Probabilistic Model Incorporating Auxiliary Covariates to Control FDR

Lin Qiu
lin.qiu.stats@gmail.com
The Pennsylvania State University
State College, PA, USA

Nils Murrugarra-Llerena
nmurrugarrallerena@weber.edu
Weber State University
Ogden, UT, USA

Vítor Silva
vitor.silva.sousa@gmail.com
Snap Inc.
Santa Monica, CA, USA

Lin Lin
l.lin@duke.edu
Duke University
Durham, NC, USA

Vernon M. Chinchilli
vchinchi@psu.edu
The Pennsylvania State University
Hershey, PA, USA

ABSTRACT

Controlling False Discovery Rate (FDR) while leveraging the side information of multiple hypothesis testing is an emerging research topic in modern data science. Existing methods rely on the *test-level covariates* while ignoring metrics about *test-level covariates*. This strategy may not be optimal for complex large-scale problems, where indirect relations often exist among *test-level covariates* and *auxiliary* metrics or covariates. We incorporate *auxiliary covariates* among *test-level covariates* in a deep Black-Box framework (named as NeurT-FDR) which boosts statistical power and controls FDR for multiple hypothesis testing. Our method parametrizes the *test-level covariates* as a neural network and adjusts the *auxiliary covariates* through a regression framework, which enables flexible handling of high-dimensional features as well as efficient end-to-end optimization. We show that NeurT-FDR makes substantially more discoveries in three real datasets compared to competitive baselines.

CCS CONCEPTS

• **Mathematics of computing → Probabilistic algorithms**.

KEYWORDS

Social Media Content Understanding, Multiple Hypothesis Testing, FDR Control

ACM Reference Format:
Lin Qiu, Nils Murrugarra-Llerena, Vítor Silva, Lin Lin, and Vernon M. Chinchilli. 2022. Probabilistic Model Incorporating Auxiliary Covariates to Control FDR. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3511808.3557672

1 INTRODUCTION

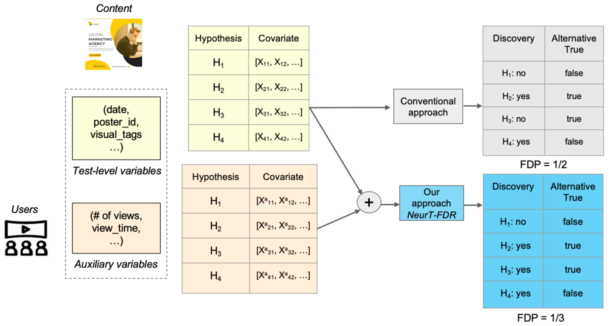
In modern statistics, from genetics, neuroimaging, to online advertising, researchers routinely test thousands or millions of hypotheses at a time [11] to discover unique data instances. Current approaches [3] solve this problem via Multiple Hypothesis Testing (MHT). MHT aims to maximize the number of discoveries while controlling the False Discovery Rate (FDR). For example, in social media, we may want to identify popular social media posts than normal ones. Also, in biology, we may want to discover which cancer cells respond positively to the treatment under a new drug.

Existing MHT approaches [8, 9, 11] only use covariate-adaptive FDR procedures on top of *test-level covariates* to improve the detection power while maintaining the target FDR. *Test-level covariates* only provide characteristics of the samples in the dataset, which can be metadata of social media posts, or genomic profiles for each cell. However, depending on the domain, we can access complementary information besides *test-level covariates* that can facilitate the work of MHT approaches. For example, as shown in Figure 1, in the social media domain, the goal is to find engaging content, and the post can be represented by visual tags and metadata information. Additionally, content consumption metrics, such as the number of views and content view time, are available. These metrics encapsulate information that facilitates MHT work. This additional information is called *auxiliary covariates* and corresponds to the samples in the dataset. More specifically, content consumption metrics do not correspond to characteristics of the sample, i.e., posted content, but how users interact in the platform to access this content. Typically, such *auxiliary covariates* are of lower dimension than those *test-level covariates* (e.g., visual tags), and are more structured.

In this paper, we present a hierarchical probabilistic black-box method which incorporates *test* and *auxiliary covariates* to control the FDR, named *NeurT-FDR*. Our main contributions can be summarized as follows:

- We pioneer the use of both *auxiliary* and the *test-level covariates* for multiple hypothesis testing problems.
- We developed a novel MHT model that jointly learns *test-level* and *auxiliary covariates* through a neural network, which enables efficient optimization and gracefully handles high-dimensional hypothesis covariates.

CIKM '22, October 17–21, 2022, Atlanta, GA, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9236-5/22/10…$15.00
https://doi.org/10.1145/3511808.3557672



Hypothesis	Covariate
H_1	$[X_{11}, X_{12}, \ldots]$
H_2	$[X_{21}, X_{22}, \ldots]$
H_3	$[X_{31}, X_{32}, \ldots]$
H_4	$[X_{41}, X_{42}, \ldots]$

Hypothesis	Covariate
H_1	$[X^a_{11}, X^a_{12}, \ldots]$
H_2	$[X^a_{21}, X^a_{22}, \ldots]$
H_3	$[X^a_{31}, X^a_{32}, \ldots]$
H_4	$[X^a_{41}, X^a_{42}, \ldots]$

Discovery	Alternative True
H_1: no	false
H_2: yes	true
H_3: no	true
H_4: yes	false

Discovery	Alternative True
H_1: no	false
H_2: yes	true
H_3: yes	true
H_4: yes	false

Figure 1: Each hypothesis has *test-level covariates* and *auxiliary covariates*. Existing covariate-adaptive FDR methods consider only *test-level covariates*, while we propose a new method (*NeurT-FDR*) to incorporate both *test-level covariates* and *auxiliary covariates* via a neural network.

2 RELATED WORK

The traditional methods for controlling FDR, such as Benjamini and Hochberg linear step-up procedure [1], and Storey's q value [7] only use the p-values and impose the same threshold for all hypotheses. To increase the statistical power, many studies have been developed to take advantage of the *test-level* information [5, 6, 9, 11]. The general formulations considered in these papers assume that each hypothesis has an associated feature vector (or called *test-level covariates*) related to the corresponding p-value.

FDRreg [6] adapts the two-groups model framework by taking into account the *test-level covariates* information to model the mixing fraction in a regression setting. IHW [5] groups the hypotheses into a pre-specified number of bins according to their associated feature space and applies a constant threshold for each bin to maximize the discoveries. One major limitation of IHW is that binning the data into groups can be tremendously difficult if the feature space is high-dimensional. NeuralFDR [9] addresses the limitation of IHW through the use of a neural network to parameterize the decision rule. This is a more general approach, and empirically it works well on a multi-dimensional feature space. AdaFDR [11] is an extension of NeuralFDR which models the discovery threshold by a mixture model using the expectation-maximization algorithm. The mixture model is a combination of a generalized linear model and Gaussian mixtures and displays improved power in comparison with IHW and NeuralFDR. However, AdaFDR only works with low-dimensional features, as its number of parameters grow linearly with respect to the covariate dimension. Thus, it is a substantial limitation for modern large-scale problems where a high-dimensional covariate setting is typical.

The recent work most relevant to ours is BB-FDR [8]. BB-FDR is the benchmark method for using a neural network to learn the true distributions of the test statistics from data in MHT. However, the existing model only deals with *test-level covariates*, while our method enables the learning from both *test-level covariates* and their associated *auxiliary covariates*, and we formulated the model in a two-stage learning structure. Our method parametrizes the test-level covariates as a neural network and adjusts the feature hierarchy through a regression framework, which enables flexible handling of high-dimensional features as well as efficient end-to-end optimization.

3 PRELIMINARIES

Consider the situation with n independent hypotheses whereby each hypothesis i produces a test statistics z_i corresponding to the test outcome. Now, each hypothesis also has k test-level covariates $\mathbf{X}_i = (X_{i1}, ..., X_{ik})' \in \mathcal{R}^k$ and q auxiliary covariates $\mathbf{X}^a_i = (X^a_{i1}, ..., X^a_{iq})' \in \mathcal{R}^q$ characterized by a tuple $(z_i, \mathbf{X}_i, \mathbf{X}^a_i, h_i)$, where $h_i \in \{0, 1\}$ indicates if the ith hypothesis is null ($h_i = 0$) or alternative ($h_i = 1$) which depends on both $\mathbf{X}_i$ and $\mathbf{X}^a_i$. The test statistics z_i is calculated using data different from $\mathbf{X}_i$ and $\mathbf{X}^a_i$. The standard assumption is that under the null ($h_i = 0$), the distribution of the test statistic z_i is from the null distribution, denoted by $f_0(z)$; otherwise z_i follows an unknown alternative distribution, denoted by $f_1(z)$. The alternative hypotheses for $h_i = 1$ are the *true signals* that we would like to discover.

The general goal of multiple hypotheses testing is to claim a maximum number of discoveries based on the observations $\{(z_i, \mathbf{X}_i, \mathbf{X}^a_i)\}^n_{i=1}$ while controlling the false positives. The most popular quantities that conceptualize the false positives are the family-wise error rate (FWER) [2] and the false discovery rate (FDR) [1]. We specifically consider FDR in this paper. FDR is the expected proportion of false discoveries, and one closely related quantity, the false discovery proportion (FDP), is the actual proportion of false discoveries. We note that FDP is the actual realization of FDR.

3.1 False discovery rate control

For a given prediction $\hat{h}_i$, we say it is a true positive or a true discovery if $\hat{h}_i = 1 = h_i$ and a false positive or false discovery if $\hat{h}_i = 1 \neq h_i$. Let $\mathcal{D} = \{i : h_i = 1\}$ be the set of observations for which the treatment had an effect and $\hat{\mathcal{D}} = \{i : \hat{h}_i = 1\}$ be the set of predicted discoveries. We seek procedures that maximize the true positive rate (TPR) also known as *power*, while controlling the false discovery rate – the expected proportion of the predicted discoveries that are actually false positives.

Definition 1. *FDP and FDR*
The false discovery proportion FDP and the false discovery rate FDR are defined as

$$FDR \triangleq:= \mathbb{E}[FDP], \qquad FDP \triangleq \frac{\#\{i : i \in \hat{\mathcal{D}} \backslash \mathcal{D}\}}{\#\{i : i \in \hat{\mathcal{D}}\}}. \tag{1}$$

In this paper, we aim to maximize $\#\{i : i \in \hat{\mathcal{D}}\}$ while controlling $FDP \leq \alpha$ with high probability.

4 METHOD

4.1 NeurT-FDR model description

As shown in Figure 2, NeurT-FDR extends the two-groups model [4] and its hierarchical probabilistic extension [8] by learning a nonlinear mapping from the *test-level covariates* (See Fig. 2-I) and their associated *auxiliary covariates* (See Fig. 2 II) jointly to model the test-specific mixing proportion (See Fig. 2-III). More specifically, the model assumes a test-specific mixing proportion λ_i which models the prior probability of the test statistics coming from the

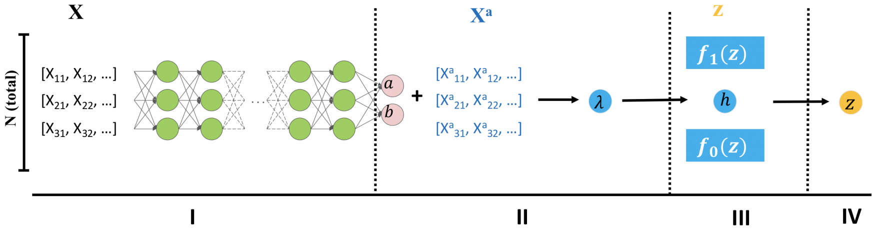


Figure 2: The graphical demonstration for NeurT-FDR. I: The deep neural network learning from the *test-level covariates* X; II: The bivariate linear regression adjustment on the beta parameters learned from the test level covariates, a, b, with the *auxiliary covariates* X^a. III: Mixing the bernoulli prior h with the estimated alternative distribution of $f_1(z)$ and $f_0(z)$ from the input z; IV: The learned statistics z from data.

alternative (i.e. the probability of the test having an effect *a priori*). Then, we place a Beta prior on each λ_i, as denoted in Eq. 2.

$$\begin{aligned} z_i &\sim h_i f_1(z_i) + (1-h_i) f_0(z_i) \\ h_i &\sim \text{Bernoulli}(\lambda_i) \\ \lambda_i &\sim \text{Beta}(a_i, b_i), \end{aligned} \tag{2}$$

Then, in order to borrow information from both $\mathbf{X}_i$ and $\mathbf{X}_i^a$ when inferencing on λ_i, ideally, one would estimate the parameters (a_i, b_i) of the Beta distribution from a neural network denoted by G using the information of both $\mathbf{X}_i$ and $\mathbf{X}_i^a$ simultaneously. However, the *test-level covariates* are usually of complex high-dimensional features, while the auxiliary features are typically more structured low-dimensional, Thus, the information contained in the high-dimensional *test-level covariates* may dominate the output of the deep neural network G [10]. Therefore, to better borrow information from the low-dimensional auxiliary features, we first learn a set of (pseudo) parameters, denoted by (a_i', b_i'), of the Beta distribution with a deep neural network G parameterized by θ_ϕ from the high-dimensional *test-level covariates* X. Then, we further propose to adjust the learned pseudo parameters from the deep neural network through a linear regression on the auxiliary features $\mathbf{X}^a$ to determine the parameters for Beta distribution, denoted by (a_i, b_i) in Eq. (3) to (4).

$$(a_i', b_i') = G_{\theta_\phi}(\mathbf{X}_i) \tag{3}$$

$$\begin{bmatrix} log_e(a_i') \\ log_e(b_i') \end{bmatrix} \sim \mathcal{N}_2 \left\{ \begin{bmatrix} \mu_a + \mathbf{X}_i^a * \boldsymbol{\delta}_a \\ \mu_b + \mathbf{X}_i^a * \boldsymbol{\delta}_b \end{bmatrix}, \begin{bmatrix} \sigma_{aa} & \sigma_{ab} \\ \sigma_{ab} & \sigma_{bb} \end{bmatrix} \right\} \tag{4}$$

Notice that $\boldsymbol{\delta}_a$ and $\boldsymbol{\delta}_b$ are the coefficients of the bivariate linear regression. After fitting the bivariate linear regression on X^a, we arrive at the fitted $\hat{a}'$ and $\hat{b}'$. Then we use the fitted mean value estimated from $\mu_a + X^a * \delta_a$, $\mu_b + X^b * \delta_b$ and the covariance matrix estimated from $cov(\hat{a}_i' - a_i', \hat{b}_i' - b_i')$ to generate the final adjusted a_i, b_i from the bivariate normal distribution.

4.2 Learning Inference

We optimize θ_ϕ by integrating out h_i from Eq. (2) and maximizing the complete data log-likelihood as follows,

$$\begin{aligned} p_\theta(z_i) = \int_0^1 & (\lambda_i f_1(z_i) + (1-\lambda_i) f_0(z_i)) \\ & \times \text{Beta}(\lambda_i | \mathbf{X}_i, \mathbf{X}_i^a) d\lambda_i \,. \end{aligned} \tag{5}$$

We opt for a beta prior because it is hierarchical and differently from other two-groups extensions, it uses a flatter hierarchy [6, 8] improving training. First, optimization is easier and more stable because the output of the function is two soft-plus activations. Second, the additional hierarchy allows the model to assign different degrees of confidence to each test, changing the model from homoskedastic to heteroskedastic.

We fit the model in Eq. (3), (4) and (2) with Stochastic Gradient Descent (SGD) on an L_2-regularized loss function,

$$\underset{\theta \in \mathcal{R}^{|\theta|}}{\text{minimize}} \quad -\sum_i \log p_\theta(z_i) + \lambda_i G_{\theta_\phi}(\mathbf{X}_i)_F^2 \,, \tag{6}$$

where $\cdot_F$ is the Frobenius norm. For computational purposes, we approximate the integral in Eq (5) by a fine-grained numerical grid. Please check the Supplementary material for estimation details.

4.3 FDR control

Once the optimized parameters $\hat{\theta}_\phi$ are chosen, we calculate the posterior probability of each test statistic coming from the alternative,

$$\begin{aligned} \hat{w}_i &= p_{\hat{\theta}}(h_i = 1 | z_i) \\ &= \int_0^1 \frac{\lambda_i f_1(z_i) \text{Beta}(\lambda_i | \mathbf{X}_i, \mathbf{X}_i^a)}{\lambda_i f_1(z_i) + (1-\lambda_i) f_0(z_i)} d\lambda_i \,. \end{aligned} \tag{7}$$

To maximize the total number of discoveries, first, we sort the posteriors in descending order by the likelihood of the test statistics being drawn from the alternative. We then reject the m hypotheses, where $0 \leq m \leq n$ is the largest possible number such that the expected proportion of false discoveries is below the FDR threshold. Formally, this procedure solves the optimization problem,

$$\begin{aligned} &\underset{m}{\text{maximize}} && m \\ &\text{subject to} && \frac{\sum_{i=1}^m (1 - \hat{w}_i)}{m} \leq \alpha \,, \end{aligned} \tag{8}$$

for a given FDR threshold α.

Table 1: Real data: # of discoveries at FDR = 0.1. Best two performers per dataset are highlighted in bold.

	Lapatinib	Nutlin-3	Airway	Visual Tags
BH[1]	117	151	4,079	312
SBH [7]	131 (+11.9%)	159 (+5.3%)	4,079	312
AdaFDR [11]	137 (+9.7%)	161 (+37.6%)	**6,050 (+48.3%)**	-
BB-FDR [8]	181 (+54.7%)	210 (+39.1%)	5,791 (+41.9%)	385 (+23.4%)
NeurT-FDRa (**ours**)	**187 (+59.8%)**	**215 (+42.3%)**	5,859 (+43.6%)	**389 (+24.7%)**
NeurT-FDRb (**ours**)	**212 (+81.2%)**	**260 (+72.2%)**	**8,820 (+116%)**	**593 (+91.9%)**

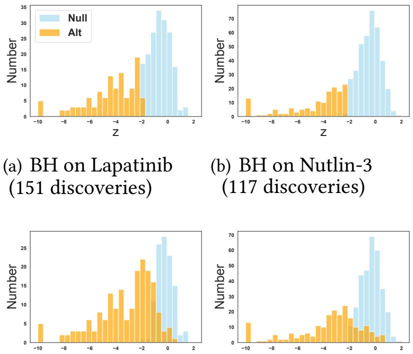
The neural network model G uses the entire *test-level* feature vector $X_{i\cdot}$ of every test to predict the prior parameters and then get adjusted by the entire *auxiliary covariate* vector $X^a_{i\cdot}$ over λ_i. The observations z_i are then used to calculate the posterior probabilities $\hat{w}_i$. The selection procedure in (8) uses these posteriors to reject a maximum number of null hypotheses while conserving the FDR.

5 CASE STUDIES

We evaluate our method [1] using three real-world scenarios. We consider BH [1], SBH [7], AdaFDR [11], BB-FDR [8], and two versions of our method, NeurT-FDRa, and NeurT-FDRb. For NeurT-FDRa, we only feed $\mathbf{X}$ into the G_{θ_ϕ}, while we stack $\mathbf{X}$ and $\mathbf{X}^a$ together and feed them into the G_{θ_ϕ} for NeurT-FDRb.

Cancer drug screening data. One goal of this analysis is to address the question of whether a given cell line responded to the drug treatment. Thus, this is a classical multiple testing problem that we need a hypothesis test for each cell line, where the null hypothesis is that the drug had no effect. We use the data preprocessed by [8] which contains genomic features and the z-score relative to mean control values for each cell line. We treat the genomic features as the *test-level covariates* and extract the rank of the z-score as the *auxiliary covariate*. For AdaFDR, we only use the *auxiliary covariates* for the model input. Table 1 (columns Lapatinib and Nutlin-3) shows for both drugs NeurT-FDRa and NeurT-FDRb achieve the largest power compared to other methods and Figure 3 shows that the *test-level covariates* and *auxiliary* features provide enough prior information that even some outcomes with a z-score above zero are still found to be significant in NeurT-FDRa.

RNA-Seq data. The original dataset contains a p-value and a log count for each gene (n=33,469), we consider the log count for each gene as the *test-level covariate* and the rank for the p-value as the *auxiliary covariate*. As the result shown in Table 1 (column Airway), where BB-FDR and NeurT-FDRa have a similar number of discoveries, AdaFDR performs slightly better and NeurT-FDRb provides 50% more discoveries than all of them. All covariate-related methods make significantly more discoveries than the non-covariate-related methods. NeurT-FDRb achieves 116% more discoveries compare to BH even when the dataset contains only one *test-level covariate*.



(a) BH on Lapatinib (151 discoveries)

(b) BH on Nutlin-3 (117 discoveries)

(c) NeurT-FDRa on Lapatinib (187 discoveries)

(d) NeurT-FDRa on Nutlin-3 (215 discoveries)

Figure 3: Discoveries found by NeurT-FDRa on the two drugs, compared to the discoveries found by a naive BH [1] approach. Blue and orange represents the null and alternative discoveries respectively.

Snap Visual Tags data. Each Snap has a visual tags vector coming from computer vision models with its corresponding content consumption metrics like how much time the particular user group spent on this Snap, number of shares, number of views, and others. So, we investigate which Snaps has the top engagements when they are compared to the normal behavior in one particular user cohort (i.e., age group and gender specification). We consider the visual tags as the *test-level covariate* and the associated 16 content consumption metrics as the *auxiliary covariates*. We used z-score as the ratio between Snap view time ratio and the number of view records to the mean values for each Snap. From our results in Table 1 (column visual tags), NeurT-FDRa and NeurT-FDRb provide significantly more discoveries than other methods, and here AdaFDR failed because it only can handle very low dimensions of covariates. AdaFDR worked in the cancer drug screening and RNA-seq data analysis when we used the rank of the test statistics as feature input. However, here we have 16 associated content consumption metrics which is a big advantage to our method since it is capable of handling both high-dimensional *test* and *auxiliary* level features' hypothesis test.

6 CONCLUSION

The neural network embedding architecture for the *test-level covariates* and the linear regression model for learning the *auxiliary covariates* enable NeurT-FDR to address modern high-dimensional problems. We believe NeurT-FDR will contribute to this field as a benchmark work for further investigation and have a wide application in neuroimaging, online advertising, and social media.

[1] https://github.com/lquvatexas/NeurT-FDR

REFERENCES

[1] Y. Benjamini and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of The Royal Statistical Society. Series B (Methodological)* 57, 1 (1995), 289–300.

[2] O. J. Dunn. 1961. Multiple comparisons among means. *J. Amer. Statist. Assoc.* 56, 293 (1961), 52–64.

[3] B. Efron. 2004. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc.* 99, 465 (2004), 96 – 104.

[4] B. Efron. 2008. Microarrays, empirical bayes and the two-groups model. *Statist. Sci.* 23, 1 (2008), 23 – 28.

[5] N. Ignatiadis, J. B. Zaugg B. Klaus, and W. Huber. 2016. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods* 13, 7 (2016), 577–580.

[6] J. G. Scott, R. C. Kelly, M. A. Smith, P. C. Zhou, and R. E. 2015. False discovery rate regression: an application to neural synchrony detection in primary visual cortex. *Journal of American Statistical Association* 110, 510 (2015), 459 – 471.

[7] J. D. Storey, J. E. Taylor, and D. Siegmund. 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 (Methodological)* 66, 1 (2004), 187–205.

[8] W. Tansey, Y. X. Wang, D. M. Blei, and R. Rabadan. 2018. Black Box FDR. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*. 4874–4883.

[9] F. Xia, M. J. Zhang, J. Zou, and D. Tse. 2017. NeuralFDR: Learning Discovery Thresholds from Hypothesis Features. In *Proceedings of the 31th International Conference on Neural Information Processing Systems (NIPS 2017)*. 1540–1549.

[10] J.L. Xu, J.W. Han, and F.P. Nie. 2017. Multi-view Feature Learning with Discriminative Regularization. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI-17)*. 3161–3167.

[11] M. J. Zhang, F. Xia, and J. Zou. 2019. Fast and covariate-adaptive method amplifies detection power in large-scale multiple hypothesis testing. https://doi.org/10.1038/s41467-019-11247-0. *Nature Communications* 10 (2019), 3433. Issue 1.