suitable robust test statistics as an alternative.

2.2 Computational problem

We agree that the main drawback of the bootstrap method is its computational burden. We are grateful for the suggestions to improve matters. However, consider expression (2). As pointed out, for the bth bootstrap data set, one has basically to compute the number of rejections determined by the critical constants $\hat{c}_i$, $i = 1, \ldots, j-1$, and the ordered test statistics $T^{*b}_{i:j}$, $i = j-1$. For a given value of c, this number, denoted by r^{*b}_j, together with $T^{*b}_{j:j}$, determines the contribution of the bth bootstrap sample to the expression $\mathrm{FDR}_{j,\hat{P}}(c)$. Actually, our software implementation is really comparable in computational complexity to this suggestion. So, unfortunately, things could not be sped up significantly along these lines.

The number of bootstrap repetitions, B, is not all that crucial in successfully controlling the FDR. Note that in our simulations we only used $B = 200$. On the other hand, consider two researchers applying the method to the same data set, both using the same value of B but a different random number generator (or a different seed value). It may well happen that, due to the randomness of the critical values which are computed sequentially, the two researchers might obtain quite different results in terms of the rejected hypotheses. It is therefore indeed desirable to pick B as large as possible, given the computational resources.

2.3 Some possible extensions

We agree that bootstrap stepup methods should be less sensitive to a few extreme outliers or a large number of skewed data sets, as typical with microarray data. However, to the best of our knowledge, no such methods have been developed yet in the multiple testing literature, even for the presumably simpler problem of controlling the FWER (at least not in the nonparametric setting under weak conditions). This remains an exciting field for future research.

As pointed out, the computation of the critical values progresses from the "bottom up" rather than "top down." The latter would save much time in case the number of false hypotheses is relatively small. Unfortunately, we have not yet been able to come up with a "top down" method.

At this point, if the number of hypotheses is very large compared to the sample size, we would not be comfortable with applying the bootstrap method. In such applications, it is probably safer to use methods based on the marginal p-values. But as much effort as possible should be made to ensure that the distribution of the null p-values is as close as possible to the uniform [0,1] distribution in finite samples. Using the usual t-test to compute individual p-values in the presence of extreme outliers or skewed data, combined with small sample sizes, does not appear prudent, yet it seems quite common in practice.

It would be very desirable to develop bootstrap methods that provide error rate control (whether FWER, FDP, or FDR) under more general asymptotics where the number of hypotheses is allowed to tend to infinity together with the sample size. This appears a very challenging task, but we hope to make some progress here in future research.

3 Reply to James F. Troendle

We fully agree that for many, if not most, applications, it would be preferable to control the FDP rather than the FDR. As pointed out, by controlling an expected value, one cannot really say anything of much use about the realized FDP for a given data set. (Of course, one can apply Markov's inequality to get some crude information; see (34) of Lehmann and Romano (2005a). In this sense, it is indeed unfortunate to see that many researches use FDR controlling methods and then interpret their results as if they had actually controlled the FDP instead.

However, control of the FDR is widespread, while control of the FDP is still used comparatively rarely. We hope that this will change over time. In the meantime, and also for those applications where control of the FDR might actually be preferred, we tried to develop a resampling method to account for the unknown dependence structure in order to improve power or the ability to detect false null hypotheses.

Notably, in our own research, we have worked on resampling methods for FDP control first; see Romano and Wolf (2007) and Romano et al. (2008). In the latter paper, inspired by the example in Korn et al. (2004), we also addressed the tail behavior of the realized FDP under FDR control. It was seen that, especially under strong dependence, high values of the FDP can become very likely, even though the FDR is perfectly controlled.

We also agree that there is potential for the subsampling method when the sample size is much larger than one considered in our simulation study, that is, $n = 100$. It is interesting that, even in testing problems involving mean-like parameters and statistics, the asymptotic behavior of the bootstrap and subsampling method are quite distinct in the behavior of critical values. Usually, their first-order asymptotic behavior is the same, but not in the setting of the present paper. It is also frustrating that we could not justify the bootstrap without the exchangeability assumption, even though

this assumption is not needed for subsampling. Future research will be dedicated to these issues.

4 Reply to Sanat K. Sarkar and Ruth Heller

In the setting of our paper, weak assumptions are imposed on the mechanism generating the data, denoted by P, with the number of data points n asymptotically tending to ∞ while the number of tests s remains fixed. It is a consequence of these assumptions (rather than a basic assumption) that all false null hypotheses are rejected with probability tending to one. As Sarkar and Heller point out, the false discovery rate, which is indeed both a function of n and P, now denoted $\mathrm{FDR}_{n,P}$, behaves asymptotically like their expression (1).

In order to interpret our asymptotic results, let us be clear. As pointed out, our results do not imply that there exists a sufficiently large $n_0 = n_0(\alpha)$ such that $\mathrm{FDR}_{n,P} \leq \alpha$ for all $n \geq n_0$. The actual statement is that, for any $\epsilon > 0$, there exists a sufficiently large $n_0 = n_0(\alpha, P)$ such that $\mathrm{FDR}_{n,P} < \alpha + \epsilon$ for all $n \geq n_0(\alpha, P)$. Notice that $n_0(\alpha, P)$ depends on the unknown P; that is, our asymptotic analysis is pointwise in P. Uniform asymptotic convergence over a broad class $\mathbf{P}$ of P would demand that n_0 not depend on $P \in \mathbf{P}$. The distinction between pointwise and uniform convergence in the case of single testing is discussed in Sect. 11.1 of Lehmann and Romano (2005b). Since P is unknown, the stronger uniform convergence results are generally more desirable, though they require additional arguments and sometimes do not hold (for example, as a consequence of the Bahadur–Savage result). Although we did not prove the stronger uniform convergence result in this paper, for the special case where the test statistics are studentized sample means like those considered in the simulations, we expect our results to hold uniformly over a broad class $\mathbf{P}$. In the single testing case, one restriction is that the underlying family of distributions have a uniformly bounded $2 + \delta$ moment, and a weaker condition is given in (11.77) in Theorem 11.4.4 of Lehmann and Romano (2005b). A multivariate extension of that theorem that is relevant for the multiple testing situation studied here is given in Lemma 3.1 of Romano and Shaikh (2008).

A certain limitation of our theoretical analysis is the assumption that n gets large while s remains fixed. We should mention that some literature has considered the large s situation; see, for example, Genovese and Wasserman (2004), Storey et al. (2004), and Efron (2008). However, note that, in some ways, the problem of large s is made easier by stronger assumptions and by the ability to average out errors over many tests. For instance, with the commonly used mixture model, the tests cannot be that different from one another in that their average behavior must settle down, so that, for example, the density of the distribution of test statistics corresponding to false null hypotheses is the same for all such test statistics and can therefore be estimated by usual techniques. Our goal here was to see what can be accomplished in a more general setting which allows for a great deal of heterogeneity (in the sense that the limiting covariance matrix of the test statistics is quite general), but with s fixed.

Sarkar and Heller present an interesting derivation of the stepdown procedure of Gavrilov et al. (2008) as an adaptive stepdown analog of the Benjamini–Hochberg

procedure. The procedure is adaptive in that it modifies the BH procedure by incorporating an estimate of the number of true null hypotheses s_0. Interestingly, the resulting stepdown critical constants, given by (2) in the discussion of Sarkar and Heller, are nonrandom, even though the motivation was based on incorporating a data-dependent estimate of s_0.

We appreciate the discussion of the choice of $\lambda = 0.5$. We also redid some of our simulations, using your suggestion of $\alpha/(1+\alpha)$; see our above rejoinder to Ferreira and van de Wiel.

Sarkar and Heller summarize the use of augmentation methods suggested by Pacifico. et al. (2004) and Dudoit and van der Laan (2008). Our experience with these methods is that they are not as powerful as other resampling methods we have considered, at least in the context of other error rates; see the comparisons in Romano and Wolf (2007). While augmentation is a general approach that exploits the relationship between the familywise error rate and a given generally weaker measure of error control, it appears that the idea behind augmentation is too crude in that the construction does not really make full use of the given measure of error control desired. Nor does it take into account the dependence structure in the problem, outside the first stage where control of the familywise error rate is used. Indeed, after the first stage, a given number of additional hypotheses are rejected at the second stage, and this number only depends on the number of rejections at the first stage and not, for example, on the dependence structure of the remaining test statistics to be tested.

Finally, it would be interesting to improve the procedure, perhaps by incorporating an estimate of s_0. An alternative but similar approach might first apply some kind of thresholding (say by a familywise error rate controlling procedure) to reduce the number of hypotheses under consideration.

5 Reply to Daniel Yekutieli

Of course, we wish we could propose a method with finite sample validity which implicitly or explicitly accounts for the dependence structure in the problem. Unfortunately, even in single testing, this is usually too much to hope for in nonparametric problems, but we believe that resampling methods can still be quite useful and reliable with sufficiently informative data. Of course, we point out the obvious fact that, in order for the BH procedure, or any other procedure which claims finite sample control based on marginal p-values, to truly exhibit finite sample control, the p-values must be exact in the sense of (1) in the paper. Of course, this requirement is almost never satisfied in practice, as p-values often rely on either asymptotic or resampling approximations.

Apparently, it is indeed quite challenging to construct a reasonable scenario where the Benjamini–Hochberg (BH) method fails to control the FDR. However, suppose we are in a situation where the exact sampling distribution of the test statistics is multivariate normal with a known covariance matrix Σ, which corresponds to an asymptotic approximation of the problem studied here. In the case $s = 2$ with both null hypotheses true and with negative correlation between the test statistics, control of the BH method reduces to the validity of Simes inequality. In this case, it is known

to fail; see, for example, Samuel-Cahn (1996) for a counterexample in the one-sided case. To the best of our knowledge, it is not known in general whether the BH method ever fails in the two-sided case, even if the covariance matrix exhibits extreme negative dependence. The statement that the FDR of the BH method approaches $\alpha s_0/s$ for large n and any P seems unsubstantiated, unless one has further knowledge of the limiting covariance matrix Σ. The validity of the BH method for multivariate normal test statistics in the two-sided case is interesting and deserves further thought. Certainly, a highlight of our work is that no assumptions are required on the limiting covariance matrix, in either the one- or two-sided cases.

Yekutieli's argument for the conservatism of FDR controlling procedures when the non-null tested effects are small is nice. The problem is essentially reduced to the study of control of the FDR under the complete null hypothesis when all null hypotheses are true. However, the argument does assume exchangeability, and one must know that the given method controls the FDR under the complete null. Of course, the BH method may not do so in general, and one is left with deciding which method is most appropriate.

To be clear, we do not assume that all false null hypotheses are rejected with probability tending to one; rather, it is a proven consequence of very basic assumptions concerning the limiting behavior of the test statistics under the fixed known data generating mechanism P. A more complete asymptotic framework would consider uniformity with respect to P, as well as s getting large (as discussed above in the response to Sarkar and Heller).

References

Dudoit S, van der Laan MJ (2008) Multiple testing procedures with applications to genomics. Springer series in statistics. Springer, New York

Efron B (2008) Microarrays, empirical Bayes and the two-groups model. Stat Sci 23:1035–1061

Gavrilov Y, Benjamini Y, Sarkar SK (2008) An adaptive step-down procedure with proven FDR control. Ann Stat (in press)

Genovese CR, Wasserman L (2004) A stochastic process approach to false discovery control. Ann Stat 32(3):1035–1061

Korn EL, Troendle JF, McShane LM, Simon R (2004) Controlling the number of false discoveries: application to high-dimensional genomic data. J Stat Plann Inference 124:379–398

Lehmann EL, Romano JP (2005a) Generalizations of the familywise error rate. Ann Stat 33(3):1138–1154

Lehmann EL, Romano JP (2005b) Testing statistical hypotheses, 3rd edn. Springer, New York

Pacifico. M, Genovese. C, Verdinelli. I, Wasserman. L (2004) False discovery control for random fields. J Am Stat Assoc 99:1002–1014

Romano JP, Shaikh AM (2008) Inference for identifiable parameters in partially identified econometric models. J Stat Plann Inference 138:2786–2807

Romano JP, Wolf M (2007) Control of generalized error rates in multiple testing. Ann Stat 35(4):1378–1408

Romano JP, Shaikh AM, Wolf M (2008) Formalized data snooping based on generalized error rates. Econom Theory 24(2):404–447

Samuel-Cahn E (1996) Is the Simes improved Bonferroni procedure conservative? Biometrika 83:928–933

Storey JD, Taylor JE, Siegmund D (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. J R Stat Soc Ser B 66(1):187–205