False Discovery Control With P-Value Weighting

Christopher R. Genovese, Kathryn Roeder and Larry Wasserman[1]
Department of Statisstics
Carnegie Mellon University
October 28, 2004

We present a method for multiple hypothesis testing that maintains control of the False Discovery Rate while incorporating prior information about the hypotheses. The prior information takes the form of p-value weights. If the assignment of weights is positively associated with the null hypotheses being false, the procedure improves power, except in cases where power is already near one. Even if the assignment of weights is poor, power is only reduced slightly, as long as the weights are not too large. We also provide a similar method to control False Discovery Exceedance.

1 Introduction

Data from DNA microarray experiments, genetic epidemiology, functional Magnetic Resonance Imaging (fMRI) experiments, and astronomical imaging have spurred renewed interest in the multiple testing problem – controlling overall error rates when performing simultaneous hypothesis tests. These data sets share two features that distinguish them from multiple-testing examples in traditional linear models. First, there are often many thousands, even millions, of null hypotheses to test. This exacerbates the trade-off between power and Type I error control, making it more difficult to detect small effects, which are often of the greatest interest. With fMRI experiments, for example, applying methods such as Bonferroni to control the familywise error rate (FWER) often wipes out any evidence for significant effects. Second, the tests are related by a scientifically meaningful structure. In fMRI, each test corresponds to a specific brain location; in microarray studies, each test corresponds to a specific gene. Both scientific and spatial prior information can thus be exploited to improve the performance of testing procedures. Put simply, all null hypotheses are not created equal.

The introduction of the False Discovery Rate (FDR) and a procedure to control it by Benjamini and Hochberg (BH, 1995) gave an effective way to address the first issue above. FDR control lets investigators increase power while maintaining a principled bound on error. The BH procedure is fast and easy to compute even with large data sets and performs well

[1]This work was partially supported by funding from National Institutes of Health grants NS047493-01 (CG) and MH057881 (KR, LW) NSF Grant DMS-0104016 (CG, LW).

in sparse cases where there are relatively few true alternatives (Genovese and Wasserman, 2002). Let $P_{(1)} < \ldots < P_{(m)}$ be the ordered p-values from m hypothesis tests, with $P_{(0)} \equiv 0$. Then, the BH procedure rejects any null hypothesis for which $P \leq T$ with

$$T = \max\left\{P_{(i)} : \ P_{(i)} \leq \frac{\alpha i}{m}\right\}; \tag{1}$$

this controls the FDR at level $\alpha m_0/m$, where m_0 is the number of true null hypotheses. Adaptive variants of the BH procedure can increase power further at little additional computational expense; see Benjamini, Krieger, and Yekutieli (2004) and Storey (2002).

But neither the BH procedure nor its variants deal with the second issue above, structure and prior information, because they treat all null hypotheses interchangeably. For example, previous studies can suggest that some null hypotheses are more (or less) likely to be false. Similarly, in spatial problems, the false nulls are more likely to be clustered than true nulls. In this paper, we consider the potentially powerful approach of expressing prior information through weights on each null hypothesis.

The idea of weighting hypotheses is not new. We distinguish two approaches: *p-value weighting*, as above, and *loss weighting*, where each weight is placed on the loss or error criterion for the corresponding incorrect decision. Holm (1979) introduced the idea of p-value weights, describing them as "positive constants indicating the *importance* of the hypotheses...". A larger weight can be used to suggest it is more likely that the null hypothesis is false a priori. Holm (1979) showed that his sequential step-down test maintains control of the family-wise error rate when the p-values are divided by weights, as long as the step-down constants are adjusted appropriately. Benjamini and Hochberg (1997) investigated the use of weighting in a variety of settings. They used weights in the definition of the error rate (loss weighting) to indicate the importance of each hypothesis. Here, we use p-value weighting as a frequentist method for including prior information about the hypotheses, leaving the error measure unchanged.

Such prior information is often available in practice. In fMRI studies, for example, information on the functional response to a stimulus can be gleaned from previous studies, pilot data, and direct neural recording in animals. Detailed anatomical information is also available from structural images of each subject. Similarly, in genetic epidemiology tens of

thousands of genomic regions may be tested in a genetic association study to locate alleles that increase the risk for complex diseases. Frequently the association study is conducted after genetic linkage studies have been published. In contrast to an association study, which is designed to pinpoint genetic variants associated with disease, linkage analysis points to very broad regions of the genome that appear to contain genetic variants of interest. These regions often contain tens or even hundreds of genes. Initially it might seem that such information would not be refined enough to offer reliable weights; however, coupled with the partial knowledge of genetic function available from the human genome project, linkage studies are likely to provide useful guidance for choosing weights in an association study.

In general, p-value weighting raises several important questions. How can we choose weights so as to maintain control of a suitable error criterion, such as FDR? How much power can we gain if we guess well in the weight assignment? How much power can we lose if we guess poorly? In this paper, we will present a p-value weighting procedure that controls FDR. We will show that under moderately informative guessing, weighting improves power nontrivially and that under even mis-informative guessing, the worst-case loss in power is small. We also explore the role of weights when controlling False Discovery Exceedance (FDX; Genovese and Wasserman 2004a, 2004b; and van der Laan, Dudoit, and Pollard 2004). The reader interested primarily in our procedures can read only Section 3 for FDR control and Section 5 for FDX Control.

2 P-value Weighting

Consider the simplest case where, based on previous studies and results, an investigator can partition the m null hypotheses into two groups, where the null is a priori more plausible in one and the alternative in the other. In this setting is seems reasonable to consider using different thresholds for hypotheses in each of the two groups. If the (random) thresholds are T_0 and T_1, say, and $j(i)$ is the group for the i^{th} null hypothesis, we can write $P_i \leq T_{j(i)}$ as $P_i \leq W_i T$, or equivalently $P_i/W_i \leq T$, where $T = (T_0 + T_1)/2$ and $W_i = 2T_{j(i)}/(T_0 + T_1)$. Thus using different thresholds for the groups corresponds to using a single threshold but weighting the p-values. Note that when $T_0 \neq T_1$, the weights will be bigger than 1 in one

group and less than 1 in the other. There is, of course, no restriction to binary weighting schemes in general.

Whatever information one uses to construct p-value weights, the weight assignment remains a guess. We treat this guess as if it is made a priori, that is, before seeing the p-values, and for the purposes of analysis, we model the weights as random variables that are related to the underlying truth or falsehood of each null hypothesis.

Let $P^m = (P_1, \ldots, P_m)$ denote the observed p-values, with $P_{(1)} < \ldots < P_{(m)}$ denoting the ordered p-values and $P_{(0)} \equiv 0$. Define hypothesis indicator variables $H^m = (H_1, \ldots, H_m)$, where $H_i = 0$ (or $= 1$) if the ith null hypothesis is true (or false). Let the p-value weights be random variables $W^m = (W_1, \ldots, W_m)$ that are conditionally independent of P^m given H^m. See Figure 1.

We will assume that the p-values are drawn independently from the following mixture model:

$$
\begin{aligned}
H_1, \ldots, H_m &\overset{\text{iid}}{\longleftarrow} \text{Bernoulli}(a) && (2)\\
\xi_1, \ldots, \xi_m &\overset{\text{iid}}{\longleftarrow} \mathcal{L} && (3)\\
P_i \mid H_i = 0, \xi_i &\longleftarrow \text{Uniform}(0,1) && (4)\\
P_i \mid H_i = 1, \xi_i &\longleftarrow \xi_i, && (5)
\end{aligned}
$$

where $0 < a < 1$ and where $\mathcal{L}$ is a probability distribution on a class of p-value probability CDFs. Let $F = \int \xi d\mathcal{L}(\xi)$ be the marginal alternative, which we assume stochastically smaller than the Uniform. Marginally, the p-values are drawn IID from the CDF $G = (1-a)U + aF$, where U is the CDF of a Uniform$(0,1)$.

We consider two models for generating W^m. The first is essentially general; the second is a special case but makes analysis easier and more concrete.

General Weighting. We assume that

$$
\begin{aligned}
W_i \mid H_i = 0 &\longleftarrow Q_0\\
W_i \mid H_i = 1 &\longleftarrow Q_1,
\end{aligned}
$$

for probability distributions Q_0, Q_1 on $(0, \infty)$. The marginal distirbution of W is then $Q =$

$(1-a)Q_0 + aQ_1$. For $j = 0, 1$, let $\mu_j = \mathsf{E}\,(W \mid H = j)$, the means of Q_0 and Q_1 respectively, and let $\mu = (1-a)\mu_0 + a\mu_1$ be the marginal mean..

Under this model, define $D(t) = \mathbb{P}\{P/W \le t\}$. Then, we have

$$\begin{aligned}
D(t) &= \int \mathbb{P}\Big\{ \frac{P}{W} \le t \;\Big|\; W = w \Big\}\, dQ(w) \\
&= \int \sum_{h=0}^{1} \mathbb{P}\Big\{ \frac{P}{W} \le t \;\Big|\; W = w, H = h \Big\}\, f(h|w)\, dQ(w) \\
&= \int \sum_{h=0}^{1} \mathbb{P}\Big\{ P \le wt \;\Big|\; H = h \Big\}\, f(h|w) dQ(w) \\
&= \int \sum_{h=0}^{1} \Big((1-h)tw + hF(tw) \Big) f(h|w) dQ(w) \\
&= \sum_{h=0}^{1} \int \Big((1-h)tw + hF(tw) \Big) dQ(w|h) f(h) \\
&= (1-a) \int tw dQ(w|h=0) f(h) + a \int F(tw) dQ(w|h=1) \\
&= (1-a)\mu_0 t \,+\, a \int F(wt) dQ_1(w). \qquad (6)
\end{aligned}$$

Binary Weighting. In this case, the weights $W_1, \ldots, W_m$ can take on two possible values $w_0 \le 1 \le w_1$ and satisfy $(1/m)\sum W_i \approx 1$. The two values w_0 and w_1 correspond to guesses that the null or alternative is true. (This can easily be extended to any fixed finite number of weight values.) This guessing scheme has three parameters: γ, which determines the prevalence of alternative guesses, η, which determines the informativeness of guessing, and r, which determines the strength of weighting.

Let $U^m = (U_1, \ldots, U_m)$ be Bernoulli random variables representing prior guesses for each of the hypotheses, with $U = 1$ corresponding to an alternative and $U = 0$ to a null. Let $\overline{U}_m = \frac{1}{m}\sum_{i=1}^{m} U_i$. Assume that U^m and P^m are conditionally independent given H^m.

We assume that each U_i is drawn marginally from a Bernoulli(γ) with

$$\begin{aligned}
\mathbb{P}\{U_i = 1 \mid H_i = 1\} &= \frac{\gamma\eta}{a\eta + 1 - a} \qquad (7) \\
\mathbb{P}\{U_i = 1 \mid H_i = 0\} &= \frac{\gamma}{a\eta + 1 - a}. \qquad (8)
\end{aligned}$$

Thus, $\eta = \mathbb{P}\{U = 1 \mid H = 1\}\,/\mathbb{P}\{U = 1 \mid H = 0\}$, a measure of the informativeness of guessing. When $\eta = 1$, U^m and H^m are independent; for $\eta > 1$, there is greater likelihood of

guessing correctly, and for $0 \le \eta < 1$, incorrectly. We will typically assume that the (P_i, U_i) pairs are independent. Note the following constraint coupling γ and η:

$$0 \le \gamma \le \min\left(a\eta + 1 - a, a + \frac{1-a}{\eta}\right). \tag{9}$$

We usually take $\gamma > 0$.

Based on the U_is, we define weights W_i as follows:

$$W_i = \frac{1 + (r-1)U_i}{1 + (r-1)\overline{U}_m}. \tag{10}$$

Each of these weights takes on one of two values: $w_0 = 1/(1 + (r-1)\overline{U}_m)$ and $w_1 = r/(1 + (r-1)\overline{U}_m)$. Note that

$$r = \frac{w_1}{w_0}$$

and that the mean weight satisifes $\overline{W}_m = 1$. When the parameter $r = 1$, we return to the unweighted case.

Because of the $\overline{U}_m$ in the denominator above, it is convenient for analysis to use weights that only approximately satisfy $\overline{W}_m = 1$. Define

$$\widetilde{W}_i = \frac{1 + (r-1)U_i}{1 + (r-1)\gamma}. \tag{11}$$

Note that

$$\frac{1}{m}\sum_{i=1}^{m} \widetilde{W}_i = 1 + \frac{1}{\sqrt{m}} \frac{\sqrt{m}(\overline{U}_m - \gamma)}{1 + (r-1)\gamma} = O_P\left(\frac{1}{\sqrt{m}}\right) \tag{12}$$

$$\frac{\widetilde{W}_i - W_i}{\widetilde{W}_i} = \frac{1}{\sqrt{m}} \frac{(r-1)\sqrt{m}(\overline{U}_m - \gamma)}{1 + (r-1)\gamma + (r-1)(\overline{U}_m - \gamma)} = O_P\left(\frac{1}{\sqrt{m}}\right), \tag{13}$$

so for now, we will ignore the difference between the two weighting schemes and write W_i for $\widetilde{W}_i$.

3 FDR Control With Prior Weighting

If we reject all hypotheses for which $P_i \le T$, for some (possibly random) threshold T, then the false discovery proportion FDP is defined to be

$$\mathrm{FDP}(T) = \frac{\text{false rejections}}{\text{rejections}} = \frac{\sum_{i=1}^{n} 1\{P_i \le T\}(1 - H_i)}{\sum_{i=1}^{n} 1\{P_i \le T\}} \tag{14}$$

where the ratio is defined to be 0 when the denominator is 0. For threshold T, FDR is defined to be $\mathbb{E}(\text{FDP(T)})$.

We now define a procedure, which we call wBH, for incorporating prior p-value weights while maintaining control of FDR. Following Holm, we weight the p-values using prior weights W_i. Define $Q_i = P_i/W_i$ where $W_i > 0$. In practice the weights adjust the threshold for rejection: rejecting when $Q_i \leq T$ is equivalent to rejecting when $P_i \leq W_i T$.

Storey (2002) and Genovese and Wasserman (2002) noted that the BH threshold can be written as

$$T_{\text{BH}} = \sup\left\{t : \ \widehat{B}(t) \leq \alpha\right\} \tag{15}$$

where

$$\widehat{B}(t) = \frac{t}{\widehat{G}_m(t)}.$$

This suggests incorporating the weights by defining:

$$T_{\text{wBH}} = \sup\left\{t : \ \widehat{R}(t) \leq \alpha\right\}. \tag{16}$$

where

$$\widehat{R}(t) = \frac{t\sum_{i=1}^m W_i}{\sum_{i=1}^m 1\{P_i \leq W_i t\}} = \frac{t\overline{W}_m}{\widehat{D}_m(t)}, \tag{17}$$

where $\overline{W}_m$ is the average of the weights and $\widehat{D}_m(t)$ is the empirical CDF of P_i/W_i.

The procedure is as follows:

1. Assign weights $W_i > 0$ to each null hypothesis such that $\frac{1}{m}\sum_i W_i = 1$. (This latter condition need only be approximately met in practice.)

2. For each $i = 1, \ldots, m$, compute $Q_i = P_i/W_i$.

3. Apply the BH procedure at level α to the Q_is.

In Section 4, we show that wBH controls FDR at the nominal level. In Sections 6 and 7, we investigate the power of the procedure.

REMARK 3.1. It is possible to replace the BH procedure in Step 3 above with an adaptive FDR-controlling procedure. We expect that this will improve power, although we do not

investigate this question fully here. Using an adaptive procedure corresponds to inserting an estimate of the proportion of true nulls in the expression for $\widehat{R}$. For example, define

$$\widehat{R}_+(t) = \frac{\widehat{(1-a)}t\overline{W}_m}{\widehat{D}_m(t)}. \tag{18}$$

The estimator in Storey (2002) is a reasonable candidate.

4 wBH Controls FDR

In this section, we show that wBH is a valid FDR-controlling procedure under the general weighting scheme. We begin with a finite-sample result and then describe the asymptotic behavior of the wBH threshold.

THEOREM 4.1. *wBH controls FDR at the level $\alpha(1-a)\mu_0$ and conditionally on H^m, at the level $\alpha\mu_0\frac{1}{m}\sum_i(1-H_i)$. If $\mu_0 \le 1/(1-a)$, which occurs for instance when $\mu \le 1$, then wBH ensures* $\mathrm{FDR} \le \alpha$.

PROOF. Our approach is based on the method of Benjamini and Yekutieli (2001). First, let $Q_i = P_i/W_i$ with sorted values $Q_{(i)}$ as usual. Note that the wBH threshold for the Q_is

$$T = \sup\left\{t : \frac{t\overline{W}_m}{\widehat{D}_m(t)} \le \alpha\right\}$$

can be equivalently written as

$$T = \sup\left\{Q_{(i)} : Q_{(i)} \le \frac{\alpha i}{\sum W_j}\right\}.$$

Let $q_k = \alpha k/\sum_j W_j$. If we require $\sum_j W_j = m$, then $q_k = \alpha k/m$.

Define the events

$$R_{k,i} = \left\{\sum_{j\neq i} 1\{Q_j \le q_k\} = k-1\right\}. \tag{19}$$

For each i, the events $R_{k,i}$ for $k = 1,\dots,m$ form a partition: they are disjoint and $\cup_{k=1}^m R_{k,i}$ must occur. To see the last point, note that for each realization $k \mapsto 1 + \sum_{j\neq i} 1\{Q_j \le q_k\}$ is a non-decreasing function from $\{1,\dots,m\}$ into $\{1,\dots,m\}$ and thus has a fixed point.

Then,

$$
\begin{aligned}
&\mathsf{E}\left(\mathrm{FDP}(T) \mid H^m\right) \\
&= \sum_{i:\, H_i=0} \sum_{k=1}^{m} \frac{1}{k} \mathbb{P}\left(\{P_i \le W_i q_k\} \cap R_{k,i} \;\middle|\; H_i = 0, H^m \right) && (20) \\
&= \sum_{i:\, H_i=0} \sum_{k=1}^{m} \frac{1}{k} \mathsf{E}\left(\mathbb{P}\left(\{P_i \le W_i q_k\} \cap R_{k,i} \;\middle|\; H_i = 0, H^m, W^m \right) \;\middle|\; H_i = 0, H^m \right) && (21) \\
&= \sum_{i:\, H_i=0} \mathsf{E}\left(\sum_{k=1}^{m} \frac{1}{k} \mathbb{P}\{P_i \le W_i q_k \mid H_i = 0, H^m, W^m\}\; \mathbb{P}(R_{k,i} \mid H^m, W^m) \;\middle|\; H^m, H_i = 0 \right) && (22) \\
&= \sum_{i:\, H_i=0} \mathsf{E}\left(\sum_{k=1}^{m} \frac{\alpha W_i}{\sum_j W_j}\, \mathbb{P}(R_{k,i} \mid H^m, W^m) \;\middle|\; H^m, H_i = 0 \right) && (23) \\
&= \sum_{i:\, H_i=0} \mathsf{E}\left(\frac{\alpha W_i}{\sum_j W_j} \sum_{k=1}^{m} \mathbb{P}(R_{k,i} \mid H^m, W^m) \;\middle|\; H^m, H_i = 0 \right) && (24) \\
&= \sum_{i:\, H_i=0} \mathsf{E}\left(\frac{\alpha W_i}{\sum_j W_j} \;\middle|\; H_i = 0, H^m \right). && (25)
\end{aligned}
$$

Equation (22) follows from the (conditional) independence of P_i and P_{-i}. In the case of discrete test statistics, equality in (23) need not hold, but it can be replaced by a $\le$ as mentioned by Benjamini and Yekutieli (2001).

Because $\sum_j W_j \equiv m$, it follows that

$$
\mathsf{E}\left(\mathrm{FDP}(T) \mid H^m\right) = \alpha \mathsf{E}\left(W_1 \mid H_1 = 0\right) \frac{1}{m} \sum_i (1 - H_i) = \alpha \mu_0 \frac{1}{m} \sum_i (1 - H_i). \qquad (26)
$$

This proves the second claim. Taking expectations under the mixture model produces $\alpha\mu_0(1-a)$ on the right hand side. If $\mu = 1$, then $\mu_0 \le 1/(1-a)$, so this bound is $\le \alpha$. $\square$

REMARK 4.1. In general, if $\mathsf{E} \sum_j W_j = m$, we have

$$
\mathsf{E}\left(\mathrm{FDP}(T) \mid H^m\right) = \alpha \mathsf{E}\left(\frac{W_1}{\overline{W}_m} \;\middle|\; H_1 = 0, H^m \right) \frac{1}{m} \sum_i (1 - H_i), \qquad (27)
$$

so $\mathrm{FDR} = \alpha\mu_0(1-a) + O(m^{-1/2})$.

Now define

$$
C(t) \;=\; \frac{D(t)}{t\mu}, \qquad (28)
$$

$$\widehat{C}_m(t) = \frac{\widehat{D}_m(t)}{t \times \overline{W}_m}, \tag{29}$$

and corresponding threshold

$$t_* = \sup\left\{t : C(t) \geq \frac{1}{\alpha}\right\}, \tag{30}$$

$$T_m = \sup\left\{t : \widehat{C}_m(t) \geq \frac{1}{\alpha}\right\}. \tag{31}$$

(Note that T_m is an equivalent expression for T_{wBH}.) Recall also that G and D are, respectively, the marginal CDFs of P_i and P_i/W_i, and that F is the marginal alternative distribution. We then have the following.

LEMMA 4.1. *If F is strictly concave on $[0,1]$, then (i) G is strictly concave on $[0,1]$, (ii) D is strictly concave on $[0,1]$, and (ii) C is monotone decreasing on $(0,1)$.*

PROOF. Because $G = (1-a)U + aF$, the first claim follows immediately. Similarly, by equation (6), $D(t) = (1-a)\mu_0 t + a\int F(wt)dQ_1(w)$. Hence, for $0 \leq \lambda \leq 1$,

$$\begin{aligned} D((1-\lambda)t_0 + \lambda t_1) &= (1-a)\mu_0((1-\lambda)t_0 + \lambda t_1) + a\int F(w((1-\lambda)t_0 + \lambda t_1))dQ_1(w) \\ &> (1-\lambda)D(t_0) + \lambda D(t_1), \end{aligned} \tag{32}$$

where the final inequality is strict because Q_1 has all its mass on $(0,\infty)$. This proves the second claim.

Finally, let $1 > t_1 > t_0 > 0$ and note that $F(0) = 0$ implies $D(0) = 0$. Note that

$$C(t_1) = \frac{D(t_1)}{t_1} = \frac{(1-\frac{t_0}{t_1})D(0) + \frac{t_0}{t_1}D(t_1)}{t_0} \leq \frac{D(t_0)}{t_0} = C(t_0), \tag{33}$$

which proves (iii). □

THEOREM 4.2. *Assume that F is strictly concave. Then,*

$$T_m \to t_* \quad \text{almost surely} \tag{34}$$

$$\mathsf{E}\,|\text{FDP}(T_m) - \text{FDP}(t_*)| \to 0 \tag{35}$$

and thus

$$\mathsf{E}\,(\text{FDP}(T_m)) \leq \alpha + o(1). \tag{36}$$

PROOF. First assume that the W_is are not constrained to have average 1, and are thus independent. Fix $b, \epsilon > 0$. Then, $C(t_* + b) < 1/\alpha$, so $C(t_* + b) = 1/\alpha - \delta$ for some $\delta > 0$. We have, using Lemma 4.1, that for every $t > t_* + b$,

$$\begin{aligned}
\widehat{C}_m(t) &= \frac{\widehat{D}_m(t)}{t \times \overline{W}_m} \\
&\leq \frac{D(t) + \sup_u |\widehat{D}_m(u) - D(u)|}{t\mu - t|\overline{W}_m - \mu|} \\
&\leq \frac{D(t) + \epsilon}{t(\mu - \epsilon)} \qquad (37) \\
&= C(t) \left(\frac{\mu}{\mu - \epsilon} \right) + \frac{\epsilon}{t(\mu - \epsilon)} \\
&\leq C(t_* + b) \left(\frac{\mu}{\mu - \epsilon} \right) + \frac{\epsilon}{(t_* + b)(\mu - \epsilon)} \\
&= \left(\frac{1}{\alpha} - \delta \right) \left(\frac{\mu}{\mu - \epsilon} \right) + \frac{\epsilon}{(t_* + b)(\mu - \epsilon)} < \frac{1}{\alpha}, \qquad (38)
\end{aligned}$$

for large enough m. Equation (37) follows for large m from the Gilvenko-Cantelli Theorem (which implies $\sup_u |\widehat{D}_m(u) - D(u)| \to 0$ a.s.) and the Strong Law of Large Numbers (which implies $\overline{W}_m \to \mu$ a.s.). Hence, $T_m < t_* + b$. Combined with a similar argument applied to $t < t_* - b$, this implies that

$$|T_m - t_*| \leq b \quad \text{almost surely}$$

for all large m.

Now,

$$\text{FDP}(T_m) = \frac{m^{-1} \sum_i (1 - H_i) 1\{P_i/W_i \leq t\}}{m^{-1} \sum_i 1\{P_i/W_i \leq t\}} = \frac{\widehat{V}_m(T_m)}{\widehat{D}_m(T_m)}. \qquad (39)$$

Then, for all large m,

$$\begin{aligned}
|\widehat{D}_m(T_m) - D(t_*)| &= |\widehat{D}_m(T_m) - D(T_m) + D(T_m) - D(t_*)| \\
&\leq \sup_u |\widehat{D}_m(u) - D(u)| + |D(T_m) - D(t_*)| \\
&\to 0
\end{aligned}$$

by the Gilvenko-Cantelli and Continuous Mapping Theorems. By similar argument, $\widehat{V}_m(T_m) - V(t_*) = o(1)$ almost surely, where

$$V(t) = \mathsf{E}\,(1 - H_i) 1\{P_i/W_i \leq t\} = (1 - a) t \mu_0 \leq t\mu.$$

If the W_is are constrained to have average 1, we proceed as follows. Write $W_i = U_i/\overline{U}_m$ for IID variables $U_1, \ldots, U_m$. Let $\widetilde{W}_i = U_i/\mathsf{E}\, U_1$. It follows that

$$\begin{aligned}
\widehat{D}_m(t) &= \frac{1}{m}\sum_i 1\Big\{\frac{P_i}{W_i} \le t\Big\} \\
&= \frac{1}{m}\sum_i 1\Big\{\frac{P_i}{\widetilde{W}_i} \le t\frac{\mathsf{E}\, U_1}{\mathsf{E}\, U_1 + (\overline{U}_m - \mathsf{E}\, U_1)}\Big\} \\
&\le \frac{1}{m}\sum_i 1\Big\{\frac{P_i}{\widetilde{W}_i} \le t\frac{\mathsf{E}\, U_1}{\mathsf{E}\, U_1 - \epsilon}\Big\} \\
&\le \mathbb{P}\Big\{\frac{P_i}{\widetilde{W}_i} \le t\frac{\mathsf{E}\, U_1}{\mathsf{E}\, U_1 - \epsilon}\Big\} + \epsilon,
\end{aligned}$$

for large enough m, uniformly in t. Similarly,

$$\begin{aligned}
\widehat{D}_m(t) &\ge \frac{1}{m}\sum_i 1\Big\{\frac{P_i}{\widetilde{W}_i} \le t\frac{\mathsf{E}\, U_1}{\mathsf{E}\, U_1 + \epsilon}\Big\} \\
&\ge \mathbb{P}\Big\{\frac{P_i}{\widetilde{W}_i} \le t\frac{\mathsf{E}\, U_1}{\mathsf{E}\, U_1 + \epsilon}\Big\} - \epsilon,
\end{aligned}$$

Because $\epsilon > 0$ is arbitrary, we conclude that $\sup_u |\widehat{D}_m(t) - \widehat{\widetilde{D}}_m(t)| \to 0$ almost surely, where $\widetilde{D}(t) = \mathbb{P}\Big\{P_i/\widetilde{W}_i \le t\Big\}$ and $\widehat{\widetilde{D}}_m$ is the corresponding empirical CDF. The remainder of the proof is the same.

Thus, $|\mathrm{FDP}(T_m) - \mathrm{FDP}(t_*)| \to 0$ almost surely, and because this is bounded, dominated convergence yields the second claim. The third claim follows immediately. □

5 Weighted Exceedance Control

In this section, we present an approach to weighted testing that controls False Discovery Exceedance (FDX). The method generalizes the approaches in Genovese and Wasserman (2004a, 2004b), van der Laan, Dudoit and Pollard (2004), and Perone Pacifico, Genovese, Verdinelli, and Wasserman (2004a, 2004b). Those methods begin with a familywise test and then augment the familywise rejection region by adding in extra rejections.

First, we introduce some notation that is helpful for this section. Let $S = \{1, \ldots, m\}$ and let

$$S_0 \equiv S_0(\mathbb{P}) = \{j : H_j = 0\} \tag{40}$$

be the set of true nulls. We call any (possibly random) $\mathcal{R} \subset S$ a *rejection region* and say that $\mathcal{R}$ controls familywise error rate at level α if

$$\mathbb{P}\{\#(\mathcal{R} \cap S_0(\mathbb{P})) > 0\} \leq \alpha,$$

where $\#(B)$ denotes the number of points in a set B. The FDP of a rejection set $\mathcal{R}$ is then

$$\text{FDP} = \frac{\sum_{j=1}^m (1 - H_j) 1\{j \in \mathcal{R}\}}{\sum_{j=1}^m 1\{j \in \mathcal{R}\}} \tag{41}$$

where the ratio is defined to be zero if the denominator is zero.

Instead of controlling the mean of the FDR, we will instead control the FDP exceedance. Specifically, our goal in this section is to use the weighted p-values to find a rejection set $\mathcal{R}$ such that

$$\text{FDX} \equiv \mathbb{P}\{\text{FDP} > c\} \leq \alpha \tag{42}$$

for given c and α. We call such an $\mathcal{R}$ a (c, α) *rejection region* and we say that $\mathcal{R}$ provides (c, α) exceedance control. The inequality (42) will be valid for all finite m and will not make assumptions about the form of the dependence between the p-values.

We begin by introducing weighted familywise tests. Then we use these famliywise tests to construct exceedance control methods.

Let us recall two popular methods for familywise control. Let

$$P_{(1)} \leq \cdots \leq P_{(m)}$$

denote the sorted p-values. The Bonferroni method uses

$$\mathcal{R}_0 = \left\{ j : \ P_j \leq \frac{\alpha}{m} \right\}.$$

Holm's (1979) method takes

$$\mathcal{R}_0 = \left\{ j : \ P_j \leq T \right\}$$

where $T = 0$ if $P_{(1)} \geq \alpha/m$ and $T = P_{(k)}$ otherwise, where $k = \max\{j : \ P_{(j)} < \alpha/(m-j)\}$.

In what follows, we assume that $\mathbb{P}\{\overline{W}_m = 1\} = 1$, for simplicity. This means that the W_is are not independent but we do continue to assume that they are (marginally) identically distributed. The weighted Bonferroni rejection set is

$$\mathcal{R}_0 = \left\{ j : \ Q_j \leq \frac{\alpha}{m} \right\}. \tag{43}$$

LEMMA 5.1. *The weighted Bonferroni procedure controls familywise error at level* $(1-a)\mu_0\alpha$*, which is* $\leq \alpha$ *if* $\mu = 1$*, as assumed above.*

PROOF.

$$\begin{aligned}
\mathbb{P}\{\#(\mathcal{R} \cap S_0(\mathbb{P})) > 0\} &= \mathbb{P}\left\{ P_j \leq \frac{\alpha W_j}{m} \text{ for some } j \in S_0 \right\} \\
&\leq \sum_{j=1}^{m} \mathbb{P}\left\{ P_j \leq \frac{\alpha W_j}{m} \text{ and } H_j = 0 \right\} \\
&= \sum_{j=1}^{m} \mathbb{P}\left\{ P_j \leq \frac{\alpha W_j}{m} \;\middle|\; H_j = 0 \right\} \mathbb{P}\{H_j = 0\} \\
&= (1-a) \sum_{j=1}^{m} \int \mathbb{P}\left\{ P_j \leq \frac{\alpha w}{m} \;\middle|\; H_j = 0, W^m = w^m \right\} dQ_0(w) \\
&= (1-a)\frac{\alpha}{m} \sum_{j=1}^{m} \int w dQ_0(w) \\
&= (1-a)\alpha\mu_0 \leq \alpha.
\end{aligned}$$

□

Holm's (1979) weighted procedure for controlling familywise error is as follows. Let $Q_i = P_i/W_i$ and let

$$Q_{(1)} \leq \cdots \leq Q_{(m)}$$

denote the sorted values. Let

$$H_{(1)}, \ldots, H_{(m)}, \quad \text{and} \quad W_{(1)}, \ldots, W_{(m)}$$

denote the H_is and W_i's sorted correspondingly. If $Q_{(1)} \geq \alpha/m$, set $\mathcal{R}_0 = \emptyset$. Otherwise, find the largest j for which

$$Q_{(j)} < \frac{\alpha}{\sum_{i=j}^{m} W_{(j)}}$$

and let $\mathcal{R}_0$ be the indices corresponding to the j smallest $Q_j's$. Holm proved that

$$\mathbb{P}\{\#(\mathcal{R}_0 \cap S_0(\mathbb{P})) > 0\} \leq \alpha$$

when the weights and H_is are regarded as fixed. Let us now prove that the same is true for random weights and random H_is by adapting his proof.

LEMMA 5.2. *The weighted Holm procedure controls familywise error at level α.*

PROOF. Let $H^m = (H_1, \ldots, H_m)$, $W^m = (W_1, \ldots, W_m)$, $N_0 = \sum_{i=1}^m W_i(1 - H_i)$ and let $\pi(h)$ denote the marginal probability mass function for the vector H^m. Also, we write $S_0 = S_0(H^m)$ to make explicit the dependence of S_0 on H^m. Define the event

$$A = \left\{ Q_i > \frac{\alpha}{N_0} \text{ for all } i \in S_0 \right\}.$$

Then,

$$\begin{aligned}
&\mathbb{P}(A) \\
&= 1 - \mathbb{P}\left\{ Q_i \leq \frac{\alpha}{N_0} \text{ for some } i \in S_0 \right\} \\
&= 1 - \sum_h \int \mathbb{P}\left\{ Q_i \leq \frac{\alpha}{N_0} \text{ for some } i \in S_0 \;\middle|\; W^m = w, H^m = h \right\} dQ_0(w)\pi(h) \\
&\geq 1 - \sum_h \int \sum_{i \in S_0(h)} \mathbb{P}\left\{ Q_i \leq \frac{\alpha}{N_0} \;\middle|\; W^m = w, H^m = h \right\} dQ_0(w)\pi(h) \\
&= 1 - \sum_h \int \sum_{i \in S_0(h)} \mathbb{P}\left\{ P_i \leq \frac{w_i \alpha}{N_0} \;\middle|\; W^m = w, H^m = h \right\} dQ_0(w)\pi(h) \\
&= 1 - \sum_h \int \sum_{i \in S_0(h)} \frac{w_i \alpha}{N_0} dQ_0(w)\pi(h) \\
&= 1 - \alpha \sum_h \int \frac{1}{N_0} \sum_{i \in S_0(h)} w_i dQ_0(w)\pi(h) \\
&= 1 - \alpha \sum_h \int \frac{N_0}{N_0} dQ_0(w)\pi(h) \\
&= 1 - \alpha \sum_h \int dQ_0(w)\pi(h) \\
&= 1 - \alpha.
\end{aligned}$$

Assume A occurs. Let $\nu = \min\{j : \; H_{(j)} = 0\}$. Then,

$$Q_{(\nu)} > \frac{\alpha}{N_0} = \frac{\alpha}{\sum_{i \in S_0} W_i} \geq \frac{\alpha}{\sum_{i=\nu}^m W_{(i)}}$$

which implies that $S_0 \cap \mathcal{R}_0 = \emptyset$. $\square$

Let $\mathcal{R}_0$ be the rejection region from either the weighted Bonferroni method or the weighted Holm method. Define $\mathcal{R}$ as follows. If $\#(\mathcal{R}_0) < (1-c)/c$ take $\mathcal{R} = \emptyset$. Otherwise take $\mathcal{R} = \mathcal{R}_0 \cup A$ where $A \subset S$ is any set of hypotheses such that $A \cap \mathcal{R}_0 = \emptyset$ and $\#(A)/(\#(A)+\#(\mathcal{R}_0)) \leq c$.

THEOREM 5.1. *If $\mathcal{R}$ is constructed as defined above then*

$$\mathbb{P}\{\text{FDP} > c\} \leq \alpha. \tag{44}$$

The proof is essentially the same as the proofs for the unweighted case in Genovese and Wasserman (2004a, 2004b) or van der Laan, Dudoit and Pollard (2004).

There is freedom in choosing the extra rejections A. Two alternatives are to choose the k hypotheses not in $\mathcal{R}_0$ with the smallest Q-values or the smallest P-values, where $k \approx \#(\mathcal{R}_0)c/(1-c)$. The former will have somewhat higher power when the weights are well chosen and the latter will be more robust to mis-specification of the weights. Based on the methods in Genovese and Wasserman (2004b) it is possible to construct versions with even higher power but we shall not pursue them here.

6 Power of the Weighted Procedures

Having established that our procedures control FDR or FDX, we next turn to the question of what effect weighting has on power. To make weighting worthwhile, power should improve substantially when guessing is informative but not drop too low when guessing is poor. The asymmetry between null and alternative makes this possible. With F stochastically smaller than the Uniform, assigning small weights to true alternatives can still allow the corresponding null hypotheses to be preferentially rejected, and similarly for large weights assigned to true nulls. This "power arbitrage" does in fact appear to hold practice. Indeed, if weighting is "informative," in that the weights are positively associated with the null hypothesis being false, we would expect weighting to improve power over the corresponding unweighted procedure. In this section, we provide some theoretical support for this idea.

Let Q_0 and Q_1 be weight distributions as in Section 2, with respective means μ_0 and μ_1. We assume that $\mu = (1-a)\mu_0 + a\mu_1 \equiv 1$, so $0 < \mu_0 < 1/(1-a)$ and $0 < \mu_1 < 1/a$. We call $\mu_1 > 1$ the informative case and $\mu_1 < 1$ the mis-informative case.

Define the type I error rate and power as a function of threshold:

$$I(t) = \mathbb{P}\{P \leq Wt \mid H = 0\} = \mu_0 t \tag{45}$$

$$H(t) = \mathbb{P}\{P \leq Wt \mid H = 1\} = \int F(wt)\, dQ_1(w). \tag{46}$$

Note that under informative weighting, the Type I error rate drops.

Now, let t^w and t^0 denote the asymptotic thresholds (defined from the population CDFs) for the weighted and unweighted method. Then, we have that the ratio of power is

$$\frac{H(t^w)}{F(t^0)} = \frac{t^w}{t_0}\left[1 + (\mu_1 - 1)\frac{a\alpha}{1-(1-a)\alpha}\right]. \tag{47}$$

To see this, note that at the specified threshold

$$\frac{t^w}{D(t^w)} = \alpha = \frac{t^0}{G(t^0)}. \tag{48}$$

Solving for $H(t^w)$ on the left side yields $H(t^w)/t^w = 1/a\alpha - \mu_0(1-a)/a = 1/a\alpha + \mu_1 - 1/a$, by the relationship between μ_0 and μ_1. A similar calculation for $F(t^0)$ shows that $F(t^0)/t^0 = 1/a\alpha - (1-a)/a$.

Equation (47) is less than satisfying on its own because the thresholds t^0 and t^w depend on F and H as well. To investigate how the threshold changes with the weighting, we introduce a one parameter family. Fix Q_1 and for $0 \leq \lambda \leq 1$, let $W^\lambda = \lambda W + 1 - \lambda$. Then $\mathsf{E}\, W^\lambda = 1$ and for $j = 0, 1$

$$\mu_j^\lambda = \mathsf{E}\,(W^\lambda \mid H = j) = 1 + \lambda(\mu_j - 1). \tag{49}$$

Let t^λ be the asymptotic threshold defined by $t^\lambda / D^\lambda(t^\lambda) = \alpha$, where t^0 is the BH threshold and $D^0 = G$. Exploiting this equality as λ varies, we can define

$$R(t, \lambda) = \int \frac{F((\lambda W + 1 - \lambda)t)}{t}\, dQ_1(w) - \lambda(\mu_1 - 1) - \frac{F(t^0)}{t^0}. \tag{50}$$

Then, $R(t^\lambda, \lambda) \equiv 0$ for $0 \leq \lambda \leq 1$. Computing the partial derivatives of R at $(t^0, 0)$ and applying the Implicit Function Theorem yields

$$\left.\frac{dt^\lambda}{d\lambda}\right|_{t^0,0} = (\mu_1 - 1)t_0\frac{f(t_0) - t_0}{F(t_0) - t_0 f(t_0)}, \tag{51}$$

where $F(t_0) - t_0 f(t_0) > 0$ by the strict concavity of F. To first order then, we have

$$\frac{t^\lambda}{t^0} = 1 + (\mu_1 - 1)t_0\frac{f(t_0) - t_0}{F(t_0) - t_0 f(t_0)}, \tag{52}$$

and plugging this in to equation (47) gives to first order that

$$\begin{aligned}\frac{H(t^\lambda)}{F(t^0)} &= \left[1 + (\mu_1 - 1)t_0\frac{f(t_0) - t_0}{F(t_0) - t_0 f(t_0)}\right]\left[1 + (\mu_1 - 1)\frac{a\alpha}{1-(1-a)\alpha}\right] \\ &= \left[1 + (\mu_1 - 1)t_0\frac{f(t_0) - t_0}{F(t_0) - t_0 f(t_0)}\right]\left[1 + (\mu_1 - 1)t_0\frac{1}{F(t_0)}\right]\end{aligned} \tag{53}$$

In the informative case, the only term that is possibly negative here is $f(t_0)-t_0$. By concavity, this is a measure of distance between the alternative F and the uniform. For fixed a and α, t^0 is determined by the intersection of F with a fixed line of slope $\frac{1-(1-a)\alpha}{a\alpha}$. An F with mass concentrated near zero intersects that line where the density is small; an F close to the uniform intersects the line at larger density. When $f(t_0)$ is very small, there is less room for improvement in power because most of the alternatives will have been rejected at t_0.

Consider the above expression for the family of Normal$(\theta, 1)$ alternatives. In this case, for all $\theta \leq 5$ at least, $f(t_0) > t_0$, so informative weighting improves power for large m. Figure 2 gives a representative contour plot of the power ratio as a function of θ and μ_1 in this family..

This also suggests that for small α, $f(t_0) > t_0$ so informative weighting should improve power in this case as well. In particular, if $f(\alpha) > \alpha$, then because $t_0 \leq \alpha$, $f(t_0) \geq f(\alpha) > \alpha \geq t_0$, and informative weighting improves power. In the Normal$(\theta, 1)$ family, $f_\theta(t) = \exp(-\frac{1}{2}\theta^2 + \theta\Phi^{-1}(1-t))$, so informative weighting improves power *at least* for all $\theta \leq \theta_\alpha$, where

$$\theta_\alpha = \Phi^{-1}(1-\alpha) + \sqrt{(\Phi^{-1}(1-\alpha))^2 - 2\log\alpha}. \tag{54}$$

For example, $\theta_{0.05} = 4.59$ and $\theta_{0.01} = 6.15$. (As Figure 2 shows, these are conservative; in practice, the boundary θ will be higher.)

7 Simulation Studies

In this section, we present simulations and power calculations to evaluate the power of wBH under a variety of weighting regimes. We limit our attention to binary weighting schemes with $\gamma = a$ but allow for a wide range on η and r.

Figures 3–5 compare the power of wBH to the standard BH method and the BH "Oracle," which takes a as known. We consider Normal$(\theta, 1)$ alternatives for $\theta \in \{2, 2.5, 3, 3.5, 4\}$ and $a \in \{0.01, 0.05, 0.1\}$. We ran 10000 iterations for each configuration, computing results for all methods on the same data. This amounts to 360000 iterations for each of the BH results because they are not affected by η or r.

FDR was controlled at the nominal level, within simulation error, for all cases. But as

expected from Theorem 4.1, wBH ensures $FDR \leq (1-a)\mu_0\alpha$. So under informative guessing, it both improves power and reduces FDR below the nominal level. This suggests an adaptive method for gaining further power by estimating μ_0 from the data.

To investigate the power of the weighted Holm-based method for FDX control, we also conducted a simulation using the same settings as above. The results are given in Figures 6–8.

We discuss the power results fully in Section 8.

8 Discussion

Scientific inquiries that aim to test vast numbers of well-defined hypotheses using a common database have become more and more common. These studies have been plagued by low power. In response to this new scientific environment, new paradigms for multiple testing that increase power are required. Methods that incorporate additional information such as the spatial structure of the hypotheses or prior information can improve the chances of detecting small, but important effects.

In cases where there is spatial structure among the tests, one approach is to focus on inference for significant regions. Pacifico Perone, Genovese, Verdinelli, and Wasserman (2004a, 2004b) show how to control the proportion of false regions in a random field context. Taylor (2004) alters the null hypothesis to account for adjacency and devises a procedure that is better able to distinguish structure signal from scattered noise.

We present a new multiple testing approach wBH that allows one to incorporate prior information in the form of weights to increase the chance of discovering the non-null hypotheses. In our analytical investigation we show that wBH controls FDR at, or below, the nominal level. Moreover, we obtain an expression that readily permits power comparisons under various conditions.

Our simulation results (Figures 3–8) confirm that, while controlling FDR or FDX, weighting can lead to substantial gains in power when the weights are well chosen ($\mu_1, \eta > 1$) whereas the potential loss in power is small even when the weights are poorly chosen ($\mu_1, \eta < 1$). Remarkably, the loss of power is not serious even if the weights are completely

wrong ($\eta = 0$).

Much research has been done on estimating a to construct adaptive FDR-controlling procedures with higher power than BH. The gain in power from such adaptive procedures, however, is bounded above by the difference in heights of the two horizontal lines in Figures 3–5. Notice that the potential gain in power from informative weighting is at least as large and often substantially larger.

Comparing Figures 3–5 to Figures 6–8, we see that FDR control typically provides higher power than FDX control, as expected given the stronger guarantee of the latter, but that the weighted FDX-controlling procedures are more robust to poorly chosen weights. The power below $\eta = 1$ drops an essentially negligible amount. This suggests using a large value of r for FDX control, whereas smaller r's seem warranted for FDR control.

Other weighted multiple testing methods have been proposed. In contrast to Benjamini and Hochberg (1997)'s weighted approach, our procedure aims to weight hypothesis highly if they are more likely to be non-null, a priori. For a threshold T, wBH defines the weighted false discovery proportion as

$$\frac{\sum_{i=1}^{m}(P_i \leq W_i T)(1 - H_i)}{\sum_{i=1}^{m}(P_i \leq W_i T)}.$$

Benjamini and Hochberg (1997) weight hypotheses based on the relative consequences of making a false discovery on the ith hypothesis. They define the weighted false discovery proportion as

$$\frac{\sum_{i=1}^{m} W_i(P_i \leq T)(1 - H_i)}{\sum_{i=1}^{m} W_i(P_i \leq T)}.$$

With the former approach every false discovery is counted equally. The weights define varying thresholds for rejection. Heavily weighted hypotheses are rejected more easily. With the latter approach, heavily weighted hypotheses count more when assessing the false discovery rate. But all hypotheses are rejected or accepted with an equal threshold value. Clearly both of these appraoches have merit and which is preferable is dependent upon the context of the experiment.

In considering the use of prior information to improve testing, a Bayesian approach comes to mind. Indeed, the Bayesian method given in Genovese and Wasserman (2003) can easily

be extended to incorporate distinct priors for each hypothesis. Storey (2002) and Efron, Storey, Tusher and Tibshirani (2002) have given Bayesian interpretations of FDR. It is an interesting question as the relationship between a weighted version of their procedures and a fully Bayesian approach.

References

Benjamini, Y. and 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, 289–300.

Benjamini, Y. and Hochberg, Y. (1997). Multiple Hypothesis Testing with Weights, *Scandinavian Journal of Statistics*, **24**, 407–418.

Benjamini, Y., Krieger, A., and Yekutieli, D. (2004). Adaptive Linear Step-up Procedures that Control the False Discovery Rate, to appear.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29, 1165-1188.

Genovese, C. R. and Wasserman, L. (2002). Operating Characteristics and Extensions of the False Discovery Rate Procedure, *J. Royal Statist. Soc. B*, **64**, 499–518.

Genovese, C. R. and Wasserman, L. (2003). Bayesian and Frequentist Multiple Testing. To appear: *Valencia VI*.

Genovese, C. R. and Wasserman, L. (2004a). A stochastic process approach to false discovery ccontrol. *The Annals of Statistics*, **32**, 1035-1061.

Genovese, C. R. and Wasserman, L. (2004b).

Efron, B., Storey, J., Tusher, V.G., and Tibshirani R. (2001). Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association*, 96, 1151-1160.

Holm S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

Perone Pacifico, M., Genovese, C., Verdienlli, I. and Wasserman, L. (2004a). False discovery control for random fields. In press: *Journal of the American Statistical Association*,

Perone Pacifico, M., Genovese, C., Verdienlli, I. and Wasserman, L. (2004b). Scan Clustering: A False Discovery Approach. Technical Report, Dept. of Statistics, Carnegie Mellon University.

Storey, J.D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, **64**, 479–498.

Storey J.D., Taylor J.E., and Siegmund D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society, Series B*, **66**, 187–205.

Taylor, J. (2004), personal communication.

van der Laan, M.J., Dudoit, S., Pollard, K.S. (2004). Multiple Testing. Part III. Procedures for Control of the Generalized Family-Wise Error Rate and Proportion of False Positives, U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 141.

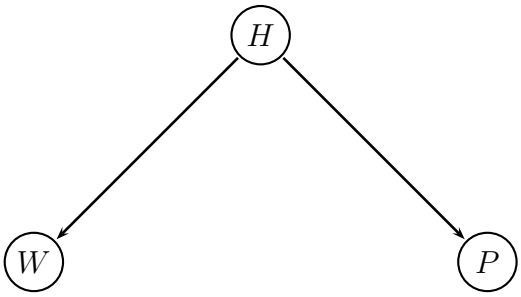

Figure 1: Graphical representation of the joint distribution of (P, W, H). Note that P and W are conditionally independent given H.

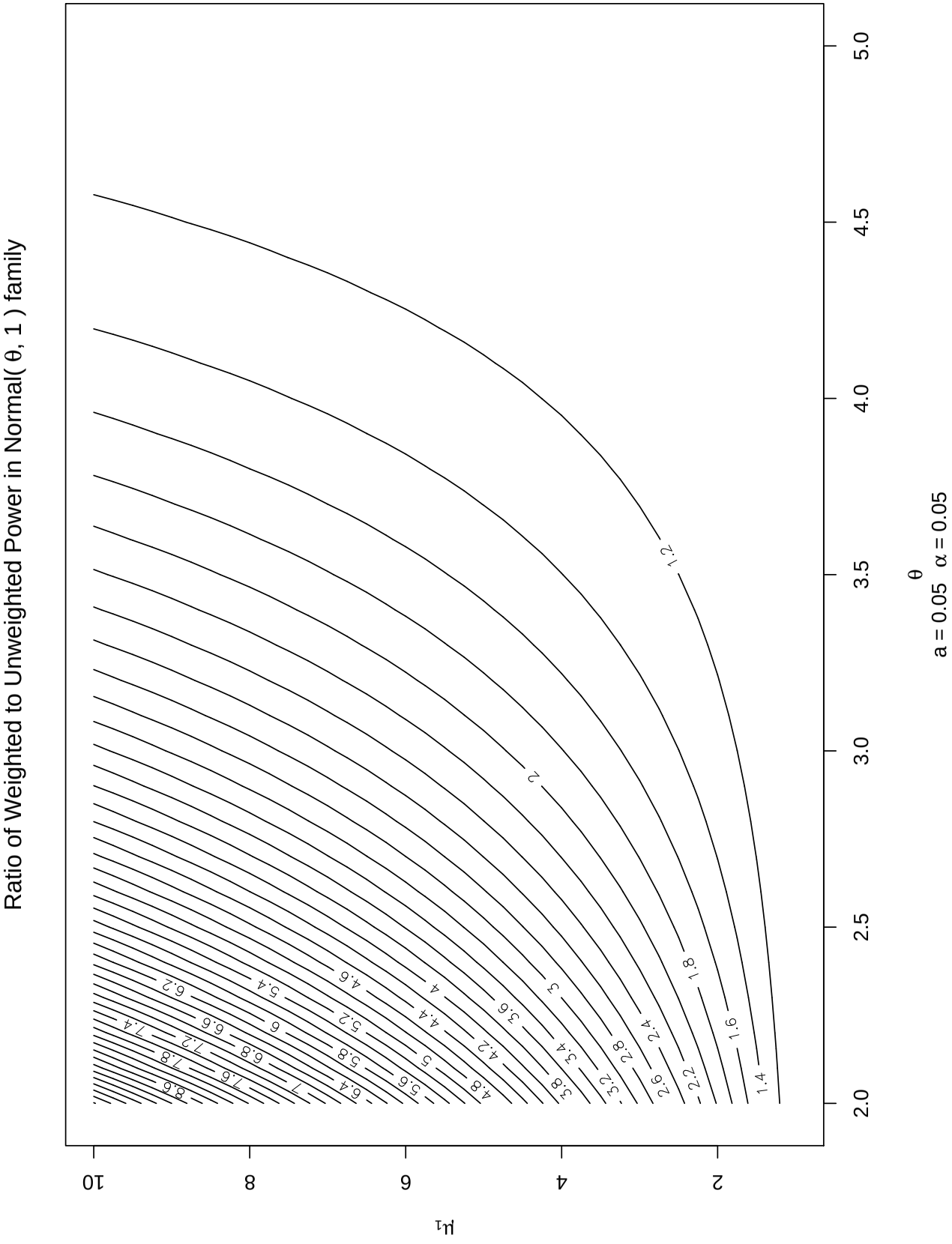

Figure 2: Ratio of weighted to unweighted power for the FDR controlling procedures from (53) as a function of μ_1 and θ, for $a = 0.05$.

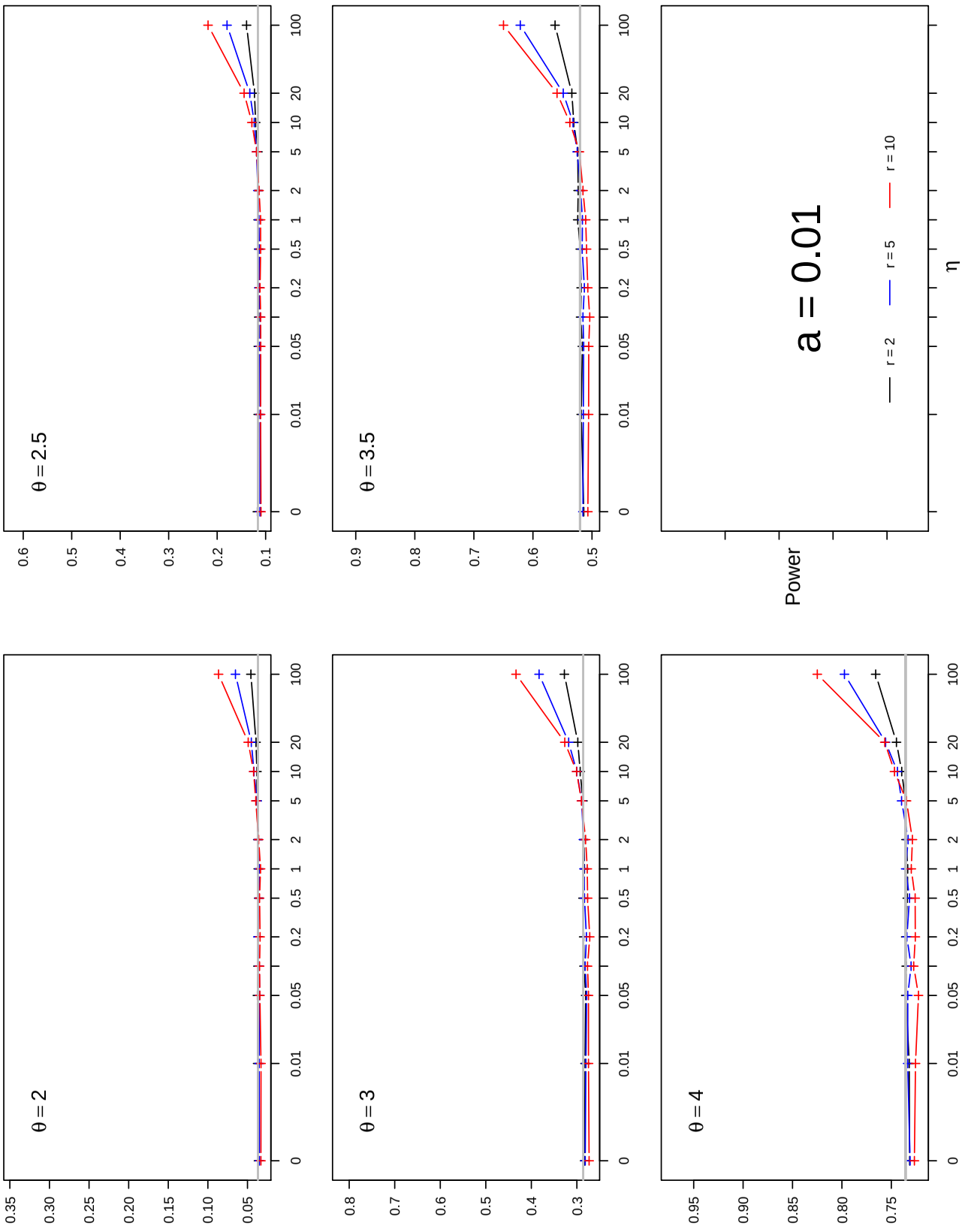

Figure 3: Power of the wBH procedure as a function of η and r. The BH and BH Oracle powers are given by the horizontal lines, with the latter always larger. Vertical scales are held fixed across θ not a.

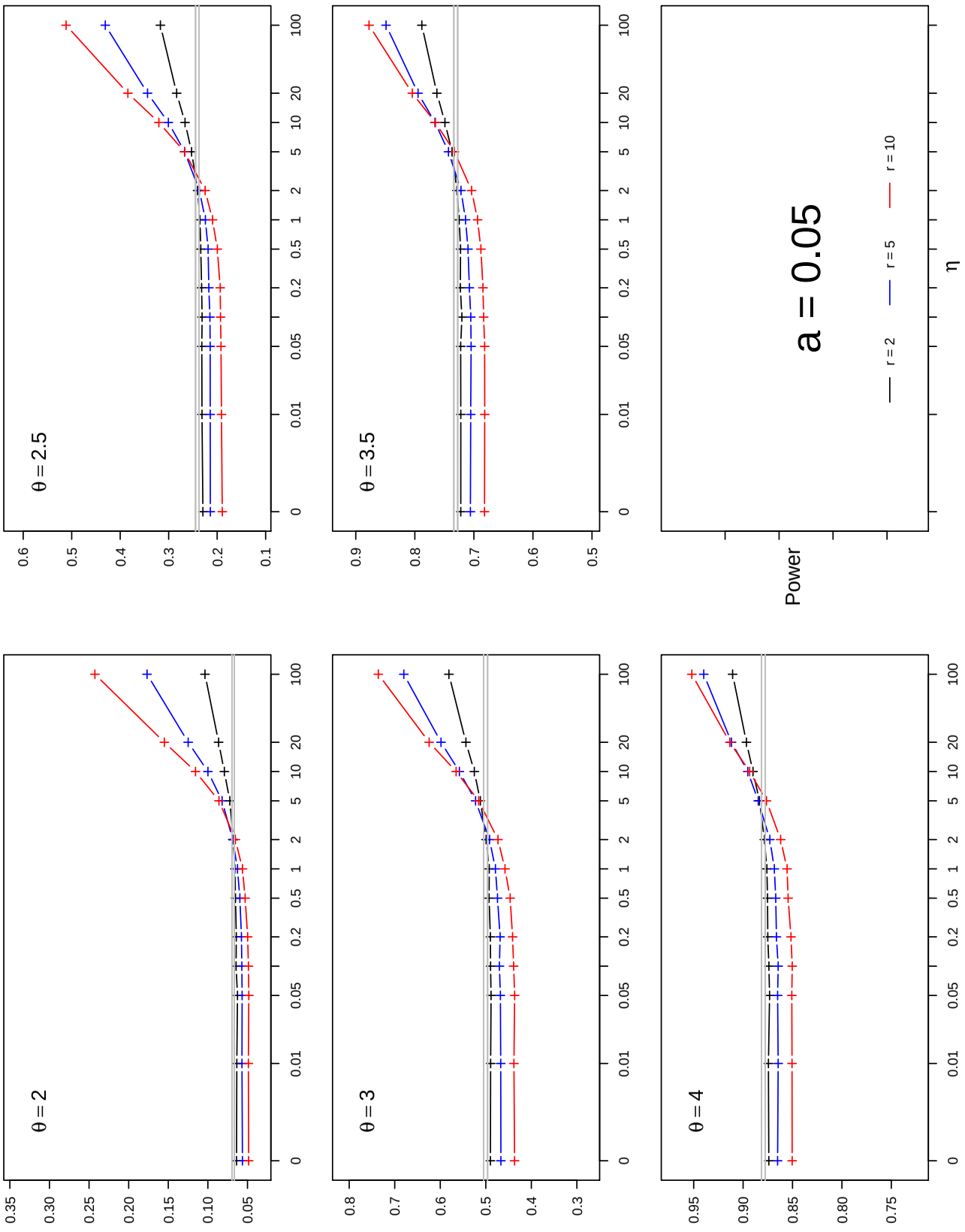

Figure 4: Power of the wBH procedure as a function of η and r. The BH and BH Oracle powers are given by the horizontal lines, with the latter always larger. Vertical scales are held fixed across θ not a.

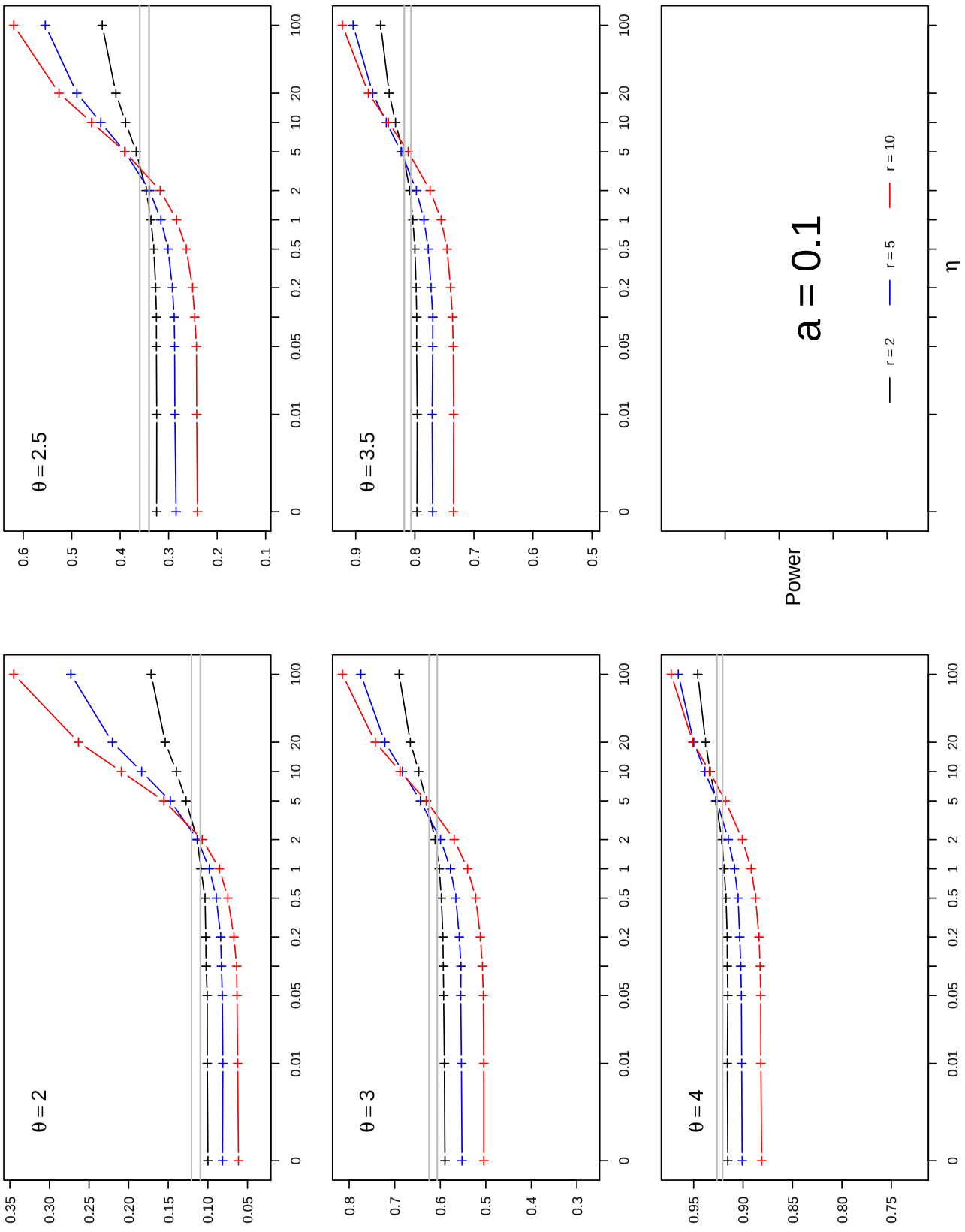

Figure 5: Power of the wBH procedure as a function of η and r. The BH and BH Oracle powers are given by the horizontal lines, with the latter always larger. Vertical scales are held fixed across θ not a.

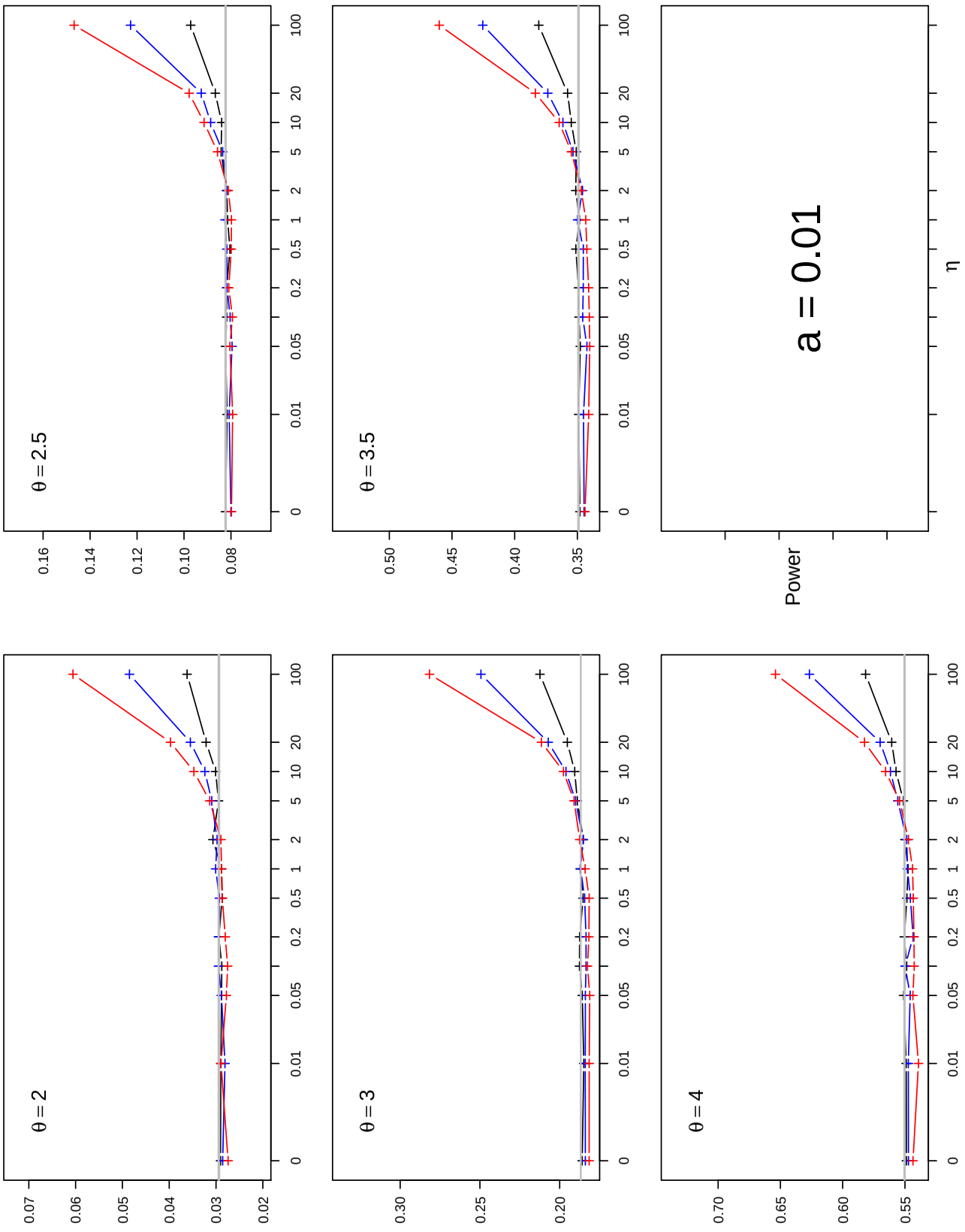

Figure 6: Power of the weighted FDX-controlling procedure as a function of η and r. The power of the weighted Holm procedure is given by the horizontal line. Vertical scales are held fixed across θ not a.

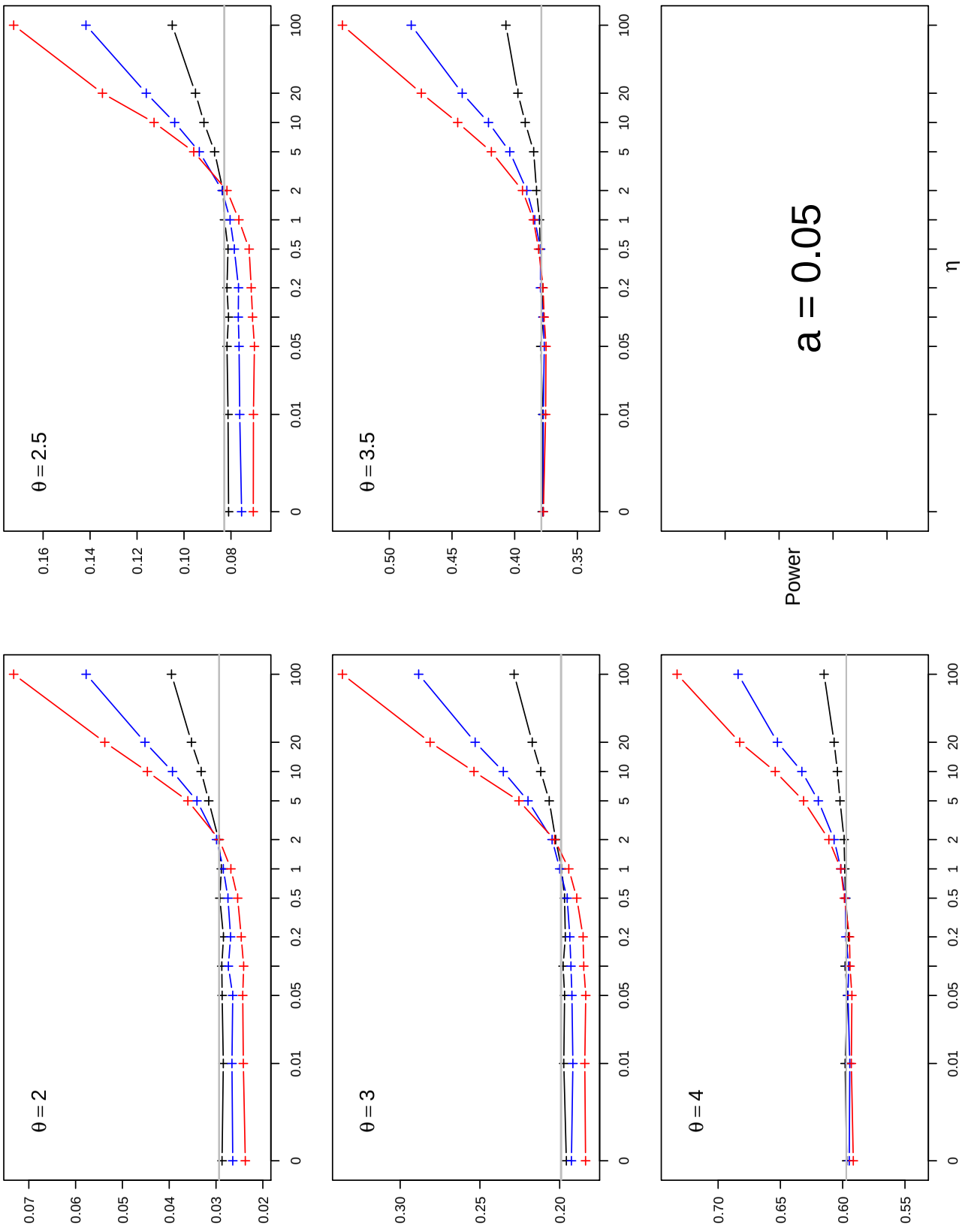

Figure 7: Power of the weighted FDX-controlling procedure as a function of η and r. The power of the weighted Holm procedure is given by the horizontal line. Vertical scales are held fixed across θ not a.

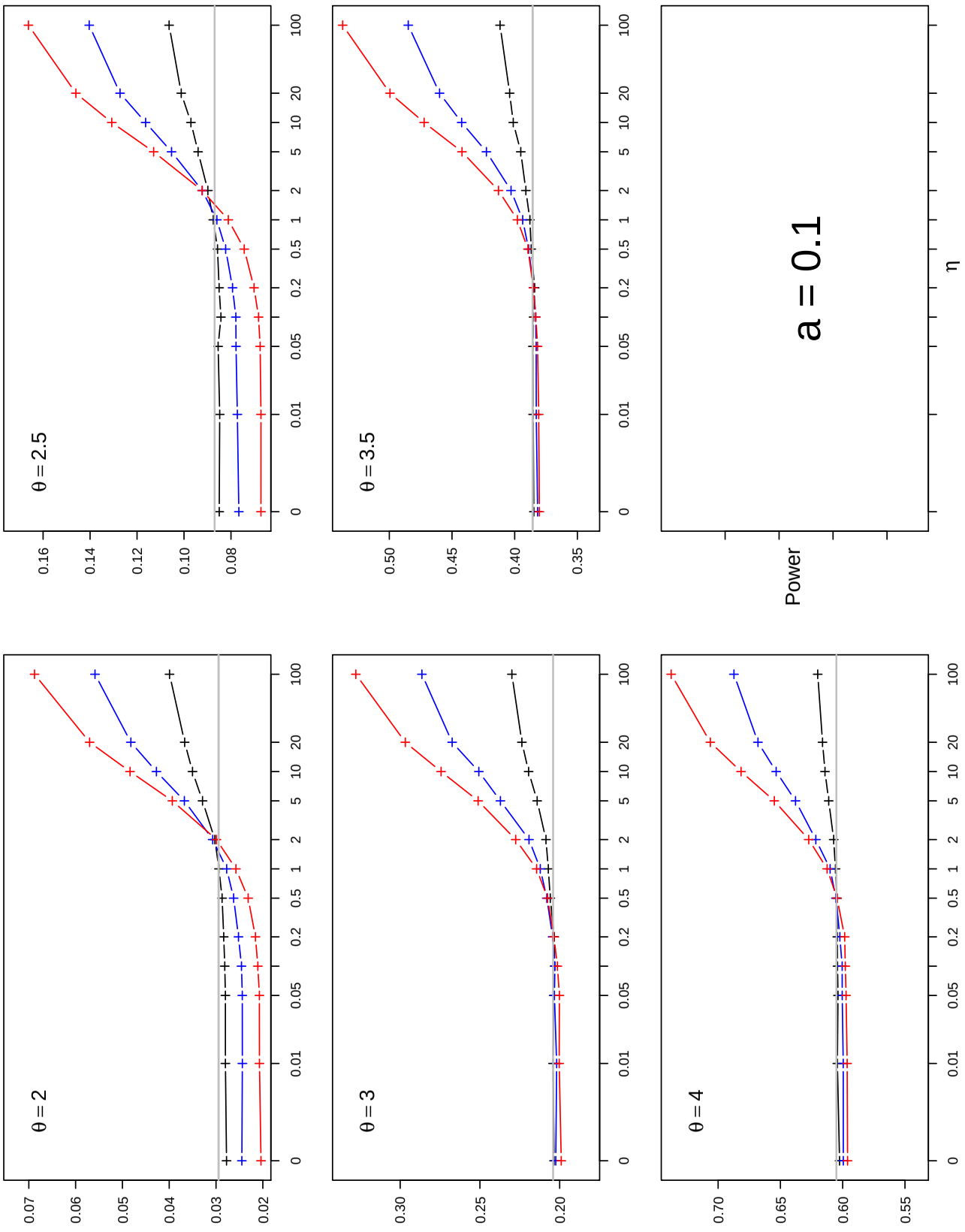

Figure 8: Power of the weighted FDX-controlling procedure as a function of η and r. The power of the weighted Holm procedure is given by the horizontal line. Vertical scales are held fixed across θ not a.