International Statistical Review

International Statistical Review (2020) doi:10.1111/insr.12430

Double Empirical Bayes Testing

Wesley Tansey[1], **Yixin Wang**[2], **Raul Rabadan**[3] **and David Blei**[2,4]

[1] *Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, New York, USA*

[2] *Department of Statistics, Columbia University, New York, New York, USA*

[3] *Department of Systems Biology, Columbia University Medical Center, New York, New York, USA*

[4] *Department of Computer Science, Columbia University, New York, New York, USA*

E-mail: david.blei@columbia.edu

Summary

Analysing data from large-scale, multiexperiment studies requires scientists to both analyse each experiment and to assess the results as a whole. In this article, we develop double empirical Bayes testing (DEBT), an empirical Bayes method for analysing multiexperiment studies when many covariates are gathered per experiment. DEBT is a two-stage method: in the first stage, it reports which experiments yielded significant outcomes and in the second stage, it hypothesises which covariates drive the experimental significance. In both of its stages, DEBT builds on the work of Efron, who laid out an elegant empirical Bayes approach to testing. DEBT enhances this framework by learning a series of black box predictive models to boost power and control the false discovery rate. In Stage 1, it uses a deep neural network prior to report which experiments yielded significant outcomes. In Stage 2, it uses an empirical Bayes version of the knockoff filter to select covariates that have significant predictive power of Stage 1 significance. In both simulated and real data, DEBT increases the proportion of discovered significant outcomes and selects more features when signals are weak. In a real study of cancer cell lines, DEBT selects a robust set of biologically plausible genomic drivers of drug sensitivity and resistance in cancer.

Key words: cancer drug studies; empirical Bayes; knockoffs; multiple testing; two-groups model.

1 Introduction

Multiple testing in most of the last century involved small-scale inference tasks, with a few dozen test statistics treated in isolation. The last 20 years have seen a fundamental change in the multiple testing landscape thanks to new high-throughput screening (HTS) techniques like DNA sequencing and fMRI scanning. Scientists are now able to perform hundreds of experiments in parallel, with each experiment containing rich contextual information about the samples under study. The statistician is then tasked with making sense of the mountain of noisy experimental outcomes while simultaneously sifting through the high-dimensional side information to find potential drivers of significance. This article addresses both of these tasks.

© 2020 International Statistical Institute.

Figure 1 shows a slice of the Genomics of Drug Sensitivity in Cancer (GDSC) dataset (Yang *et al.*, 2012), an HTS study investigating how cancer cell lines (cells taken from a malignant tumour and cultured in the lab) respond to different cancer therapeutics. The response measures how many cancer cells die over a period of time; low numbers indicate a possible sensitivity to the drug. The left panel of the figure shows the relative response of 30 different cancer cell lines ($C_1, C_2, \ldots, C_{30}$), each treated with the drug Nutlin-3. For each cell-line experiment, the treatment response (black triangles) is overlayed on top of the distribution of untreated controls (grey box plots). Even when no drug is applied, each cell line still exhibits natural variation. In analysing this data, the first question a scientist asks is whether each cell line responded to the treatment. Answering this question involves multiple hypothesis tests, one for each cell line, where the null hypothesis is that the drug had no effect.

The GDSC data also contain covariates that describe the molecular profile of each cell line under experimentation. Each covariate corresponds to a binary property of a specific gene in a cell line. The right panel of Figure 1 shows a subset of the molecular profile, with a black dot indicating the cell line is positive for that molecular feature in that gene. Biologically, molecular differences in a cell line can lead to different phenotypic behaviour that may, in turn, cause sensitivity or resistance to a drug. Statistically, this means that the likelihood of a cell line responding to treatment (the answer to the first question) is an unknown function of the cell line's molecular profile. A second scientific question is which features of cancer cells potentially drive the response of the drug. Answering this question again involves multiple hypothesis tests, one for each feature, where the null hypothesis is that the feature does not drive the response.

In a series of seminal papers (Efron *et al.*, 2001; Efron, 2003; 2004; 2008; 2019), Efron has mapped out a practical and powerful framework for multiple hypothesis testing, focusing on the types of inferences for the first question above. The foundation of this framework is the formulation of multiple testing as an empirical Bayes problem, through the ‘two-groups model’. The extra wrinkles in this article are the per-experiment covariates and the additional task of

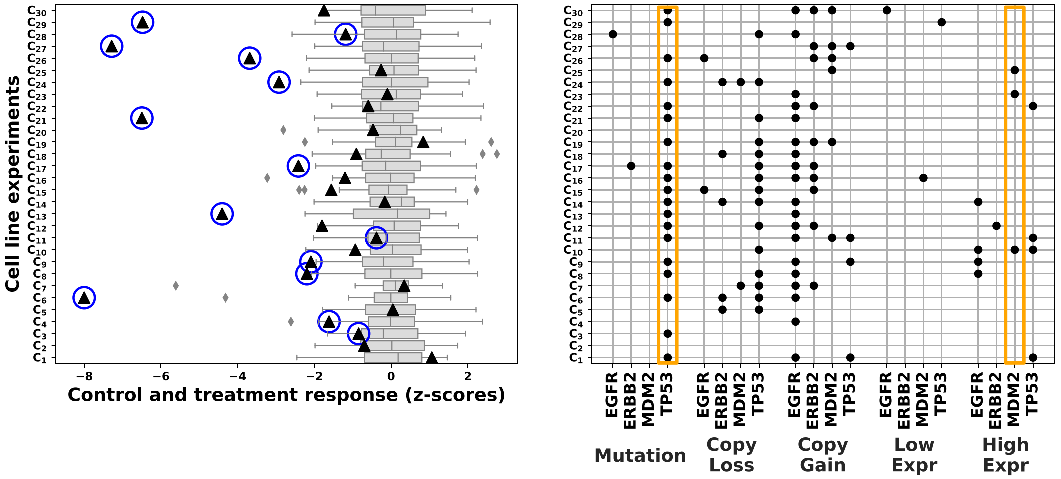


Figure 1. *Left: a subset of 30 cell line experiments from the Nutlin-3 case study in Section 7. Control replicates (grey box plots) and cell line responses (black triangles) are measured as z-scores relative to mean control values. Right: a subset of the corresponding molecular features for each experiment; black dots indicate a cell line has a recurrent mutation in the gene labelled on the x-axis. The goal in Stage 1 analysis is to select cell lines that showed a significant response (double empirical Bayes testing selections are circled in blue). In Stage 2, the molecular features are analysed to understand the mutations driving drug response (double empirical Bayes testing selections are circled in orange). [Colour figure can be viewed at* wileyonlinelibrary.com*]*

© 2020 International Statistical Institute.

finding the significant ones. Extending the two-groups model to handle the GDSC dataset, and others like it, is the goal.

The main contribution of the article is a method for analysing large, multiexperiment datasets like the GDSC in a reliable fashion and without breaking the bank on the statistician's computational budget. More generally, consider data that are a large set of experiments and a large set of covariates for each one. Formulate the two scientific questions above as a two-stage inference:

- **Stage 1**: Which individual experiments produced 'significant' or 'interesting' results?
- **Stage 2**: Which of the covariates are worthy of follow-up investigation as potential causal mechanisms of experimental significance?

The proposed method uses Efron's empirical Bayes model at both stages, and so we call it double empirical Bayes testing (DEBT).[1]

In Stage 1, DEBT builds a model of whether a given experiment generates significant outcomes a priori, where the prior depends on the per-experiment covariates through a fitted neural network. It then uses this prior model to adaptively select significant outcomes in a manner that controls the overall false discovery rate (FDR) at a specified Stage 1 level. (These are the blue circles in the left panel of Figure 1.) In Stage 2, DEBT builds a probabilistic model of the covariates themselves and uses it to perform variable selection on the Stage 1 results, while conserving a specified Stage 2 FDR threshold. (These are the orange columns in the right panel of Figure 1.) This second stage introduces an empirical Bayes extension of the model-X knockoffs framework of (Candes *et al.*, 2018) to perform simultaneous conditional independence testing on all covariates in a fast 'one-shot' manner.

1.1 Related Work

Controlling the FDR in multiple hypothesis testing has a long history in statistics. Many approaches handle the Stage 1 problem, where a collection of test statistics are observed and selecting among them is the goal. The Benjamini–Hochberg (BH) procedure (Benjamini & Hochberg, 1995) is the classic technique and still the most widely used in science. Many other methods have since been developed to take advantage of study-specific information to increase power. Recent examples include accumulation tests for ordering information (Li & Barber, 2017), the p-filter for grouping and test statistic dependency (Ramdas *et al.*, 2017), FDR-smoothing for spatial testing (Tansey *et al.*, 2017), FDR-regression for low-dimensional covariates (Scott *et al.*, 2015) and, most recently, NeuralFDR (Xia *et al.*, 2017) and AdaPT (Lei & Fithian, 2018) for high-dimensional covariates from a frequentist perspective. The empirical Bayes approach proposed here follows FDR-regression but handles data with high-dimensional covariates, as studied with NeuralFDR and AdaPT.

The Stage 2 problem of FDR-controlled covariate selection is more recent. The key idea in most methods is to introduce uninformative variables that serve as a control or null sample for comparison. Wu *et al.* (2007) was one of the earliest to consider the idea of such pseudovariables and FDR control. Linear knockoffs (Barber & Candès, 2015) brought finite-sample FDR control to the low-dimensional, linear model case with a fast one-shot procedure for selection. The generalisation to model-X knockoffs (Candes *et al.*, 2018) covers arbitrary models and arbitrary response functions, while still providing finite-sample control over FDR. In the same article, Candes *et al.* (2018) introduce the conditional randomisation test (CRT) as a more-powerful alternative to knockoffs but discard it due to computational cost. A number of papers have aimed at making CRTs more computationally efficient (Tansey *et al.*, 2018; Katsevich & Ramdas, 2020a; Liu *et al.*, 2020), but knockoffs remain significantly faster if one is willing to

© 2020 International Statistical Institute.

accept the potential reduction of power. The covariate selection method proposed here bridges the knockoff-CRT gap from the other direction: rather than make CRTs faster, it uses empirical Bayes to make knockoffs more powerful when signals are weak. Furthermore, we leverage empirical null estimation to make the selection procedure more robust to imperfect estimation of the null.

Algorithm 1 Double Empirical Bayes Testing (DEBT)

```
 1: Input: Test statistics $(z_1, \ldots, z_n)$, covariate vectors $\{x_1, \ldots, x_n\}$, thresholds $(\alpha_1, \alpha_2)$
 2: Stage 1
 3:     Estimate the empirical null of the $z$-scores $\hat{f}_0(z)$ as in Efron (2004).
 4:     Estimate the marginal $\hat{f}(z)$ (Eq. 9) with predictive recursion.
 5:     Deconvolve $\hat{f}_1(z)$ from $\hat{f}_0(z)$ and $\hat{f}(z)$ (Eq. 10).
 6:     Fit the prior model $\hat{\theta}$ (Eqs. 12 and 13).
 7:     Calculate posteriors $(\hat{w}_1, \ldots, \hat{w}_n)$ (Eq. 14).
 8:     Select discoveries $(\hat{h}_1, \ldots, \hat{h}_n)$ at the $\alpha_1$ (Eq. 15).
 9: Stage 2
10:     Fit a factor model $(\{\hat{\boldsymbol{\omega}}_i\}, \{\hat{\boldsymbol{\nu}}_j\})$ on $X$ (Eq. 19).
11:     Sample knockoffs $\tilde{x}_{ij}$ (Eq. 19).
12:     Calculate test statistics $(t_1, \ldots, t_m)$ (Eq. 22).
13:     Estimate the marginal $k(t)$ from the observed statistics (Eq. 9).
14:     Estimate the empirical null $\hat{k}_0$ (Eq. 24).
15:     Repeat lines 4–7 using $(t_1, \ldots, t_m)$ and $\hat{k}_0$ to calculate knockoff posteriors.
16:     Select discoveries $(\hat{g}_1, \ldots, \hat{g}_m)$ at the $\alpha_2$ level (Eq. 15).
17: Output: Discoveries $(\hat{h}_1, \ldots, \hat{h}_n)$ and $(\hat{g}_1, \ldots, \hat{g}_m)$.
```

Algorithm 1 presents the complete DEBT algorithm. The final algorithm runs in approximately 5 min on a 2018 MacBook Pro for each of the drugs in the GDSC dataset. For concision, we did not list the entire set of hyperparameters for the algorithm. In addition to the main parameters, the user must also provide the kernel bandwidth for predictive recursion, the stochastic gradient descent (SGD) learning rate and the number of latent factors. In our implementation, we add auto-tuning procedures over a range of values to alleviate this burden.[2]

1.2 Paper Outline

The paper builds the DEBT approach around the specific GDSC case study. Section 2 describes the data, giving details about the experiments and sample sizes. Section 3 reviews the two-groups model and then augments it to involve covariates. Section 4 leverages the results of Stage 1 to perform inference in Stage 2 (covariate selection) with an empirical Bayes knockoffs procedure. Section 5 provides theory about the augmented two groups model in Stage 1. Section 6 evaluates DEBT in a simulation setting where ground truths are known. Section 7 demonstrates both stages of DEBT on the GDSC dataset, analysing drug response in cancer cell lines. In both simulation and real data, DEBT outperforms the conventional choices, BH for Stage 1 and knockoffs for Stage 2. Section 8 presents the final discussion.

2 Setup

The data we study come from the GDSC (Yang *et al.*, 2012; Garnett *et al.*, 2012; Iorio *et al.*, 2016), an HTS study on therapeutic response in cancer cell lines. The goal of these experiments is to map the landscape of how different drugs respond to cancer and how different genes interact to drive the sensitivity or resistance to therapy.

2.1 The Data

The GDSC data comprise the results of testing 1072 cancer cell lines against 265 cancer therapeutic drugs, *in vitro*. The full data contain nearly every combination of drug and cell line. We study a single exemplar drug, Nutlin-3, for which there are 832 cell line experiments.

In each experiment, the scientist places cells from the cancer cell line in several wells, treating some with the drug and leaving others as control. They then measure the growth of the cells in each well after a 72 ho using a fluorescence assay. The outcome of the experiment is a measurement of the cell growth among treated and untreated cells. (If the drug prevents the cancer from growing, then this number will be negative.) The complete results of all the experiments are the z-scores of the treated cells, derived using the outcomes of the control wells as the null distribution,

$$\text{outcomes: } \{z_i\,, i = 1, \ldots, n = 832\}. \tag{1}$$

Each cell line is also associated with molecular information about gene mutations, copy number variations and gene expression. These covariates are preprocessed to be binary and then filtered down to those that tend to appear frequently in large-scale observational cancer studies. The binary features approximate binary biomarkers used in the clinical setting to stratify and treat patients with targeted therapies. This process leaves 236 molecular features per cell line,

$$\text{features: } \{x_{ij}\,, j = 1, \ldots, m = 236\}. \tag{2}$$

Figure 1 (left) shows outcome for treatment and control; Figure 1 (right) shows the covariates associated with these cell line experiments.

2.2 The Analysis Task

As we discussed above, there are two goals for the scientist. The first goal is the Stage 1 task: determine whether, for each experiment, the drug had an effect ($h_i = 1$) or not ($h_i = 0$),

$$\text{Stage 1 effects: } \{h_i\,, i = 1, \ldots, n\}. \tag{3}$$

This requires statistical inference. Cell line growth and response vary substantially between replicates, and measurements of cell survival are imprecise and sometimes even erroneous. Denote the Stage 2 discoveries as $\{\hat{h}_i\}$.

The second goal is the Stage 2 task: determine the relationship between the covariates x_i and the efficacy of the drug h_i. More formally, the goal is to determine whether covariate j carries unique predictive power ($g_j = 1$) or not ($g_j = 0$),

$$\text{Stage 2 effects: } \{g_j\,, j = 1, \ldots, m\}. \tag{4}$$

Mathematically, $g_j = 0$ if and only if covariate j is independent of the response, conditioned on all other features. As in Stage 1, noisy samples and measurements (let alone noisy outcomes $\hat{h}_i$) preclude us from detecting the important covariates without error. Denote the Stage 2 discoveries as $\{\hat{g}_j\}$.

© 2020 International Statistical Institute.

2.3 False Discovery Rate Control

Both goals target a similar end: maximising the number of true discoveries while minimising the number of spurious ones. There are many ways one may formalise this objective. In this article, we focus on FDR control (Benjamini & Hochberg, 1995).

A prediction $\hat{h}_i$ is called a true positive or a true discovery if $\hat{h}_i = 1 = h_i$; it is called a false positive or false discovery if $\hat{h}_i = 1 \neq h_i$. Let $\mathcal{S} = \{i : h_i = 1\}$ be the set of observations for which the treatment truly had an effect and $\hat{\mathcal{S}} = \{i : \hat{h}_i = 1\}$ be the set of predicted discoveries. We would like a method that maximises the true positive rate (TPR), also known as *power*, while controlling the FDR. The FDR is the expected proportion of the predicted discoveries that are actually false positives,

$$\text{FDR} := \mathbb{E}[\text{FDP}], \quad \text{FDP} = \frac{\#\{i : i \in \hat{\mathcal{S}} \backslash \mathcal{S}\}}{\#\{i : i \in \hat{\mathcal{S}}\}}. \tag{5}$$

FDP in (5) is the *false discovery proportion*: the proportion of false positives in the predicted discoveries from a specific experiment. While we would ideally like to control the FDP, the inherent randomness of the outcome variables makes this impossible. Modern scientific analysis typically controls for the FDR.

Similarly, the TPR is the expected proportion of true positives that are selected by the model,

$$\text{TPR} := \mathbb{E}[\text{TPP}], \quad \text{TPP} = \frac{\#\{i : i \in \hat{\mathcal{S}} \bigcap \mathcal{S}\}}{\#\{i : i \in \mathcal{S}\}}, \tag{6}$$

where TPP is the *true positive proportion*: the proportion of true positives actually selected. In both Stages 1 and 2, the objective is to maximise the TPR while controlling for FDR.

3 DEBT Stage 1: Finding Experiments With a Significant Response

Stage 1 in DEBT performs multiple testing on individual experimental outcomes. The model extends the classic two-groups formulation and approach outlined in Efron (2008).

3.1 The Augmented Two-Groups Model and Empirical Bayes

The two-groups model is a simple model of test statistics (Efron *et al.*, 2001; Efron, 2008). It posits that each statistic z_i comes from a mixture of two distributions: the null f_0 and the alternative f_1,

$$\begin{aligned} z_i &\sim h_i f_1(z_i) + (1 - h_i) f_0(z_i) \\ h_i &\sim \text{Bernoulli}(c). \end{aligned} \tag{7}$$

Whether z_i comes from the null or the alternative is coded by the latent variable h_i, this model deviates from the frequentist perspective on multiple testing by putting a prior on this variable. The BH method (Benjamini & Hochberg, 1995), for instance, sets $c = 1$ and provides a bound on the FDR, based on tail probabilities of f_0. This bound ensures the FDR will be below the target level for any choice of f_1 but sacrifices power when information about f_1 is available.

When the number of statistics is large, it is feasible to estimate (c, f_0, f_1) and perform multiple hypothesis testing (Efron, 2008). For the null distribution, estimation typically means assuming a parametric form for f_0, either a theoretical null or one where the parameters are estimated

© 2020 International Statistical Institute.

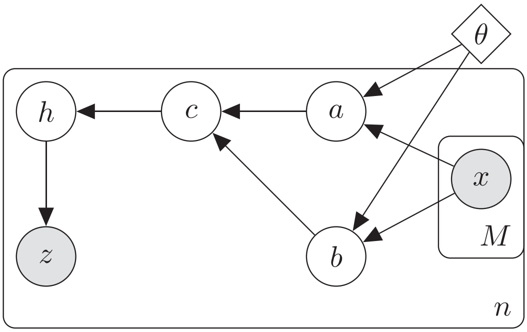


Figure 2. *The graphical model for double empirical Bayes testing.*

from data. The prior c and alternative f_1 are estimated with empirical Bayes, by estimating the marginal $f(z)$ then backing out c and f_1 from f_0 and f (Efron, 2019). Finally, the fitted two-groups model is used for multiple hypothesis testing: calculate posterior probabilities of each statistic coming from the alternative and reject those above a threshold to control for FDR.

Stage 1 of DEBT extends the two-groups model in (7) with experiment-specific priors, conditional on the covariates. An experiment-specific weight c_i models the prior probability of the test statistic coming from the alternative, that is the probability of the treatment having an effect a priori. We place a beta prior on each experiment-specific prior c_i and model the parameters of the beta with a black-box function of the covariates, $\phi(x;\theta)$. To handle the large number of covariates and the possible nonlinear interactions between them, we choose ϕ to be a deep neural network, where θ are the weights of the network.

The augmented two-groups model is

$$\begin{aligned} z_i &\sim h_i f_1(z_i) + (1-h_i) f_0(z_i) \\ h_i &\sim \text{Bernoulli}(c_i) \\ c_i &\sim \text{Beta}(a_i+1, b_i+1) \\ a_i, b_i &= \phi(x_i;\theta)\,. \end{aligned} \tag{8}$$

Figure 2 shows the DEBT graphical model.

The beta prior is a departure from other two-groups extensions to covariates, which use a flatter hierarchy and directly learn a predictive model for c_i (Scott *et al.*, 2015; Tansey *et al.*, 2017). We found the flat approach to be difficult to fit and lead to degenerate functions $\phi(x;\theta)$ that always predicted the global mean as the prior. In contrast, the hierarchical prior allows the model to assign different degrees of confidence to each experiment (via the beta distribution), which is a heteroskedastic model. The beta parameters are both incremented by 1 to ensure concavity of the beta distribution.

3.2 *Inference*

DEBT takes four steps to complete Stage 1 inference.

First, DEBT estimates f_0, the distribution of the statistic for ineffective drugs. We follow the zero assumption approach of Efron (2008), assuming most z-scores near zero are null and $f_0(z)$ follows $\mathcal{N}(\mu_0, \nu_0)$ with unknown shape and scale. DEBT fits a $df = 5$ degree polynomial around the central region ($z \in [-3, 3]$) of the observed statistics. The central peak of the fit is used to estimate (μ_0, ν_0) using a second-order Taylor approximation to the normal density.

© 2020 International Statistical Institute.

DEBT then fixes f_0 and estimates the alternative f_1. To estimate f_1, it temporarily considers the original two-groups model and follows the g-modelling approach with predictive recursion. In detail, use predictive recursion (Newton, 2002) to estimate the marginal $f(z)$,

$$f(z) = (1-c)f_0(z) + c\int \mathcal{N}(z;\delta,\sigma_1^2)d\delta. \tag{9}$$

Using predictive recursion in this way follows other recent extensions to the two-groups model (Scott *et al.*, 2015; Tansey *et al.*, 2017). Predictive recursion enjoys strong empirical performance and consistency guarantees (Tokdar *et al.*, 2009).

Bayes' rule then yields an estimate of $f_1(z)$ via the following identities:

$$f_1(z) = 1 - \frac{(1-c)f_0(z)}{f(z)} \tag{10}$$

$$c = \mathbb{E}_{z\sim f(z)}\left[\frac{f_1(z)}{f(z)}\right]. \tag{11}$$

Modelling the marginal instead of f_1 directly reduces solving a difficult deconvolution problem to a simpler density estimation problem (Efron *et al.*, 2001; Efron, 2008).

Returning to the augmented model, DEBT fixes the null f_0 and alternative f_1 and fits the conditional prior $\phi(x_i;\theta)$. It optimises θ by integrating out the significance indicator h_i and maximising the complete data log-likelihood,

$$p_\theta(z_i) = \int_0^1 (c_i f_1(z_i) + (1-c_i)f_0(z_i))\text{Beta}(c_i|\phi(x_i;\theta))dc_i. \tag{12}$$

Specifically, DEBT uses SGD with L_2-regularisation,

$$\underset{\theta\in\mathbb{R}^{|\theta|}}{\text{minimise}} \quad -\sum_i \log p_\theta(z_i) + \lambda\phi(x_i;\theta)_F^2, \tag{13}$$

where $||\cdot||_F$ is the Frobenius norm. (In pilot studies, we found adding a small amount of L_2-regularisation prevented overfitting at virtually no cost to statistical power.) For computational purposes, we approximate the integral in (12) by a fine-grained numerical grid. More optimisation details are available in Appendix A1.

Finally, DEBT follows the two-groups method to estimate the treatments with significant responses. It calculates the posterior probability of each test statistic coming from the alternative,

$$\begin{aligned}\hat{w}_i &= p_{\hat{\theta}}(h_i = 1|z_i, x_i)\\ &= \int_0^1 p(h_i = 1|c_i, z_i)p_{\hat{\theta}}(c_i|x_i)dc_i\\ &= \int_0^1 \frac{c_i f_1(z_i)\text{Beta}(c_i|\phi(x_i;\hat{\theta}))}{c_i f_1(z_i) + (1-c_i)f_0(z_i)}dc_i,\end{aligned} \tag{14}$$

which uses the fitted per-experiment prior.

Assuming the posteriors are accurate, rejecting the i-th hypothesis will produce $1-\hat{w}_i$ false positives in expectation. Therefore, DEBT maximises the total number of discoveries by a step-down procedure, as in the original two-groups model (Efron, 2008). Sort the posteriors

© 2020 International Statistical Institute.

in descending order by the likelihood of the test statistics being drawn from the alternative; then reject the first q hypotheses, where $0 \leq q \leq n$ is the largest possible index such that the expected proportion of false discoveries is below the FDR threshold. This procedure solves the optimisation problem,

$$\begin{aligned} \underset{q}{\text{maximise}} \quad & q \\ \text{subject to} \quad & \frac{\sum_{i=1}^{q}(1-\hat{w}_i)}{q} \leq \alpha\,, \end{aligned} \tag{15}$$

for a given FDR threshold α. (By convention $\frac{0}{0} = 0$.)

4 Stage 2: Identifying Important Covariates

The Stage 1 portion of DEBT produces two quantities:

1. Posterior probability estimates $(\hat{w}_1, \ldots, \hat{w}_n)$ for each of the experiments being a success, 14.
2. Parameter estimates $\hat{\theta}$, for the neural-network prior that maps the covariates x to the probability of an experiment being a success.

With these new quantities, DEBT enters a second stage of inference. In this stage, the goal is to use $\hat{w}$ and ϕ to understand which of the covariates x are responsible for driving the outcomes z.

Figure 1 (right) illustrates the second stage. Unlike the Stage 1 inference task (Figure 1, left), identifying the important covariates is a cross-cutting concern. For each variable, its importance is assessed across all experiments. As with Stage 1, the goal will be to select as many true positives (i.e. important variables) as possible while controlling the FDR.

4.1 Conditional Randomisation Tests, Knockoffs and Variable Selection in DEBT

We formalise inference as a multiple testing problem, where the null hypothesis is conditional independence,

$$X_j \perp\!\!\!\perp Z | X_{-j}\,, \tag{16}$$

where X_{-j} is every feature in X except X_j.[3] Under the null, the feature X_j contains no additional information about Z that is not contained in the other features X_{-j}.

Testing 16 is challenging when using the deep neural network for ϕ in 8. Simply inspecting the parameter weights θ, as one might in a linear model, will not be sufficient to extract non-null features reliably. Neural network models are black boxes, with no analytic null distribution.

A recent work in the frequentist testing literature has produced two methods for variable selection: CRTs and model-X knockoffs (Candes *et al.*, 2018). The CRT repeatedly resamples from the null distribution for X_j and calculates the test statistic using the null sample,

$$\tilde{X}_j \sim P(X_j | X_{-j}), \quad \tilde{T}_j = \mathcal{T}(\tilde{X}_j, X_{-j}, Z)\,. \tag{17}$$

Given a collection of r null test statistics $(\tilde{t}_j^{(1)}, \ldots, \tilde{t}_j^{(r)})$, a one-sided p value can be calculated for the true feature test statistic t_j, and standard multiple testing tools can be brought to bear. The CRT is powerful, but computationally expensive. A number of approaches have sought to reduce this computational burden (Tansey *et al.*, 2018; Katsevich & Ramdas, 2020a; Liu *et al.*, 2020), but CRTs are still too expensive to run (on a laptop) in a few minutes for high-dimensional machine learning models.

Candes *et al.* (2018) addresses this computational issue with the model-X knockoffs approach, which is a single-shot procedure. To perform inference, generate a 'knockoff' feature $\tilde{X}$ for every feature in the dataset and then fit a distribution of both the original and knockoff features. Valid knockoffs satisfy two key conditions,

$$\begin{aligned} X &\overset{d}{=} \tilde{X} \\ (X, \tilde{X}) &\overset{d}{=} (X, \tilde{X})_{\mathrm{swap}(\mathcal{J})}, \end{aligned} \tag{18}$$

where the second condition states that any subset of columns $\mathcal{J} \subseteq \{1, \ldots, m\}$ can be swapped between X and $\tilde{X}$ and the joint distribution remains the same. With a single knockoff sample in hand, Barber & Candès (2015) propose a step-up selection procedure to select features with finite sample (frequentist) control over the FDR. Typically, the procedure selects on the difference of variable importance heuristics $(\eta_1, \ldots, \eta_m)$, such as the difference in lasso coefficient magnitudes between the original feature and its corresponding knockoff.

By avoiding the need to resample, knockoffs dramatically lower the computational cost of simultaneous conditional independence testing. But this speed-up comes at a cost: power. Generally, knockoffs are less powerful than CRTs (Candes *et al.*, 2018), and so once one has valid knockoffs, maximising power becomes a top criterion. Stage 2 of DEBT boosts power in this stage by merging model-X knockoffs with empirical Bayes. We first discuss two requirements of model-X knockoffs: generating the knockoffs and choosing a test statistic.

4.1.1 Generating knockoffs with a logistic factor model

To generate knockoffs, it suffices to find a latent factor model such that the X_j are conditionally independent (Liu & Zheng, 2018; Bates *et al.*, 2020),

$$P(X_1, \ldots, X_m | U) = \prod_{j=1}^{m} P(X_j | U). \tag{19}$$

Conditioned on finding the latent factors U, sampling from 19 generates valid knockoffs.

In the GDSC data, the molecular features are binary. We use a logistic factor model,

$$P(x_{ij} | \boldsymbol{\omega}_i, \boldsymbol{\nu}_j) = \prod_{j=1}^{m} \mathrm{Bern}(\sigma(\boldsymbol{\omega}_i^{\top} \boldsymbol{\nu}_j)), \tag{20}$$

where σ is the logistic function,

$$\sigma(\rho) = \frac{1}{1 + e^{-\rho}}. \tag{21}$$

The number of factors is a modelling decision; for the GDSC dataset, we use 20 latent factors. The model is fit by maximum likelihood, running an alternating minimisation algorithm until numerical convergence to a local minimum. Once the factor model has been fit, each knockoff feature $\tilde{x}_{ij}$ is drawn independently, conditioned on the latent variables $(\boldsymbol{\omega}_i, \boldsymbol{\nu}_j)$.

© 2020 International Statistical Institute.

4.1.2 The choice of test statistic

As a measure of variable importance, DEBT uses the change in the posterior probability of z_i coming from the Stage 1 alternative f_1. For each knockoff, $\tilde{x}_{ij}$, DEBT swaps out the original feature for the knockoff and calculates the posterior $\tilde{w}_i^{(j)} = p_{\hat{\theta}}(h_i = 1|z_i, \tilde{x}_i)$ from 14.

The test statistic is the difference between posterior entropies, with and without the knockoff,

$$\begin{aligned} \text{Entropy}(w) &= -\sum_i w_i \log w_i - \sum_i (1 - w_i) \log(1 - w_i) \\ t_j &= \text{Entropy}(\hat{w}) - \text{Entropy}(\tilde{w}) . \end{aligned} \tag{22}$$

If a feature is useful in predicting an outcome, then it should (stochastically) reduce the overall entropy of the posterior, relative to the null. By definition, a feature sampled from the null adds no new information to the model; it cannot systematically reduce the entropy. The entropy statistic in 22 can analogously be thought of as a difference in empirical risk, which has been shown to be an optimal choice of test statistic (Katsevich & Ramdas, 2020b).

4.2 Empirical Bayes Knockoffs

Given the collection of test statistics $\{t_1, \ldots, t_m\}$, the last step of Stage 2 is to select among them. This is where DEBT differs from the frequentist knockoffs approach.

Again, DEBT takes an empirical Bayes view. We assume that, given the choice of test statistic, the *distribution* of null statistics is symmetric. (This is a stronger assumption than the frequentist knockoff filter where, under the null hypothesis, only the sign of each knockoff statistic is independent and a symmetric fair coin flip.) With this assumption in place, we then use the familiar empirical Bayes method of Efron (2008) to boost power in the variable selection problem.

To model the knockoffs, adopt the original two-groups model,

$$\begin{aligned} t_j &\sim g_j k_1(t_j) + (1 - g_j) k_0(t_j) \\ g_j &\sim \text{Bernoulli}(\zeta) . \end{aligned} \tag{23}$$

This is the classic multiple testing setup considered in Efron *et al.* (2001), because there is no side information about the knockoff statistics. We estimate the null k_0, the alternative k_1, and the prior ζ. We calculate posterior probabilities of coming from the alternative, and then use the step-down procedure to control FDR. As in Stage 1, we use predictive recursion to estimate the margial $k(t)$, including an estimate of the prior ζ; we estimate k_1 using k and k_0.

The wrinkle here, however, is that the null distribution is estimated in a different way, which is particular to the knockoffs setting. Similar to Efron (2004), there are reasons to doubt the theoretical null in the empirical Bayes knockoffs setup.

DEBT leverages two properties of the knockoff statistics to estimate the empirical null. First, the null distribution, although not necessarily normal, is assumed to be symmetric if the factor model has been estimated well. Second, the alternative distribution should concentrate nearly all its mass on the left of the central peak of the null distribution. This motivates an extension of the traditional empirical null of Efron (2004) to a *one-sided* empirical null estimator.

The first step is to estimate $k(t)$, the marginal distribution of the observed statistics. We follow Efron (2004) and use Lindsey's method (Efron & Tibshirani, 1996) with a 5th-degree polynomial Poisson regression on the histogram bin counts. The central peak c is then located in the polynomial by filtering down to the centre 35% of the data and taking the value of c for which

© 2020 International Statistical Institute.

$\hat{k}(t)$ is maximised. Once the central peak is located, we assume all points to the right represent null observations and use a mirrored estimation approach,

$$k_0(t) = \begin{cases} \frac{k(t)}{2\int_c^{\inf} k(t)dt}, & \text{if } t \geq c \\ \frac{k(2c-t)}{2\int_c^{\inf} k(t)dt}, & \text{if } t < c\,. \end{cases} \tag{24}$$

The estimate of the null distribution for points above c in 24 is proportional marginal probability. The density on the left-hand side of c is proportional to the mirrored image of the marginal on the right-hand side of c. This is similar in design to the empirical null used in Efron *et al.* (2001).

4.3 The Choice of a Null Hypothesis

Why do we estimate the empirical null in this way? Theoretically, the null distribution for test statistics in both stages should be centred at 0 and symmetric. Efron (2004) outlines how the null distribution may deviate from the theoretical null due to latent confounders, technical error or a large number of small-but-uninteresting effects may cause the null distribution. In practice, estimating the null empirically from the data is essential to prevent spurious false positives.

In Stage 2, the null knockoff statistics in 22 are summations of differences of negative log-likelihoods. Standard central limit theorem assumptions could be applied to straightforwardly derive an asymptotic normal approximation,

$$k_0(t/n) \overset{d}{\rightarrow} \mathcal{N}(0, \sigma^2)\,, \tag{25}$$

where σ^2 is the population variance. Similar to Efron (2004), there are reasons to doubt this theoretical null in the empirical Bayes knockoffs setup.

First, the experimental design of the HTS experiments produces correlations between presumed-independent test statistics. Each HTS experiment is conducted on microwell plates, similar to the DNA microarray setup. These plates contain a single batch of controls but many different treatments. The z-scores for each treatment on the same plate are estimated using the same set of controls, inducing dependence between the plates. Further, wells near each other on the same plate, or run on different plates but in the same lab on the same day, tend to have similar z-scores independent of the treatment applied. These 'batch effects' are a common problem in HTS experiments (Mazoure *et al.*, 2017), which leads to violations of the independence assumption.

Second, the empirical Bayes approach used in DEBT explicitly reuses the data. The same dataset is used first to estimate $\hat{\theta}$, the prior parameters, and then again to calculate the knockoff statistics. This double-dipping is routine in empirical Bayes methods but is lamented by those demanding strict independence in the test statistics (Efron, 2019).

Third, the factor model parameters $(\hat{\omega}, \hat{\nu})$ must be estimated from the data. If the estimate is too aggressive, capturing independent sources of variation as well as shared variation, the power of the model will be low. It therefore behoves the statistician to estimate a minimal factor model—one that accounts for all of the shared variation and no more. If, in the pursuit of higher power, the estimated factors fail to render the covariates conditionally independent, there will be lingering dependence between the test statistics.

© 2020 International Statistical Institute.

5 FDR Control and Power Analysis

We discuss the FDR control and the power of Stage 1 of DEBT (DEBT-1). We show that the existing theory about the two-groups model still holds in the augmented model of 8. (Stage 2 of DEBT adopts the original two groups model, so the existing theory about the two groups model applies.)

The existing theory about the two groups model says that local FDR controls average FDR at a specified level (Efron, 2005), assuming that the prior probability of the null is known. Moreover, Lei & Fithian (2018) show that under FDR control, local FDR is the most powerful rejection rule. Below, we present two analogous propositions for the augmented model.

Begin with the FDR control. Assume that we know the local probability of the alternative c_i. Suppose the set of independent test statistics $z = (z_1, \ldots, z_n)$ are drawn from

$$z_i \sim c_i^* f_1^*(z_i) + (1 - c_i^*) f_0^*(z_i), i = 1, \ldots, n.$$

Further consider the ideal setting where DEBT-1 learns the true data-generating prior probabilities $(c_1^*, \ldots, c_n^*)$ and null and alternative distributions f_0^*, f_1^*. In this setting, the average FDR will be bounded by α if we apply DEBT-1 with an FDR threshold α.

Proposition 1 (FDR control of DEBT-1). *Assume DEBT-1 learns the true data-generating prior probabilities, Beta*$(\phi(x_i; \theta)) = \delta_{c_i^*}, i = 1, \ldots, n$ *and the true data-generating null and alternative distributions* $(f_0, f_1) = (f_0^*, f_1^*)$. *Then, DEBT-1 with an FDR threshold* α *controls FDR at* α,

$$|Z_1|^{-1} \sum_{i: z_i \in Z_1} FDR(z_i) \leq \alpha,$$

where $Z_1 \subset \{z_1, \ldots, z_n\}$ *denotes the set of rejected hypotheses and* $FDR(z_i)$ *denotes the FDR of hypothesis i.*

Proposition 1 shows that DEBT-1 is tight in FDR control, that is DEBT-1 controls FDR at the nominal rate α a user sets. To prove Proposition 1, we show that the expected posterior probabilities of the rejected hypotheses coincides with their FDR $\mathrm{P}(H_i = 0 | z_i$ is rejected). The proof is in Appendix B1.

Next, we study the power of DEBT-1. We will show that DEBT-1 maximises the power under FDR α by connecting it to rejecting with local FDR surfaces. Again assume that c_i is known.

Proposition 2 (Power analysis of DEBT-1). *Assume DEBT-1 learns the true data-generating prior probabilities, Beta*$(\phi(x_i; \theta)) = \delta_{c_i^*}, i = 1, \ldots, n$ *and the true data-generating null and alternative distributions* $(f_0, f_1) = (f_0^*, f_1^*)$. *Then, DEBT-1 maximises the power under FDR control at level* α.

Proposition 2 shows that DEBT-1 maximises power. To analyse the power of DEBT-1, we draw a connection between DEBT-1 and rejecting hypotheses based on the local FDR surface. We show that any solution to DEBT-1 can be generated by rejecting hypotheses based on the local FDR surface. Because DEBT-1 is tight in FDR control and the local FDR rejection rule maximises the power (Lei & Fithian, 2018), DEBT-1 also maximises the power. The proof is in Appendix C1.

Propositions 1 and 2 establish the FDR control of DEBT-1 and show it maximises power. They provide theoretical guarantees for DEBT-1 and demonstrate its optimality. Note that both theorems rely on the key assumptions that DEBT-1 learns the true prior probabilities

© 2020 International Statistical Institute.

and null/alternative distributions. While these assumptions are never true with finite datasets, previous work shows Stage 1 of DEBT-1 to be robust and empirically powerful (Tansey *et al.*, 2018).

6 Simulation

We study the performance of DEBT on a simulation setup that resembles the real data case study in Section 7. The simulated data contains $n = 500$ samples and $m = 100$ features. The features are binary and drawn according to a correlated latent variable model,

$$\begin{aligned} \Sigma_{jk} &= e^{|j-k|} \\ \boldsymbol{\xi} &\sim \mathcal{N}(\mathbf{0}, \Sigma) \\ X_j &\sim \mathrm{Bern}(\mathrm{logistic}(\xi_j)), \end{aligned} \tag{26}$$

where the function logistic is the logistic function,

$$\mathrm{logistic}(x) = \frac{1}{1 + e^{-x}} . \tag{27}$$

The responses z are drawn from a sparse, second-order interaction model,

$$\begin{aligned} \psi_i &= \sum_{j \in \mathcal{S}} \beta_{j,0} X_{ij} + \sum_{(j,k) \in \mathrm{pairs}(\mathcal{S})} \beta_{j,k} X_{ij} X_{ik} - 1 \\ h_i &\sim \mathrm{Bern}(\mathrm{logistic}(\psi_i)) \\ z_i &\sim \mathcal{N}(-2h_i, 1) . \end{aligned} \tag{28}$$

The pairs function divides $\mathcal{S}$ into two sets uniformly at random to create pairwise interactions in addition to the linear terms. The alternative distribution is a normal distribution with mean -2, whereas the null is centred at 0; both distributions have variance 1. Offsetting the logits ψ_i by -1 leads to a sparser set of non-null results in Stage 1, representing the sparse results we typically expect in real data. Each coefficient $\beta_{j,0}$ and $\beta_{j,k}$ is drawn i.i.d. from a centred normal,

$$\mathcal{N}\left(0, \left(\frac{s}{\frac{3}{2}|\mathcal{S}|}\right)^2\right) , \tag{29}$$

where s is the signal strength. We set $s = 3$ and select $|\mathcal{S}| = 20$ features uniformly at random without replacement. These settings lead to 20% true non-null features and approximately 20% non-null z scores on average. We run 100 independent trials and compare the results of DEBT in Stages 1 and 2 to BH and knockoffs, respectively, with a nominal FDR of 0.2 in both stages.

Figure 3 shows the number of discoveries made by DEBT in comparison with the baseline methods. In Stage 1 (left panel), DEBT consistently outperforms BH, with roughly twice as many discoveries in most trials. In Stage 2 (right panel), DEBT performs similarly to knockoffs when both methods discover large numbers of features. This is indicated by the overlapping lines starting at the minimum of $x = 6$ true discoveries. When the number of discoveries is small, however, DEBT outperforms knockoffs. This is due to the knockoff filter requiring an offset term in the numerator and denominator that makes rejecting a small number of features impossible while still controlling FDR. Thus, when signals are sparse (as in many biological experiments), DEBT will be able to provide an answer while knockoffs cannot.

© 2020 International Statistical Institute.

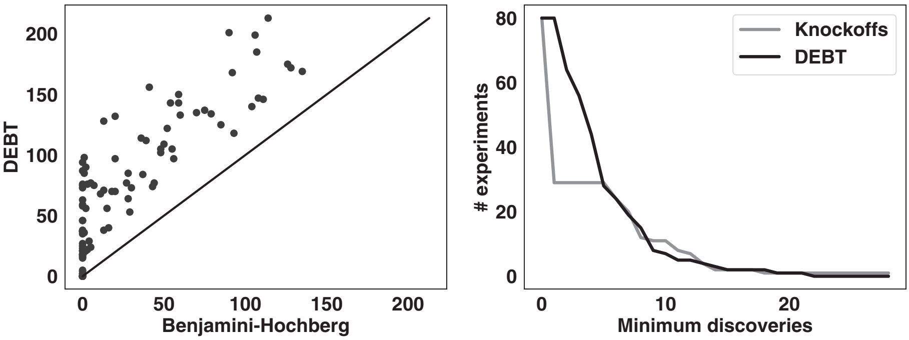


Figure 3. *Power comparisons for DEBT in simulation. Left: Stage 1 results compared with Benjamini–Hochberg (BH); each point is a single trial with the number of true positives discovered by each method. Right: Stage 2 results compared with model-X knockoffs (Knockoffs); the curves show the number of trials where the method selected at least x features for each point on the x-axis. DEBT, double empirical Bayes testing.*

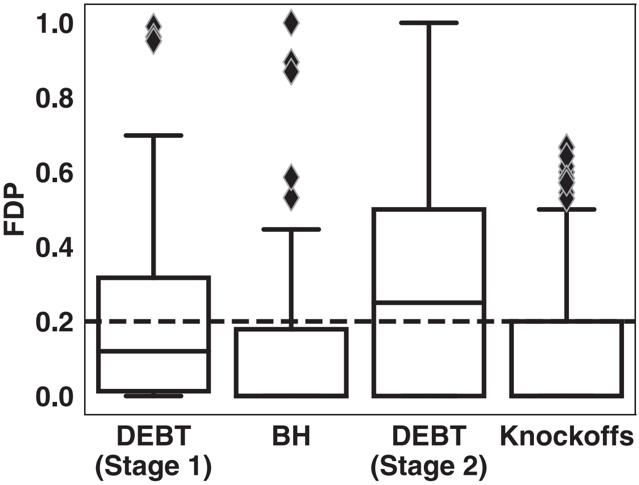


Figure 4. *False discovery proportions for DEBT and the two baseline methods in each of the two stages of inference. The baseline methods, BH and knockoffs, tend to be conservative; DEBT has an FDP closer to the target (dashed line). BH, Benjamini–Hochberg; DEBT, double empirical Bayes testing.*

Figure 4 shows the false discovery proportions for each method across all 100 simulations. The two baseline methods (BH and knockoffs) are generally conservative. Both baseline methods typically have a much lower FDP than the nominal FDR target of 0.2, as indicated by the central line in the box plots. The two stages of DEBT inference, however, have results much closer to the target level. The box plot central lines show the median rather than the average FDP. In DEBT, the average FDP is 20.22% in Stage 1 and 30.86% in Stage 2. While the Stage 2 average FDP is above the nominal target, the small number of true positives in Stage 2 leads to a high-variance FDP (standard error 3.20% after 100 trials). Thus, it is possible that the FDR is closer to the target rate and the inflation is a side effect of finite trials. Further, although the knockoffs method controls the FDR in Figure 4, the real data analysis in Section 7.2 shows knockoffs are more brittle under less ideal data modelling conditions and DEBT produces more robust Stage 2 discoveries.

© 2020 International Statistical Institute.

7 Cancer Drug Screening

We use DEBT to study cancer drug effects in the GDSC (Yang *et al.*, 2012) study. We first focus on the drug Nutlin-3, a well-studied drug designed to target a specific protein. Nutlin-3 has many known biomarkers of sensitivity and resistance, making it an ideal case study for a deep dive into the meaningfulness of the discoveries. We then run on all drugs in the GDSC and report high-level discovery statistics mirroring those presented in the simulation study. In both the deep dive and multidrug study, DEBT finds more Stage 1 discoveries than BH and more plausible Stage 2 results than model-X knockoffs.

The data for each drug are the results of experiments on n cell lines with m features. In each experiment, some cells are treated with the drug, and others are untreated (control) cells. The result of each experiment is the difference in cell growth between the treated and untreated. Following Algorithm 1, DEBT analyses this data in two stages. In the first stage, it determines which cell lines responded to the drug; in the second stage, it determines which molecular features drive sensitivity and resistance to the drug.

7.1 Case Study: Nutlin-3

We first focus on the experiments for the drug Nutlin-3, with $n = 832$ cell lines and $m = 236$ features. Nutlin-3 is a well-studied drug with a known target and mechanism of action. Further, many biomarkers for predicting the effectiveness of Nutlin-3 have been discovered experimentally in previous studies. Thus, true positive features in Stage 2 are likely to appear in the literature, making assessing their biological plausibility more viable.

7.1.1 Stage 1 analysis

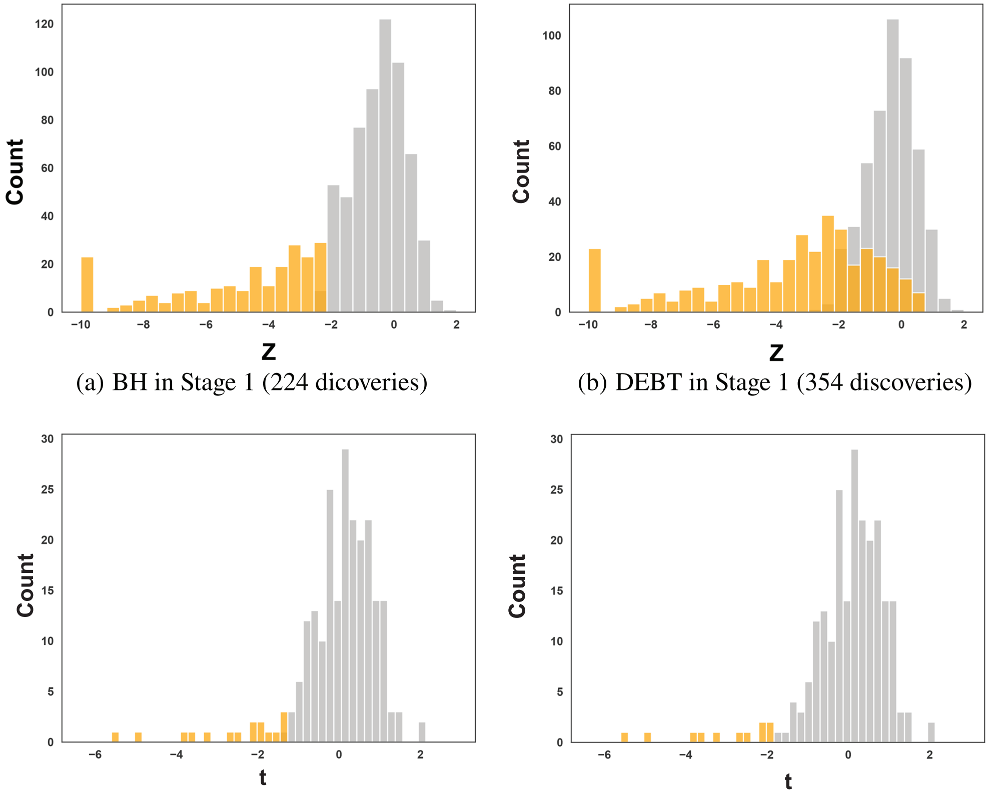
Figure 5 shows the aggregate number of treatment effects discovered by BH, model-X knockoffs and DEBT. DEBT reports 130 more discoveries in Stage 1 discoveries compared with BH.

The molecular profiles of the cell lines provide enough prior information that even some outcomes with a z-score above zero are found to be significant. This is possible for two reasons. First, the flexibility provided by the prior model of the covariates has the ability to lower the bar for test statistics. In extreme cases, where the model predicts a high prior probability of success, the data are unlikely to outweigh the alternative. To see this visually, Figure 6 shows the prior, outcomes and posteriors. On the far right, points with high prior probability can have positive z-scores but still have high probability of posterior success.

The other reason for rejecting z-scores with high values is due to an idiosyncrasy of FDR. In the Bayesian setup, FDR is formulated as an average over posterior probabilities. This average creates a rolling budget for each successive posterior. If the posteriors are sorted in decreasing order $(\hat{w}_{(1)}, \ldots, \hat{w}_{(n)})$, then the rejection threshold for the i-th posterior is a function of the posteriors,

$$\alpha^{(i)} = i * (1 - \alpha) - (i - 1) \sum_{\ell=1}^{i-1} \hat{w}_{(\ell)} . \tag{30}$$

As long as $\hat{w}_{(i)} \geq 1 - \alpha^{(i)}$, the null hypothesis will be rejected while conserving FDR. Thus, if the first few posteriors indicate high probability that the treatment is significant, then even experiments with a dubious chance of significant treatment can be added without violating FDR.



(a) BH in Stage 1 (224 dicoveries)

(b) DEBT in Stage 1 (354 discoveries)

(c) Knockotts in Stage 2 (11 discoveries)

(d) DEBT in Stage 2 (16 discoveries)

Figure 5. *Discoveries found by DEBT on the drug Nutlin-3, compared with the discoveries found by a naive BH approach. DEBT leverages the molecular profiling information of the cell lines to identify more discoveries at the same FDR threshold. BH, Benjamini–Hochberg; DEBT, double empirical Bayes testing. [Colour figure can be viewed at* wileyonlinelibrary.com*]*

7.1.2 Stage 2 analysis

Table 1 lists the genes reported by DEBT in Stage 2. The first column lists the discoveries reported by both DEBT and the knockoffs filter; the second column lists the discoveries reported by DEBT but not selected by the knockoffs filter. Interpreting the quality of the results requires familiarity with genomics and cancer biology. Below, we briefly detail the scientific rational behind the biological plausibility of some of the Stage 2 results and refer the reader to Weinberg (2013) for a full review.

Nutlin-3 is an inhibitor of the oncogene MDM2. The MDM2-encoded protein tags the P53 protein for ubiquitylation. When highly overexpressed, MDM2 can functionally inactivate TP53. By targeting MDM2, Nutlin-3 enables a nonmutated (‘wild type’) TP53 to trigger apoptosis in cancer cells. However, if TP53 is mutated, Nutlin-3 will be ineffective, and hence, its mutation state is an important driver of Nutlin-3 sensitivity. The discovery by both DEBT and knockoffs that TP53 mutation and MDM2 overexpression are important features is a good indication that the selected features match biological processes underlying the cells.

Five genes were selected by DEBT but not by the knockoffs filter. Each of these five cases has biological plausibility and scientific evidence to support its selection as a driver of response. In cases where Nutlin-3 produces TP53-independent effects, it has been shown that Nutlin-3 interacts with RB1 to induce tumour suppression (Laurie *et al.*, 2006); inactivating

© 2020 International Statistical Institute.

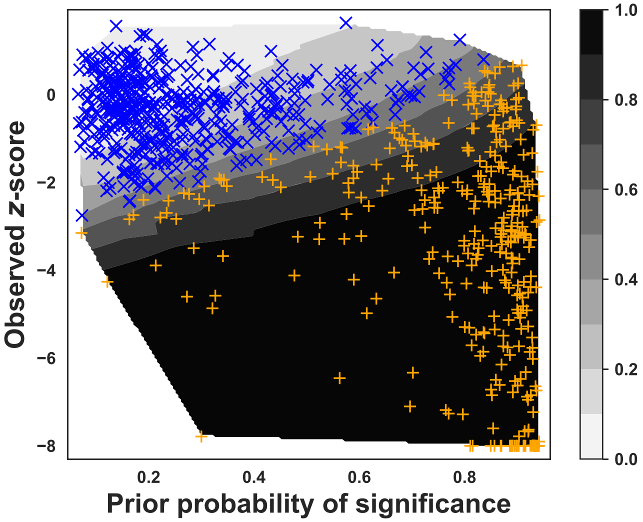


Figure 6. *Visualisation of the Stage 1 data for the Nutlin-3 dataset. The x-axis is the prior $P(h_i = 1|x_i)$; the y-axis is the observed z-score; the background gradient is the posterior $P(h_i = 1|x_i, z_i)$. Orange crosses are selected discoveries; blue x's are data points not selected. The strong prior information leads to some points being selected even with z-scores greater than zero (upper right corner). [Colour figure can be viewed at* wileyonlinelibrary.com*]*

Table 1. *Significant molecular features identified by DEBT that are predictive of sensitivity or resistance to Nutlin-3.*

Feature type	Selected by DEBT & KO	Selected by DEBT only
Mutation	TP53	CDKN2A, RB1
Copy loss	RB1	
Copy gain		CYLD
Low expression	JAK2, VHL, CDKN2A	FBXW7
High expression	CCND2, CCND3, MDM2, CDKN2A, MLLT3, SMARCB1	PTEN

DEBT, double empirical Bayes testing.

mutations in RB1 would remove this Nutlin-3 response pathway. CDKN2A mutations have both been shown to MDM2-based inactivation of P53 (Walter *et al.*, 2015). Both PTEN and CYLD are tumor suppressor genes that have been shown to directly regulate P53 (Puszynski *et al.*, 2014; Fernández-Majada *et al.*, 2016). FBXW7 cooperates with PTEN in suppression (Mao *et al.*, 2008), but its direct link to Nutlin-3 response has not been established, making this a potential new discovery.

7.2 All Drugs in the GDSC Dataset

We replicate the case study above for 214 drugs in the GDSC dataset. We use the same settings as in the simulation study in Section 6.

Figure 7 shows the results of both Stages 1 and 2 across the entire GDSC. As in the simulation results, DEBT discovers substantially more Stage 1 discoveries than BH. Also similar to the simulations, DEBT experiences the same small-scale discovery boost in Stage 2, indicated by the offset of black curve in the right panel of Figure 7. Unlike in the simulations, knockoffs appear to have a heavier tail, suggesting it discovers more features than DEBT in Stage 2. The

© 2020 International Statistical Institute.

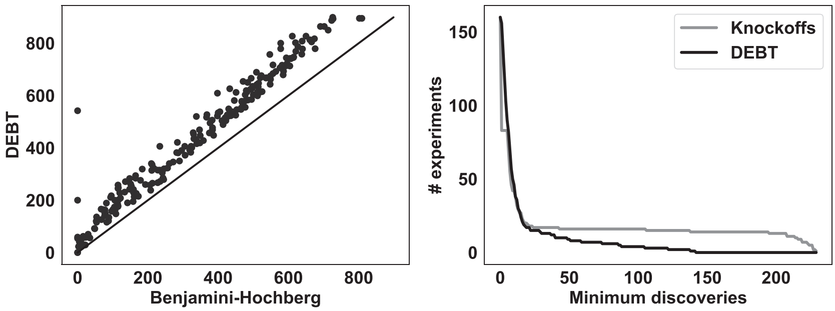
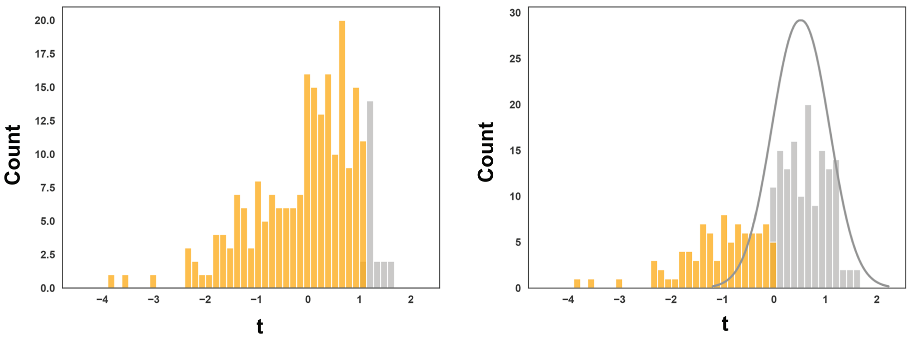


Figure 7. *Power comparisons for DEBT on the Genomics of Drug Sensitivity in Cancer dataset. Results mirror those of Figure 3 but show model-X knockoffs with a heavier tail of discoveries. DEBT, double empirical Bayes testing.*



(a) Knockoffs

(b) DEBT

Figure 8. *An example where the empirical null technique in DEBT prevents inflated false discoveries. Gray bars indicate unselected features; orange bars indicate selected features. The knockoffs approach selects an implausibly-large number of features because it assumes the null test statistics are well-calibrated and symmetric around the origin. The empirical null in DEBT (gray line) is centered around the central peak, which is shifted to the right of the origin. DEBT, double empirical Bayes testing. [Colour figure can be viewed at* wileyonlinelibrary.com*]*

knockoff results seem implausible, however, as it is unlikely that there are truly hundreds of driver features in many of the experiments.

A closer look at the results provides a resolution to this issue: a failure of the theoretical null. Figure 8 shows the Stage 2 results for a single drug (Camptothecin) where the knockoffs approach selects a large number of features based on a null distribution assumed to be symmetric about the origin (left panel). The DEBT empirical null (right panel, grey line) recalibrates the null distribution to be shifted to the right, centred at roughly 0.5. Consequently, DEBT selects a much smaller—but more biologically plausible—set of features than knockoffs.

The failure of the theoretical null is due to insufficient factor modelling. In practice, such factor modelling will always fail to completely capture the true null distribution of the target features. The empirical null adds a layer of robustness to DEBT that protects against an imperfect null model. The trade-off is a more conservative estimate of the null distribution. We consider this worth the cost, as the theoretical null is too brittle to be trusted in real data applications.

© 2020 International Statistical Institute.

8 Conclusion

This article proposed DEBT, a method that increases statistical power in multiexperiment scientific studies when side information is available for each experiment. DEBT leverages empirical Bayes testing techniques to boost power without sacrificing interpretability. In a case study on high-throughput drug screening, DEBT finds more experimental discoveries than conventional testing approaches and provides robust scientific insight into the mechanisms associated with differential treatment response.

The investigation here was the first for the DEBT method. A number of open questions remain:

- **How should one choose the predictive prior model?** DEBT uses a deep neural network, but the best model will be application-specific.
- **What are the theoretical properties of Stage 2?** We have provided theory that is centred on Stage 1, but an investigation of Stage 2 would be useful to better understand the asymptotic properties of DEBT.
- **Can the factor model be calibrated empirically?** Choosing the factor model component counts was chosen arbitrarily here at a reasonably large number. Principled choice of the number of latent factors and the specific latent factor model is an open problem.
- **Can Stage 2 be extended to CRTs?** Using CRTs instead of knockoffs in Stage 2 could potentially boost power. However, the computational cost of CRTs may prohibitive without further modifications to the approach. Adapting CRTs for Stage 2 inference is left for future work.

Appendix A: Optimisation Details

The hierarchical prior enables more stable optimisation. This is partially due to low-level details of stochastic gradient descent. With the flat prior, the output of the neural network at the last layer must be mapped through a logistic transform. This can lead to saturated or exploding gradients, resulting in poor convergence. The output of the neural network in the hierarchical model is a two-vector that is passed through an elementwise soft-plus transform,

$$\text{Softplus}(x) = \log(1 + \exp(x)).$$

This function is more numerically stable as it is numerically identical to a linear function for values of x above around 10.

We divide the data into K folds and learn for a different model for each fold. Fold-specific models are trained on all other data, with 10% of the training data used as a validation set; fold data are used as a holdout dataset for prediction. After every epoch, we evaluate the model on the validation set and use the best-performing model across all epochs to predict on the test set.

When the model is well-fit, the SGD gradients are small and make effectively random minute perturbations to the model. By random chance, then it is likely that a slightly-overfit model will perform slightly better on the validation set, leading to overestimation of confidence on the holdout set and a violation of the FDR threshold. This phenomenon is known as Freedman's paradox (Freedman, 1983). The regularisation term in (13) is added to correct for this concern that the model will overfit due to the training evaluation procedure.

Appendix B: Proof of Proposition 1

Proof. DEBT-1 rejects hypothesis by controlling the Bayesian local false discovery rate (BFLDR). To establish its FDR control, we show that the expected posterior probabilities of the rejected hypotheses coincides with their FDR $\text{P}(H_i = 0|z_i$ is rejected).

© 2020 International Statistical Institute.

In more detail, DEBT-1 solves the following optimisation problem when it learns the true prior probabilities and null/alternative distributions:

$$\begin{aligned} \max \quad & |Z_1| \\ s.t. \quad & bfldr(z) = |Z_1|^{-1} \sum_{i:z_i \in Z_1} (1 - w(z_i)) \leq \alpha, \end{aligned}$$

where

$$w(z_i) = \frac{c_i f_1(z_i)}{c_i f_1(z_i) + (1 - c_i) f_0(z_i)}.$$

Denote the marginal distribution of z_i as $f(z_i) = c_i f_1(z_i) + (1 - c_i) f_0(z_i)$ and the rejection region as R_w. (We note that $(f_0, f_1) = (f_0^*, f_1^*)$ and $c_i = c_i^*, i = 1, \ldots, n$ per the assumptions of Proposition 1.) Below, we show that expected posterior probabilities $1 - w(z_i)$ is equal to FDR(z_i) when hypothesis i is rejected.

$$\begin{aligned} \mathbb{E}((1 - w_i)|z_i \in Z_1) = & \frac{\int_{w(z_i) \in R_w} f(z_i)(1 - w(z_i)) dz_i}{\int_{w(z_i) \in R_w} f(z_i) dz_i} \\ = & \frac{\int_{w(z_i) \in R_w} (1 - c_i) f_0(z_i) dz_i}{\int_{w(z_i) \in R_w} f(z_i) dz_i} \\ = & \frac{P(\text{is rejected and} H_i = 0)}{P(z_i \text{is rejected})} \\ = & \text{FDR}(z_i). \end{aligned}$$

This calculation implies that the DEBT-1 constraint $|Z_1|^{-1} \sum_{i:z_i \in Z_1} (1 - w(z_i)) \leq \alpha$ controls the average FDR,

$$|Z_1|^{-1} \sum_{i:z_i \in Z_1} FDR(z_i) = \mathbb{E}(|Z_1|^{-1} \sum_{i:z_i \in Z_1} (1 - w_i)) \leq \alpha. \tag{B1}$$

That is, the mean false discovery rate of the rejected hypotheses is α. In other words, DEBT-1 is tight in the FDR control.

Appendix C: Proof of Proposition 2

Proof. To analyse the power of DEBT-1, we draw a connection between DEBT-1 and rejecting hypotheses based on the local FDR surface. We show that any solution to DEBT-1 can be generated by rejecting hypotheses based on the local FDR surface. Because DEBT-1 is tight in FDR control and the local FDR surface leads to the most powerful rejection rule (Lei & Fithian, 2018), DEBT-1 also maximises the power.

More specifically, we first show that any solution $|Z_1|$ to DEBT-1 can be achieved by rejecting using the local FDR surface:

$$\text{local FDR } = 1 - w(z_i) \leq \alpha'$$

for some α'. Given a set of hypotheses Z_1 that achieves the optimal $|Z_1|$, consider the $|Z_1|$ hypotheses with the smallest local FDR:

$$Z_1' := \{i : |\{i' : w(z_{i}') \geq w(z_i)\}| \leq |Z_1|\},$$

© 2020 International Statistical Institute.

where $|\{i' : w(zi') \geq w(z_i)\}|$ counts the number of hypotheses with a smaller or equal local FDR. The set of rejected hypotheses $Z1'$ must also be a solution to the DEBT-1 optimisation because

$$|Z_1'|^{-1} \sum_{i:z_i \in Z_1'} (1 - w(z_i)) \leq |Z|^{-1} \sum_{i:z_i \in Z} (1 - w(z_i)) \qquad \forall Z \text{ s.t. } |Z| = |Z_1'|$$

and $|Z_1'| = |Z_1|$. This way, we can turn any DEBT-1 solution Z_1 to an equally good one $Z1'$ generated by rejecting using the local FDR surface.

This equivalence between DEBT-1 and rejecting using the local FDR surface implies that DEBT-1 maximises the power. The reason is that the most powerful rejection threshold for FDR control is local FDR surfaces (theorem 2 of Lei and Fithian, 2018). Together with Proposition 1 that shows DEBT-1 is tight in FDR control, it implies that DEBT-1 maximises power under FDR control at α.

Finally, we note that DEBT-1 being tight in FDR control is essential for this argument. A method with FDR control at α' strictly smaller than α would not maximises the power even if it uses the most powerful rejection rule. The reason is the same method with a relaxed FDR control to reach α would be more powerful.

Notes

[1] This article is an expansion of our earlier work(Tansey *et al.*, 2018). This article expands on the ideas, scales up the second stage and provides new theory about the first stage.

[2] Code is available at https://github.com/tansey/debt.

[3] We will use capital letters to denote random variables, with feature indexing (e.g. X_j).

References

Barber, R.F. & Candès, E.J. (2015). Controlling the false discovery rate via knockoffs. *Annals Stat.*, **43**(5), 2055–2085.

Bates, S., Sesia, M., Sabatti, C. & Candes, E. (2020). Causal inference in genetic trio studies. arXiv preprint arXiv:2002.09644.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J.Royal Stat. Society: Ser. B (Stat. Methodology)*, 289–300.

Candes, E., Fan, Y., Janson, L. & Lv, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J. Royal Stat. Society: Ser. B (Stat. Methodology)*.

Efron, B. (2003). Robbins, empirical Bayes and microarrays. *Annals Stat.*, **31**(2), 366–378.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. American Stat. Association*, **99**(465), 96–104.

Efron, B. (2005). Local false discovery rates. Available at: http://statweb.stanford.edu/~ckirby/brad/papers/2005LocalFDR.pdf

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Stat. Sci.*, 1–22.

Efron, B. (2019). Bayes, oracle Bayes and empirical Bayes. *Stat. Sci.*, **34**(2), 177–201.

Efron, B. & Tibshirani, R. (1996). Using specially designed exponential families for density estimation. *Annals Stat.*, **24**(6), 2431–2461.

Efron, B., Tibshirani, R., Storey, J.D. & Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Stat. Assoc.*, **96**(456), 1151–1160.

Fernández-Majada, V., Welz, P.-S., Ermolaeva, M.A., Schell, M., Adam, A., Dietlein, F., Komander, D., Büttner, R., Thomas, R.K. & Schumacher, B. (2016). The tumour suppressor CYLD regulates the p53 DNA damage response. *Nat. Commun.*, **7**(1), 1–14.

Freedman, D.A. (1983). A note on screening regression equations. *Amer. Stat.*, **37**(2), 152–155.

Garnett, M.J., Edelman, E.J., Heidorn, S.J. & other (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, **483**(7391), 570.

Iorio, F., Knijnenburg, T.A., Vis, D.J., Bignell, G.R., Menden, M.P., Schubert, M., Aben, N., Gonçalves, E., Barthorpe, S. & Lightfoot, H. (2016). A landscape of pharmacogenomic interactions in cancer. *Cell*, **166**(3), 740–754.

Katsevich, E. & Ramdas, A. (2020a). The leave-one-covariate-out conditional randomization test. *arXiv preprint arXiv:2006.08482*.

Katsevich, E. & Ramdas, A. (2020b). A theoretical treatment of conditional independence testing under Model-X. *arXiv preprint arXiv:2005.05506*.

Laurie, N.A., Donovan, S.L., Shih, C.-S. & other (2006). Inactivation of the p53 pathway in retinoblastoma. *Nature*, **444**(7115), 61–66.

Lei, L. & Fithian, W. (2018). AdaPT: an interactive procedure for multiple testing with side information. *J. Royal Stat. Society*.

Li, A. & Barber, R.F. (2017). Accumulation tests for FDR control in ordered hypothesis testing. *J. Amer. Stat. Assoc.*, **112**(518), 837–849.

Liu, M., Katsevich, E., Janson, L. & Ramdas, A. (2020). Fast and powerful conditional randomization testing via distillation. *arXiv preprint arXiv:2006.08482*.

Liu, Y. & Zheng, C. (2018). Auto-encoding knockoff generator for FDR controlled variable selection. *arXiv preprint arXiv:1809.10765*.

Mao, J.-H., Kim, I.-J., Wu, D. & other (2008). FBXW7 targets mTOR for degradation and cooperates with PTEN in tumor suppression. *Science*, **321**(5895), 1499–1502.

Mazoure, B., Nadon, R. & Makarenkov, V. (2017). Identification and correction of spatial bias are essential for obtaining quality data in high-throughput screening technologies. *Sci. Reports*, **7**(1), 11921.

Newton, M.A. (2002). A nonparametric recursive estimator of the mixing distribution. *Sankhya Ser. A*, **64**, 306–22.

Puszynski, K., Gandolfi, A. & d'Onofrio, A. (2014). The pharmacodynamics of the p53-MDM2 targeting drug Nutlin: the role of gene-switching noise. *PLoS Comput. Biol.*, **10**(12), e1003991.

Ramdas, A., Barber, R.F., Wainwright, M.J. & Jordan, M.I. (2017). A unified treatment of multiple testing with prior knowledge. *arXiv preprint arXiv:1703.06222*.

Scott, J.G., Kelly, R.C., Smith, M.A., Zhou, P. & Kass, R.E. (2015). False discovery rate regression: an application to neural synchrony detection in primary visual cortex. *J. Amer. Stat. Assoc.*, **110**(510), 459–471.

Tansey, W., Koyejo, O., Poldrack, R.A. & Scott, J.G. (2017). False discovery rate smoothing. *J. Amer. Stat. Assoc.*

Tansey, W., Veitch, V., Zhang, H., Rabadan, R. & Blei, D.M. (2018). The holdout randomization test: principled and easy black box feature selection. *arXiv preprint arXiv:1811.00645*.

Tansey, W., Wang, Y., Blei, D. & Rabadan, R. (2018). Black box FDR. In *International conference on machine learning*, pp. 4874–4883.

Tokdar, S., Martin, R. & Ghosh, J.K. (2009). Consistency of a recursive estimate of mixing distributions. *Annals. Stat.*, **37**(5A), 2502–22.

Walter, R.F.H., Mairinger, F.D., Ting, S. et al. (2015). MDM2 is an important prognostic and predictive factor for platin–pemetrexed therapy in malignant pleural mesotheliomas and deregulation of P14/ARF (encoded by CDKN2A) seems to contribute to an MDM2-driven inactivation of P53. *British J. Cancer*, **112**(5), 883–890.

Weinberg, R. (2013). *The Biology of Cancer*. Garland science.

Wu, Y., Boos, D.D. & Stefanski, L.A. (2007). Controlling variable selection by the addition of pseudovariables. *J. Amer. Stat. Assoc.*, **102**(477), 235–243.

Xia, F., Zhang, M.J., Zou, J.Y. & Tse, D. (2017). NeuralFDR: learning discovery thresholds from hypothesis features. In *Advances in neural information processing systems*, pp. 1540–1549.

Yang, W., Soares, J., Greninger, P. & other (2012). Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.*, **41**(D1), D955–D961.

[*Received August 2020, Revised October 2020, Accepted October 2020*]

© 2020 International Statistical Institute.