Accurate error control in high dimensional association testing using conditional false discovery rates

James Liley and Chris Wallace (ajl88@cam.ac.uk, cew54@cam.ac.uk)

August 24, 2019

Abstract

High-dimensional hypothesis testing is ubiquitous in the biomedical sciences, and informative covariates may be employed to improve power. The conditional false discovery rate (cFDR) is widely-used approach suited to the setting where the covariate is a set of p-values for the equivalent hypotheses for a second trait. Although related to the Benjamini-Hochberg procedure, it does not permit any easy control of type-1 error rate, and existing methods are over-conservative. We propose a new method for type-1 error rate control based on identifying mappings from the unit square to the unit interval defined by the estimated cFDR, and splitting observations so that each map is independent of the observations it is used to test. We also propose an adjustment to the existing cFDR estimator which further improves power. We show by simulation that the new method more than doubles potential improvement in power over unconditional analyses compared to existing methods. We demonstrate our method on transcriptome-wide association studies, and show that the method can be used in an iterative way, enabling the use of multiple covariates successively. Our methods substantially improve the power and applicability of cFDR analysis.

1 Introduction

In the 'omics' approach to biology, a large number n of descriptive variables are considered in the analysis of a biological system, intended to provide a near-exhaustive characterisation of the system under consideration. Typically only a small proportion of the investigated variables are associated with the behaviour of the system, and we seek to identify this subset of variables, along with the magnitude and direction of their associated effect sizes. A first step is generally to test each hypothesis in a frequentist framework, generating a corresponding set of p-values. Often, additional information is available in the form of an external covariate, which assigns a numerical value to each hypothesis which has different (unknown) distributions amongst associations and non-associations. Information from such covariates can be incorporated into hypothesis testing to improve power in detecting associations.

A range of procedures have been proposed for this type of analysis. An important consideration is the form of the (two-dimensional) rejection rule applied to the p-value-covariate pairs. An optimal procedure (in terms of minimising type 2 error and controlling type 1 error) determines rejection regions on the basis of a ratio of bivariate probability densities (PDFs) of the p-value and covariate under the null and under the alternative. One approach to the problem at hand is to estimate this ratio directly [1, 2, 3]. Other approaches include 'filtering' on covariate values [4], weighting hypotheses according to the value of the covariate [5, 6, 7, 8, 9], modulating a univariate test of p-values in response to the covariate in some other way [10, 11, 12], and binning covariates in order to treat each bin separately [13]. Since covariates can be of many types (continuous, categorical; univariate, multivariate; known or unknown distributional properties) and can relate to the p-values in a range of ways, this array of methods is necessary to manage the range of problem types.

The conditional false discovery rate (cFDR) circumvents the difficulties of estimating PDFs by approximating the optimal ratio using cumulative density functions (CDFs) [14]. In this case, the covariate is generally a set of p-values arising from an analogous procedure on the same variables for a second 'conditional' trait with an unknown degree of similarity to the trait giving rise to the primary set of p-values (which we call the 'principal' trait). The method has been extensively used in genomics [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38]. Formally, the cFDR is a posterior probability of association with the principal trait given that p-values for the principal and conditional traits fall below p-value thresholds p, q respectively. It is readily estimated using empirical CDFs (ECDFs) [14].

The cFDR is a useful Bayesian quantity in its own right. Generally, the cFDR is used in effectively a frequentist way: roughly, for each observed p-value pair (p_i, q_i), we estimate the cFDR at $(p, q) = (p_i, q_i)$ and reject the null hypothesis if this estimated value is less than some threshold α. This process is nearly analogous to the Benjamini-Hochberg procedure (B-H) [39] on a single set of of p-values p_i, but unlike BH, it does not control the FDR at α (nor any other conventional measure of type-1 error rate). In a previous paper [23] we proposed a rough method to approximately control FDR in this setting, but our method was drastically conservative.

The main contribution of this paper is to propose a much improved type-1 error rate control strategy for cFDR, which improves power relative to previous methods. In four secondary contributions, we a) propose an improvement to the existing estimator which improves power, b) show several asymptotic results about the method and demonstrate that the effect of certain troublesome properties is small, c) enable and demonstrate iterative use of the procedure, and d) compare the general cFDR method with PDF-based, parametric and kernel density estimator (KDE)-based approaches. An R package is provided.

In this paper, we begin by describing a motivating example using transcriptome-wide association studies (TWAS). We then summarise the cFDR and its estimator, and describe its relation to the B-H procedure. We then describe our method to transform cFDR

estimates into p-value-like quantities, and discuss how the cFDR approach relates to similar methods in the field. We evaluate the type-1 error rate control and power of the method, and finally describe an iterated form of the procedure for use with multiple sets of covariates.

1.1 Motivating example

We consider a transcript-wide association study (TWAS) [40] of breast cancer BRCA [41]) and ovarian cancer (OCA [42]), which are epidemiologically and biologically similar diseases [43].

The TWAS aims to identify which transcripts (gene products) in which tissues differ between cases (BRCA or OCA) and controls, and generates a set of $\approx 10^5$ p-values p_{BRCA}, p_{OCA} from association tests of the same sets of tissue/transcript pairs with BRCA and OCA respectively (further detail is given in supplementary material section 9.1). We will assume that we have no prior knowledge that any variables are more likely to be BRCA- or OCA-associated, that absolute Z-scores $z_{BRCA} = \Phi^{-1}(p_{BRCA}/2)$, $z_{OCA} = -\Phi^{-1}(p_{OCA}/2)$ have a block-diagonal correlation structure where block locations are known, and that under a null hypothesis H^0_{BRCA} of no association with BRCA, z_{BRCA} and z_{OCA} are independent. We wish to find which of the variables are associated with BRCA, and thus investigate a null hypothesis H^0_{BRCA} of non-association.

A straightforward approach is to apply the Benjamini-Hochberg (B-H) procedure to the values p_{BRCA} (figure 1, panel A). BRCA and OCA tend to have associations at the same variables, suggesting that a rejection region should reflect this to improve power. A natural way to do this is to only consider those variables for which z_{OCA} exceeds some threshold, which allows rejection of H^0_{BRCA} at a looser z_{BRCA} threshold (figure 1, panel B). FDR control is maintained under the independence assumption above [4]. However, this procedure is problematic: a threshold on z_{OCA} must be chosen *a priori*, and variables with z_{OCA} falling below the red line have H^0_{BRCA} rejected automatically.

The cFDR procedure circumvents this problem (figure 1, panel C). The associated rejection region, which we term an L-region, 'adapts' to the joint distribution of z_{BRCA} and z_{OCA}. For small α, the L-region approximates an optimal rejection region (appendix 8.1). A major shortcoming is that although a B-H procedure with FDR controlled at α is repeatedly used to generate the L-region, the overall FDR is not controlled at α, nor any straightforward function of α ([23]; appendix 8.5).

In this paper, we demonstrate a straightforward and effective way to control the type-1 error rate (as FDR or otherwise) in the cFDR procedure. As α varies from 0 to 1, the leftmost boundary of the L-region 'sweeps' across the entire (+,+) quadrant, and for $\alpha_1 < \alpha_2$, we have $L(\alpha_1) \subseteq L(\alpha_2)$. Thus we can associate each point (x,y) in the (+,+) quadrant with the smallest L-region containing it, which will generally have (x,y) on its leftmost border. Loosely, we control FDR by estimating the probability that each point would lie within its associated region under H^0_{BRCA}. We term this the v-value, which has similar properties to a p-value and can be used in the B-H procedure.

Care must be taken when applying rejection rules to the same data on which those rules were determined, so we use a leave-one-out procedure which avoids this problem (section 3, appendix 8.3). We show that the rejection region generated by the cFDR approximates the best-possible rejection region (section 2.2, appendix 8.1), and that rejection regions converge reasonably fast as the number of variables under consideration increases (section 8.2). The rejection region is non-parametric, and we show that the cFDR method can outperform parametric methods (section 5.2). Finally, the v-values may be considered 'adjusted' p-values, which enables straightforward iteration of the method with further sets of p-values at the same variables, discussed in section 5.3.

2 Review of cFDR estimator

2.1 Definitions

Assume that we have results from n pairs of hypothesis tests against two series of null hypotheses $(H_0^p(i), H_0^q(i))$ in the form of a set S of bivariate p-values $S = (p_i, q_i)$, $i = 1..n$. In our motivating example, $H_0^p(i)$ and $H_0^q(i)$ are non-association of the ith tissue-gene pair with BRCA and OCA respectively. We consider $(H_0^p(i), H_0^q(i))$ to be realisations of independent Bernoulli random variables H_0^p, H_0^q satisfying $P(H_0^p) = \pi_0^p$, $P(H_0^q) = \pi_0^q$, and p_i, q_i to be IID realisations of random variables P, Q satisfying:

$$\begin{aligned} P|H_0^p &\sim U(0,1) \\ P &\perp\!\!\!\perp Q|H_0^p \end{aligned} \tag{1}$$

although assumption (1) can be relaxed. We denote

$$F_0(p,q) = P(P \leq p, Q \leq q|H_0^p) = pF_0^q(q) \tag{2}$$

$$F(p,q) = P(P \leq p, Q \leq q)$$

$$f_0(p,q) = f(P = p, Q = q|H_0^p) = f_0^q(q) \tag{3}$$

$$f(p,q) = f(P = p, Q = q)$$

where the separability of (2) and (3) is due to assumption (1).

2.2 Optimal procedure

Under H_0^p, the probability of a random instance of (P, Q) falling in a region R is $\int_R f_0(p,q)dpdq$. To find an ideal two-dimensional rejection region for hypothesis testing, we wish to fix this value at a level α while maximising the probability $\int_R f(p,q)dpdq$. This optimal region (or one such optimal region) is given by the set of points $\{(p,q) : f_0(p,q)/f(p,q) \geq k_\alpha\}$ for some k_α (a formal statement and proof are given in appendix 8.1, and this is also shown in various forms in [1, 2, 3]). In equivalent terms, an optimal decision rule for the set S would rank p-value pairs according to $f_0(p_i, q_i)/f(p_i, q_i)$ or equivalently $P(H_0^p|P = p_i, Q = q_i)$.

A natural approach is to estimate f_0 and f using a parametric approximation [1, 3] or local approximations using kernel density estimates (KDEs) [2] or spline models [44]. However, PDFs are difficult to estimate in general, and there may be little reason to believe parametric assumptions are satisfied; in our motivating example (figure 1, panels A-C) there is little reason to think that a smooth rejection region would be optimal.

2.3 Conditional false discovery rate

The conditional false-discovery rate [14] takes an alternative approach of instead ranking points by an estimate of $F_0(p,q)/F(p,q)$. This estimate is obtained by estimating the monotonically-related quantity:

$$cFDR(p,q) = P(H_0^p | P \leq p, Q \leq q) \tag{4}$$

$$= \frac{P(P \leq p | H_0^p, Q \leq q)}{P(P \leq p | Q \leq q)} P(H_0^p | Q \leq q) \tag{5}$$

Suppose we have a multiset X of p-value pairs (p_i, q_i). If almost all these pairs are iid realisations (p_i, q_i) of (P, Q), then for fixed p, q, the empirical CDFs

$$\frac{1}{|X|} \left|\{i : (p_i, q_i) \in X, p_i \leq p, q_i \leq q\}\right|$$

$$\frac{1}{|X|} \left|\{i : (p_i, q_i) \in X, q_i \leq q\}\right|$$

are consistent estimators of $P(P \leq p, Q \leq q)$, $P(Q \leq q)$ respectively. Given assumption (1), we have $p = P(P \leq p | H_0^p) = P(P \leq p | Q \leq q, H_0^p)$ and (for the moment) we may conservatively approximate $P(H_0^p | Q \leq q) = 1$. Given X, we thus define the estimated cFDR (denoted $\widehat{cFDR}$), as a function of two variables $(p, q) \in (0, 1)$:

$$\widehat{cFDR}_X(p,q) = p \frac{\max(\left|\{i : q_i \leq q, (p_i, q_i) \in X\}\right|, 1)}{\max(\left|\{i : p_i \leq p, q_i \leq q, (p_i, q_i) \in X\}\right|), 1)} \tag{6}$$

For fixed p, q, $\widehat{cFDR}_X(p,q)$ is a generally-biased but consistent estimator of $cFDR(p,q)/P(H_0^p | Q \leq q)$, which converges uniformly on fixed regions at a rate of $O(n^{-1/2})$ (see appendix 8.2), and it is usually a downwards-biased (conservative) estimator of $cFDR(p,q)$.

Approximating $P(H_0^p | Q = q) = 1$ in equation (6) disregards any variation on $P(H_0^p | Q = q) = 1$ with q, so we introduce at this stage an estimate of $P(H_0^p | Q \leq q)$, which we can multiply with $\widehat{cFDR}_X(p,q)$ to improve the accuracy of approximation of $cFDR(p,q)$. Our

estimate is

$$\begin{aligned} P(H_0^p|Q \leq q) &= P(H_0^p)\frac{P(Q \leq q|H_0^p)}{P(Q \leq q)} \\ &\approx \pi_0 \frac{P(Q \leq q|P > 1/2)}{P(Q \leq q)} \\ &\approx \frac{\min(1, |\{i : (p_i, q_i) \in X, q_i \leq q, p_i > 1/2\}|)}{\min(1, |\{i : (p_i, q_i) \in X, q_i \leq q\}||\{i : (p_i, q_i) \in X, p_i > 1/2\}|)} \\ &= \widehat{Pr}_X(H_0^p|Q \leq q) \end{aligned} \tag{7}$$

where we approximate $\pi_0 = P(H_0^P) = 1$. We denote

$$c\widehat{FDR}_X^n(p, q) = c\widehat{FDR}_X(p, q)\widehat{Pr}_X(H_0^p|Q \leq q) \tag{8}$$

Estimating π_0 (rather than setting $\pi_0 = 1$) would uniformly scale all estimates of $cFDR(p, q)$, which has no effect on our rejection procedure.

In the hypothesis testing setting, we aim to use $c\widehat{FDR}$ or $c\widehat{FDR}^n$ to construct a decision rule on our set S of observed p-value pairs (we will forego the n superscript from now, with the understanding that it may be added). A simple approach is to reject $H_0^p(i)$ if $c\widehat{FDR}_S(p_i, q_i) \leq \alpha$, but $c\widehat{FDR}_S(p, q)$ is not monotonically increasing with p and we do not want to reject H_0^p for some (p_i, q_i) and not for some other pair (p_j, q_j) with $q_i = q_j$ but $p_j < p_i$. Hence, we can use the decision rule (as per [14])

$$\text{Reject } H_0^p \text{ if: } \exists\, p' \geq p_i : c\widehat{FDR}_S(p', q_i) \leq \alpha \tag{9}$$

This enables a rejection region with a single rightmost boundary, as shown in panels D in figure 1. It closely parallels the B-H [39] procedure (B-H) on a set of p-values. Suppose we have a set S_1 of p-values $p_1, p_2, ..., p_n$, and define

$$BH_{S_1}(p) = p\frac{|S_1|}{\max(1, |\{i : p_i \leq p, p_i \in S_1\}|)}$$

Then the B-H procedure can be written as

$$\text{Reject } H_0^p \text{ if: } \exists\, p' \geq p_i : BH_{S_1}(p') \leq \alpha \tag{10}$$

The B-H procedure controls FDR at α, and if it is performed with S_1 as the subset of S for which $q_i \leq \gamma$ (where γ is a threshold chosen independently of S), the FDR will still be controlled at α if assumption (1) is satisfied [4, 5]. The rejection procedure (9) is equivalent to repeatedly performing this 'thresholded' B-H at $\gamma = q_i$, and using that decision rule for point (p_i, q_i) (panel C, figure 1).

When procedure (9) is used, the FDR is no longer controlled at α, and indeed can exceed α by an arbitrary proportion (appendix 8.5). Previous work using the cFDR generally

interprets it in a Bayesian context, without requiring a bound on FDR or FWER. In [23] we introduced a method to choose an α^* such that rejection criterion (9) roughly controlled the FDR at α, but that method was overly conservative, generally controlling FDR at a far lower level than needed.

3 Map from p-value pairs to v-values

We identify a 'rejection region' associated with $c\widehat{FDR}$ by adding a 'test point' (p, q) to a set of points X and considering the region for which a hypothesis corresponding to (p, q) is rejected under (9) with $S = X + (p, q)$.

The function $c\widehat{FDR}_{X+(p,q)}(p, q)$ is now defined on the unit square. We now define the 'L-region' $L_X(\alpha)$:

$$L_X(\alpha) = \{(p, q) : \exists p' \geq p : c\widehat{FDR}_{X+(p',q)}(p', q) \leq \alpha\} \tag{11}$$

and the 'L-curve' as it's rightmost border. Under minor assumptions on f and f_0, the L-curve 'converges' to a contour of $cFDR(p, q)P(H_0^p|Q \leq q)$ (see appendix 8.2).

For $\alpha \leq \beta$, $L_X(\alpha) \subseteq L_X(\beta)$, so the regions are nested. Define c_{pq} so that $L_X(c_{pq})$ s the smallest such region containing (p, q) (so $c_{pq} \approx c\widehat{FDR}(p, q)$). Given the PDF f_0 of $P, Q|H_0^p$, the probability under H_0^p of a newly-sampled realisation (p, q) of P, Q falling in $L_X(c_{pq})$ is readily computable by integrating f_0 over this region:

$$v_X(p, q) = P(P, Q \in L_X(c_{pq})|H_0^p) = \int_{L_X(c_{pq})} f_0(p, q)dp\, dq \tag{12}$$

The PDF f_0 can be readily estimated given assumption (1) (as will be shown in the subsequent section). The value $v_X(p, q)$ can be roughly interpreted as a p-value against H_0^p using $c\widehat{FDR}(p, q)$ as a test statistic in that

$$P(v_X(p, q) \leq \alpha) \leq \alpha \tag{13}$$

where the '$\leq \alpha$' rather than '$= \alpha$' is because (p, q) may be on the interior, rather than border, of $L_X(c_{pq})$.

However, although 13 holds for p, q newly sampled from f_0, it does not hold for the (Lesbegue-measure-zero) set of points $(p_i, q_i) \in X$, since points in X are used to define the function v_X. In order for (13) to hold, the point (p, q) must be independent of the definition of v_X. This is easily managed: to test (p_i, q_i), we simply leave (p_i, q_i) out of the points used to define the map we use on (p_i, q_i) itself. That is, given our set S of datapoints as above, we define the 'leave-one-out' v-value

$$v(p_i, q_i) = v_{S-(p_i,q_i)}(p_i, q_i) \tag{14}$$

The problem can also be managed by leaving out blocks of points; for a partition of $1..n$ into blocks $1, 2, ..., k$, supposing i is in block $b(i)$, the 'block-out' v-value is defined as

$$v(p_i, q_i) = v_{S-\text{block } b(i)}(p_i, q_i) \tag{15}$$

If observations (p_i, q_i) are not independent but have a block-diagonal correlation structure, then this procedure is necessary in order to ensure property (13) holds for $(p, q) = (p_i, q_i)$: since each observation (p_i, q_i) carries information about the other p-value pairs it is correlated with, removing it will not remove the influence of point (p_i, q_i) on the map. In this case, blocks should be chosen so that p-value pairs are independent between blocks, but possibly dependent within blocks. Such structure arises often in -omics experiments; in genetics, independence of allele counts may be assumed between chromosomes, but not generally within.

As $n \to \infty$, v-values based on $c\widehat{FDR}$ converge to the values they would obtain if L-regions were constructed with $c\widehat{FDR}(p, q) = cFDR(p, q)/P(H_0^P|Q \leq q)$. The variance of the maximum error across all v-values is $O(n^{-1/2})$ (see appendix 8.2). We note that consistency of L-region estimation is not necessary for type-1 error rate control, and that the rejection procedure is identical under monotonic transformations of c_{pq}. Consequently, type 1 error rate control is maintained for $c\widehat{FDR}^n$, even though the estimate is not necessarily conservative, and the approximation $P(H_0^P) \approx 1$ in equation 7 does not affect the method.

3.1 Estimation of $P, Q|H_0^p$

Recalling equation (3), we may write $f_0(p, q) = f_0^q(q)$ To estimate f_0^q, we assume that $(Q|H_0^p) \sim (Q|P \geq 1/2)$, and approximate the latter with a mixture-Gaussian distribution

$$-\Phi^{-1}\left(\frac{Q}{2}\right)\Bigg|P > \frac{1}{2} \sim \begin{cases} |N(0,1)| & \text{prob} = \pi_0 \\ |N(0,\sigma_0^2)| & \text{prob} = 1 - \pi_0 \end{cases} \tag{16}$$

where $N(\mu, \sigma)$ is the normal distribution with mean μ and variance σ^2. Estimates $\widehat{\pi_0}$, $\widehat{\sigma_0}$ of π_0 and σ_0 can be readily made using an expectation-maximisation algorithm [45], using the values q_i for which the corresponding p_i is $\geq 1/2$. We then set

$$\widehat{f_0}(p, q) = \mathbf{1}f_0^q(q) = \widehat{\pi_0} + (1 - \widehat{\pi_0})\phi\left(\frac{\Phi^{-1}(q/2)}{\widehat{\sigma_0}}\right) \tag{17}$$

If P, Q have a known dependence under H_0^p, an alternative distribution can be used for computing $v(L)$ (see supplementary material, section 9.2). The PDF f_0^q could be estimated in other ways; for example, a kernel density estimate [46].

3.2 Correlation between v-values

Decision rules based on multiple p-values generally require adjustment if p-values are dependent (eg [39]). If v-values are obtained by the leave-one-out procedure (14) they are slightly

pairwise dependent. The dependence is small; if $X' = X - (p_i, q_i) - (p_j, q_j)$ the values $v_{X'}(p_i, q_i)$, $v_{X'}(p_j, q_j)$ are independent, so the pairwise dependence between v-values corresponding to (p_i, q_i), (p_j/q_j) only arises from the differences $v_{X'+(p_j,q_j)}(p_i, q_i) - v_{X'}(p_i, q_i)$, $v_{X'+(p_i,q_i)}(p_j, q_j) - v_{X'}(p_j, q_j)$; that is, the effect of a single point ((p_j, q_j), (p_i, q_i) respectively) on the map $v_{X'}$ defined by $|X'| = |X| - 2$ points. The expected effect on values v_X from adding a single point to X is $O\left(\frac{1}{n^2}\right)$ where $n = |X|$ (see appendix 8.3). When v-values are defined using block-out as in (15), v-values are independent within-block but dependent between blocks.

3.3 Algorithm

We can now present our final algorithm.

Algorithm 1 Controlling type-1 error rate in cFDR

Input: 'principal' p-values $p_1, p_2, ...p_n$; 'conditional' p-values $q_1, q_2, ...q_n$; optionally fold assignment $b : 1..n \to 1..k$ such that $(p_i, q_i) \perp\!\!\!\perp (p_j, q_j) \big| b(i) = b(j)$

Output: v-values $v_1, v_2...v_n$

1: Identify the set $\{q_i : p_i > 1/2\}$ and make estimates $\widehat{\pi_0}$, $\widehat{\sigma_0}$ of π_0, σ_0 as per (16)
2: Set $\widehat{f_0}(p, q)$ as per equation (17)
3: **for** $i \in 1..n$ **do**
4: Set $S' = \{(p_j, q_j) : j \neq i\}$ (leave-one-out) or $S' = \{(p_j, q_j) : b(j) \neq b(i)\}$ (block-out)
5: Find $c_i = \arg\min\{c : (p_i, q_i) \in L_{S'}(c)\}$
6: Set $v_i = \int_{L_{S'}(c_i)} \widehat{f_0}(p, q) dp dq$
7: Return $(v_1, v_2, ..., v_i)$

We can interpret v_i as 'the probability that a randomly-chosen (p, q) pair has a more extreme $c\widehat{FDR}$ value than $c\widehat{FDR}(p_i, q_i)$'; that is, as a p-value. This allows straightforward FWER or FDR control, especially as v-values are almost independent. The v-values order hypotheses such that a rejection rule {reject $H_0^P(i)$ if $v_i \leq \alpha$} has near-optimal power, in terms of corresponding to near-optimal forms for rejection regions.

4 Relation to other methods

A wide range of approaches have been proposed for the problem of high-dimensional association testing using an informative covariate. Given the correspondingly wide variation in problems of this type, the optimal method is likely to depend on circumstance. In general, we will take P, p, H_0^p to refer to p-values and hypotheses for the trait of primary interest, and Q, q to refer to the covariate.

4.1 Determination of rejection region form

The simplest approach to covariate-based testing is 'independent filtering' [4] in which attention is restricted to the set $\{(p_i, q_i) : q_i \geq q_0\}$, with the B-H procedure then applied to the corresponding subset of values of p_i. This procedure is equivalent to rejection regions which are a series of rectangles with upper border at $q = q_0$. Independent filtering is clearly non-optimal, but is well-suited to some problem types [4].

As discussed above, a range of approaches aim to approximate the optimal rejection regions based on f_0/f. In [1] and [3], parametrisation leads to rejection regions constricted to a particular parametric class; in [1] that of oracle rejection regions under mixture-Gaussian forms of f_0 and f. In [2] and [44], boundaries of rejection regions are necessarily smooth at a scale corresponding to the smoothing kernel width, but can take otherwise arbitrary forms. An alternative approach is to 'bin' covariates [5, 13] which leads to L-curves which are step-functions with steps spaced according to the resolution of the bins.

An approach in [11] estimates $P(H_0^p|Q = q)$ for each q to modulate a B-H type test for each observation. The entire effect of the covariate in this method is encompassed through the value of $P(H_0^p|Q = q)$, which necessarily relies on point-estimates of the PDF $f(Q = q|H_0^p)$, and hence the method is dependent on the accuracy of this estimate.

Another common approach to covariate modulation is the weighted Benjamini-Hochberg procedure [7], in which each p-value p_i is reweighted to a value p_i/w_i (where $\sum w_i = 1$) and the standard B-H procedure is then applied to the values p_i/w_i. Our method can be interpreted in these terms, setting $w_i = v_i/p_i$, but this is rather unnatural; there is no clear way to interpret what the ratio v_i/p_i means, and this approach does not make use of the 'p-value property' in equation (13).

The use of empirical CDFs to generate rejection regions has the advantage of making use of the global distribution of P, Q, while spline- and kernel-density based estimates can generally only use local observations. The cFDR-based method has the obvious disadvantage of not converging to the optimum rejection region, and it can be less powerful than parametric approaches if parametric assumptions hold. However, using CDFs rather than PDFs allows faster convergence of rejection regions with n, and this favours the cFDR approach if n is small, particularly if CDF- and PDF- based regions are similar (see appendix 8.4) and PDFs are difficult to model well.

4.2 Censoring of points

In general terms, the process determining a decision rule to be used on observation (p_i, q_i) cannot easily make use of the datapoint (p_i, q_i) itself, since the use of the point biases the choice of decision rule in some way. Approaches by [1, 2] censor the points used in the decision rule to those already rejected in a stepwise approach, and a method in [3] masks the information available for the decision rule by effectively adding the point $(1 - p_i, q_i)$ to the dataset.

Since cFDR uses the entire dataset to estimate empirical CDFs, complex censoring can require that the cFDR estimator be changed in a non-trivial way. In particular, there is no obvious way to apply the methods proposed by [2] or [3]. We propose avoiding the problem by leaving out the point (p_i, q_i) directly (equations (14),(15)), at the cost of residual correlation in resultant v-values. While crude, this corresponds to a near-minimum censorship of points, and the resultant correlation tends to be small enough to ignore (see appendix 8.3).

4.3 Asymmetry and management of extreme outliers

An important property of the cFDR-based method is asymmetry, in that H_0^p cannot generally be rejected based on a low q_i alone (this can be seen by noting that $\widehat{cFDR}(p_i, q_i) \geq p_i$, and p_i can only exceed $\widehat{cFDR}^n$ in rare circumstances). Parametric approaches such as those in [1, 3] are not generally robust to this; for example, in [1], an extremely low q_i could lead to rejecting H_0^p even if the corresponding p_i were close to 1 and P, Q were independent (given that the degree of dependence is estimated). This property of the cFDR is very important when p_i and q_i are derived from GWAS on different diseases; it is entirely possible and even expected that a very strong association in the conditional trait is not an association with the principal trait. This property also differentiates our approach from meta-analysis of two sets of p-values.

4.4 Relation to original FDR-controlling method for cFDR

In a paper in 2015 [23], we identified the problem of failure of FDR control at α when using a rejection rule $\widehat{cFDR} \leq \alpha$ and proposed a rough solution. We proposed identifying L-curves and estimating f_0 as above, and for each L-region $L_S(\alpha^*)$, identifying a rectangle $R(\alpha^*)$ contained within it with vertices $(0,0)$, $(0,q_r)$, $(p_r,0)$, (p_r,q_r). Since $R(\alpha^*) \subseteq L_S(\alpha^*)$, we have

$$Pr\left((P,Q) \in L_S(\alpha^*)\right) \geq Pr\left((P,Q) \in R(\alpha^*)\right) \tag{18}$$

and $\widehat{(p_r, q_r)} \leq \alpha^*$, so the FDR associated with rejecting any (p,q) pairs falling in $L_S(\alpha^*)$ was approximately

$$\begin{aligned}
E\left(\frac{|\{i : p_i, q_i \in L_S(\alpha^*), H_0^p\}|}{\min\left(|\{i : p_i, q_i \in L_S(\alpha^*)\}|, 1\right)}\right) &\approx \frac{Pr\left((P,Q) \in L_S(\alpha^*)|H_0^p\right)}{Pr\left((P,Q) \in L_S(\alpha^*)\right)} \\
&\leq \frac{Pr\left((P,Q) \in L_S(\alpha^*)|H_0^p\right)}{Pr\left((P,Q) \in R(\alpha^*)|H_0^p\right)} \frac{Pr\left((P,Q) \in R(\alpha^*)|H_0^p\right)}{Pr\left((P,Q) \in R(\alpha^*)\right)} \\
&\approx \frac{Pr\left((P,Q) \in L_S(\alpha^*)|H_0^p\right)}{Pr\left((P,Q) \in R(\alpha^*)|H_0^p\right)} \widehat{cFDR}_S(p_r, q_r) \qquad (19) \\
&\leq \frac{\int_{L_S(\alpha^*)} f_0 dpdq}{\int_{R(\alpha^*)} f_0 dpdq} \alpha^* \qquad (20)
\end{aligned}$$

To approximately control FDR at α, our procedure found α^* so that expression (20) was $\leq \alpha$ and rejection H_0^p whenever $(p_i, q_i) \in L_S(\alpha^*)$.

As well as being approximate, this procedure was conservative due to inequality (18). Our new method avoids this conservative assumption, and is on firmer theoretical ground. Furthermore, our old method precluded use of $c\widehat{FDR}^n$ given approximation (19). We show by simulation below that this results in substantial improvement in power in our new method.

5 Assessment of performance

In this section, we address three main points. Firstly, we demonstrate that our new method controls type-1 error rate (as FDR) appropriately, and that the censoring approach of (14) and (15) is necessary. Secondly, we demonstrate that power is substantially improved relative to our previous method for fixed level of FDR control, and that use of $c\widehat{FDR}^n$ over $c\widehat{FDR}$ improves power further. Finally, we demonstrate that at least in some settings, rejection regions based on $c\widehat{FDR}^n$ can correspond to a more powerful procedure than rejection regions based on alternative CDF or PDF estimators.

In each simulation, we generated a set of values $S = (p_i, q_i)$, $i \in 1..n$ from random variables P, Q. We considered an extensive range of underlying parameters governing the distributions of P, Q. Details of the simulation strategy are shown in supplementary material, section 9.3. In general, each dataset contained n_1^p associations only with the principal trait (that is, for which H_P^0 was false but H_Q^0 was true), n_1^q associations only with the conditional trait (for which H_Q^0 was false and H_P^0 was true), and n_1^{pq} with both. We used various alternative distributions based on p-value transforms of normal, t (3df) and Cauchy distributions.

For each simulation, we considered the following null-hypothesis rejection procedures, aiming to control the FDR at either $\alpha = 0.1$ or $\alpha = 0.01$:

1. the B-H method applied to the values p_i
2. the B-H method applied to 'naive' v-values $v(p_i, q_i) = v_S(p_i, q_i)$
3. the B-H method applied to 'leave-one-out' v-values $v(p_i, q_i) = v_{S-(p_i,q_i)}(p_i, q_i)$
4. the B-H method applied to block-out v-values (after randomly separating observations into three equally sized subdivisions, so (p_i, q_i) is in subdivision $b(i)$, defining v-values $v_{S-b(i)}(p_i, q_i)$)
5. our previous method for FDR control applied to (p_i, q_i)

Given a rejection procedure, we then defined

$$FDP = \begin{cases} 0 & \text{if no rejections} \\ \frac{\text{number of falsely rejected null hypotheses}}{\text{total number of rejections}} & \text{if } \geq 1 \text{ rejection} \end{cases}$$

$$TDP = \begin{cases} 0 & \text{if no rejections} \\ \frac{\text{number of correctly rejected null hypotheses}}{\text{true number of associations}} & \text{if } \geq 1 \text{ rejection} \end{cases}$$

We analysed type 1 error in terms of the estimated FDR, $FDR = E(FDP) \approx \overline{FDP}$, and power in terms of the corresponding true-discovery rate $TDR = E(TDP) \approx \overline{TDP}$.

5.1 New FDR-controlling procedure leads to greater power than previous method, and adjustment improves power further

Expected FDP was consistent with the FDR control level when using leave-one-out v values or 'block-out v-values (rejection procedures 3,4). When using the 'naive' v-values $v_S(p_i, q_i)$ (rejection procedure 2), FDR was not controlled at the requisite level. When using our original method (rejection procedure 5), FDR was controlled at lower than the prescribed level, indicating that the method was conservative. FDR control was maintained when using the approximation of f_0 in equation (16). Results are shown in figure 2.

Having established the validity of rejection methods 3, 4, we compared the power of method 3 on 'adjusted' cFDR ($\widehat{cFDR}^n$) and on non-adjusted cFDR ($\widehat{cFDR}$), and the power of our previous method, rejection procedure 5, applied to $\widehat{cFDR}$. The use of $\widehat{cFDR}^n$ led to greater power than the use of $\widehat{cFDR}$, which in turn led to substantially greater power than our previous method (figure 3).

5.2 With current parameters, PDF- estimation leads to a less powerful procedure than CDF- based estimation

As $n \to \infty$, consistent estimators of $P(H_0^p | P = p, Q = q)$ will converge to optimal rejection regions while estimators of $P(H_0^p | P \leq p, Q \leq q)$ will not, and hence the former will ultimately be more powerful. However, we demonstrate that for a range of parameter values, the ECDF-based estimator is considerably more powerful than two consistent PDF-based estimators of $P(H_0^p | P = p, Q = q)$.

We considered parametric and KDE-based estimators of $P(H_0^p | P = p, Q = q)$. The parametric model was based on a four-Gaussian model detailed in supplementary material, section 9.4. We only considered the case in which parametric assumptions were satisfied. Both estimators of $P(H_0^p | P = p, Q = q)$ led to substantially lower power than $\widehat{cFDR}^n$ (figure 4). The performance of an oracle CDF procedure (using exact contours of F_0/F as rejection regions) and an oracle PDF procedure (using exact contours of f_0/f as rejection regions) are shown for comparison.

We also considered the performance of kernel-density and parametric estimates for cFDR (that is, of the CDF). These were less powerful than the ECDF-based cFDR estimator (supplementary material, sections 9.4.2, 9.5.1)

5.3 Iterated cFDR

Since our proposed method for type-1 error rate control maps p-value/covariate pairs to v-values preserving the p-value property, we are free to use the resultant v-values in a second cFDR-based analysis against a second covariate. This enables immediate and simple adaptation to a setting in which more than one set of covariates are available. In our motivating example, this would allow us to subsequently 'condition' on other potentially related diseases as well as OCA.

We simulated a set of p-values $\{p\} = \{p_i, i \in 1..1000\}$, with 100 true associations (H_1^p) and 900 non-associations (H_0^p). We then repeatedly simulated sets of covariates $\{q^j\} = \{q_i^j, i \in 1..1000\}$ with 100 associations, which for even j were either randomly spaced amongst the 1000 variables (uninformative covariates) and for odd j overlapped more-than-randomly with associations with principal p-values (informative covariates), with around 54 shared associations on average (see supplementary material, section 9.3.2 for more details). We then repeatedly conditioned p_i on the values q_i^j.

On repeated conditioning, all v-values when $H^p = 1$ tended to 0, while v-values when $H^p = 0$ remained uniform on $(0, 1)$ (figure 5). This indicated the potential to greatly strengthen the power of a high-dimensional association analysis by repeated conditioning in this manner, even when only half of the sets of covariates are informative.

5.4 Summary of BRCA analysis

Finally, we return to the motivating example. cFDR rejects more null hypotheses for BRCA (724) than B-H on BRCA data alone (678, figure 1A) or the subset of variables with OCA association (280, figure 1B). The procedure is asymmetrical in that it will not reject a BRCA null hypothesis for a low OCA p-value alone, and can readily be reversed: supplementary figure 15 shows a similar analysis analysing association with OCA.

6 Discussion

We present an improvement to the conditional false discovery rate method, a widely-used procedure in genetic discovery. Our new methods essentially involve computing an analogy of the p-value corresponding to the ranking of hypotheses defined by the cFDR estimator. Our method enables the cFDR to be used definitively in the discovery phase of -omics studies with control of a type-1 error rate. The general procedure of multiple p-value testing with a covariate has wide scientific application; see [1, 2, 11] for examples.

The $\widehat{cFDR}$ and $\widehat{cFDR}^n$ estimators make no distributional assumptions on P,Q. The type-1 error rate controlling method requires modelling of the PDF of $P, Q|H_0^p$, but this requires approximating a univariate PDF $Q|H_0^p$. Furthermore, this PDF is only used as an integrand rather than for direct point-estimates. It is reasonable to expect that for approximations to complex PDFs, relative average errors over intervals will be smaller than relative errors at individual points; parametric approximations tend to be smoother than the true distribution at a fine scale, and KDE-based approximations rougher. An obvious shortcoming of cFDR-based methods is the lack of asymptotic optimality. Methods based on consistent estimators of f_0/f will eventually outperform any estimator of F_0/F for large enough n (see supplementary material, section 9.5). However, the ECDF-based cFDR estimator was far stronger than PDF-based estimators at the values of n we simulated at $(10^3 - 10^4)$. In practical terms, it is important to note that n, being the number of variables, cannot generally be increased indefinitely, as opposed to, for instance, sample size. Essentially, the fitting of L-curves corresponds to a procedure by which the similarity between P, Q is assessed, and the degree of modulation when moving from p to v values corresponds to this similarity. Moreover, this assessment of similarity occurs intrinsically on the basis of the joint CDF rather than relying on a parametric description.

Our proposed 'iterated cFDR' procedure can be thought of as a meta-analysis of a series of experiments $E_P, E_{Q_1}, E_{Q_2}, ...$ giving rise to p-value sets $\{p_i\}$, $\{q_i^1\}$, $\{q_i^2\}$, ... when only the first set ($\{p_i\}$) are known to test the correct hypotheses; that is, be $U(0, 1)$ for null hypotheses. It enables us to find the set of non-null hypotheses corresponding to E_P (denoted $\{H_1^p\}$), even though the set of non-null hypotheses corresponding to E_{Q_j} (denoted $\{H_1^{Q_j}\}$) may only partly overlap $\{H_1^p\}$, may contain hypotheses not in $\{H_1^p\}$, and (half the time) may even carry no information about H_1^p at all. This could be used to refine the set of association statistics $\{p_i\}$ for a disease of interest by using sets of association statistics $\{q_i^1\}$, $\{q_i^2\}$, ... at the same variables for a range of separate traits. It could also be used to improve power when repeating an -omics study in a new ethnic group by levering on previous studies in different ethnicities.

In summary, our method improves the power of cFDR analyses and allows it to be used confidently in the setting of multiple hypothesis testing. This can enable more efficient use of data, and more information to be gained from the same datasets. Our method contributes to a set of tools for high-dimensional statistical analysis and has wide application across a range of fields in biomedicine and elsewhere.

7 Code availability

All functions necessary to apply the methods detailed in this work are available in the R package

`https://github.com/jamesliley/cfdr`

A full pipeline to generate the results in this paper is available in the git repository

`https://github.com/jamesliley/cfdr_pipeline`.

8 Acknowledgments

We thank Professor Sylvia Richardson for helpful comments on the structure of this work. CW is funded by the Wellcome Trust (WT107881) and the MRC (MC_UU_00002/4). JL was funded by WT107881 and by Johnson and Johnson, and for part of this work was on the Wellcome Trust PhD programme in Mathematical Genomics and Medicine, funded by the NIHR Cambridge BRC. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest: None declared.

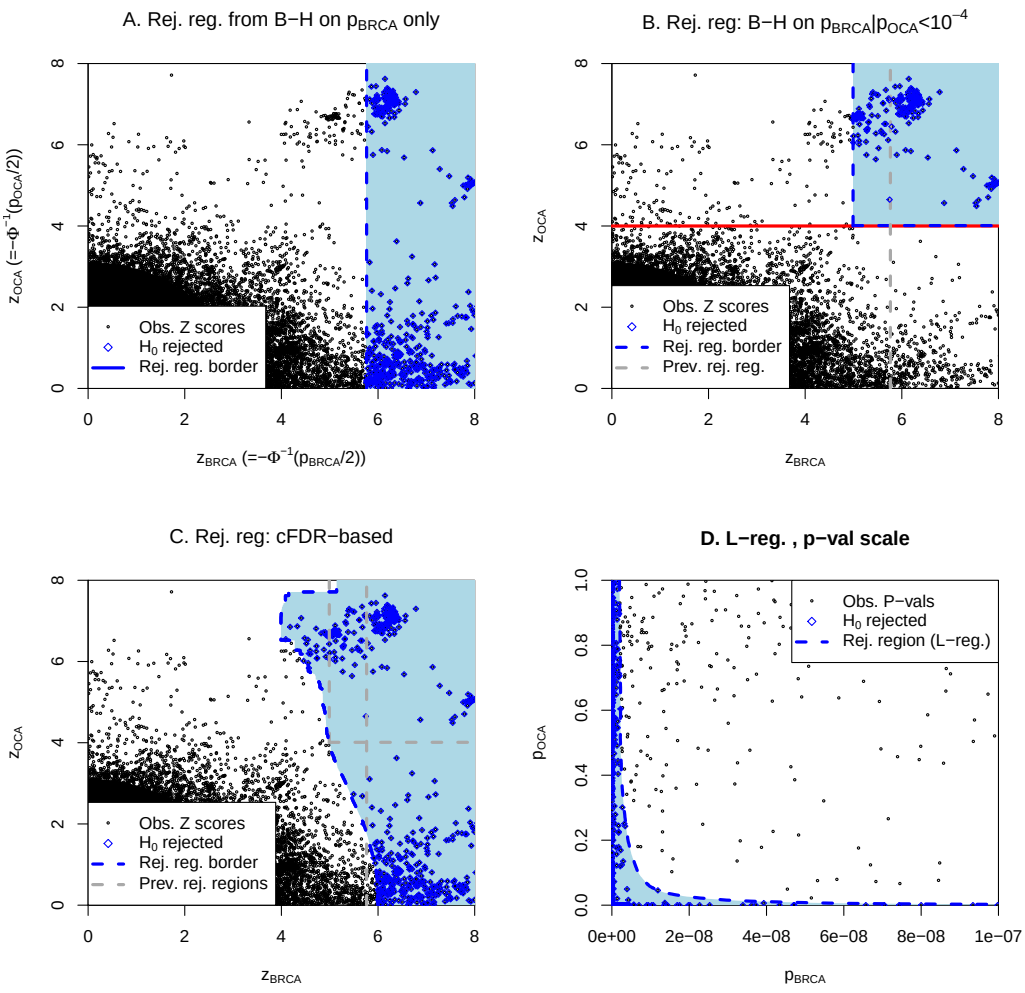

Figure 1: Illustration of cFDR approach using example data from TWAS study of breast cancer (BRCA), conditioning on ovarian cancer (OCA). Each plot shows test statistics from BRCA (x axis) and OCA (y axis) on either the Z (**A**-**C**) or p-value (**D**) scale. All rejection regions use methods to control FDR at $< 1 \times 10^{-6}$. **A** Benjamini-Hochberg (B-H) procedure applied to BRCA statistics alone, leads to a rejection region to the right of the blue dashed vertical line. **B** B-H applied to those variables for which z_{OCA} exceeds the threshold shown by a solid red line. **C** cFDR procedure: for the ith values $z_{BRCA}(i)$ $z_{OCA}(i)$, a B-H procedure aiming to control the FDR at α is conducted on only the variables for which $z_{OCA} \geq z_{OCA}(i)$, and if the ith null hypothesis is rejected during this procedure, it is rejected overall. We term the rejection region corresponding to this value α an 'L-region' $L(\alpha)$, shown as the shaded region. **D** The exposition that follows using p-values rather than Z scores, and so we reproduce the data and $L(\alpha)$ on the p-value scale. On this scale, the estimated cFDR at a point p_{BRCA}, p_{OCA} can be considered an estimate of the FDR corresponding to a fixed rejection region given by the box with with p_{BRCA}, p_{OCA} as its top-right corner, and the L-region $L(\alpha)$ roughly as the locus of top-right corners of boxes with estimated cFDR equal to α. Two such boxes are illustrated on the figure.

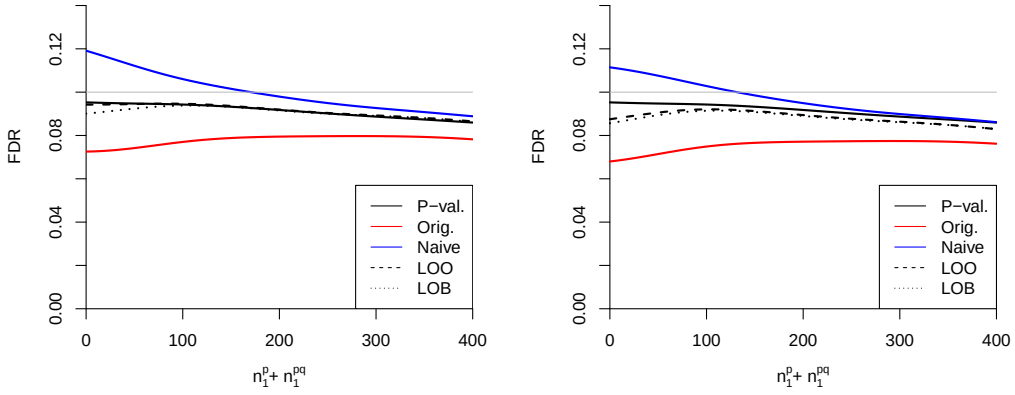

Figure 2: FDR control of various methods against $n_1^p + n_1^{pq}$, the total number of variables associated with P (the primary study under consideration). The horizontal line shows $\alpha = 0.1$, the desired FDR control level. Simulations in the left panel integrate L-regions over the the true distribution f_0; simulations in the right panel integrate over the estimated distribution as per equation (16). A corresponding plot with $\alpha = 0.01$ is shown in supplementary figure 16. Curves show moving weighted averages using a Gaussian kernel with SD 3/10 of the X axis range.

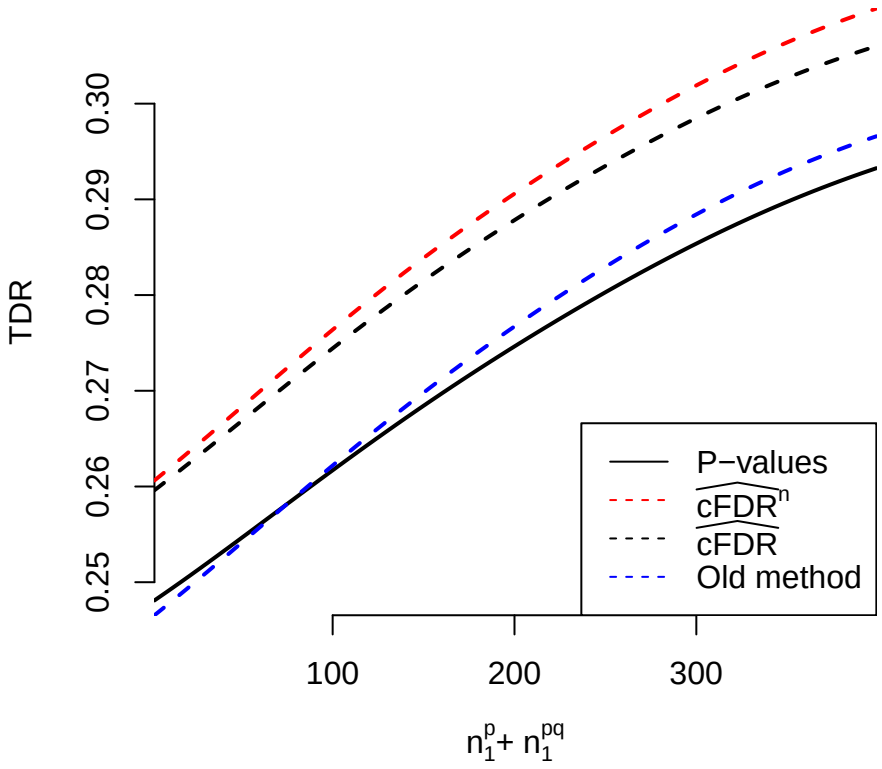

Figure 3: TDR of various methods against $n_1^p + n_1^{pq}$, the total number of variables associated with P (the primary study under consideration), at FDR control level $\alpha = 0.1$. A corresponding plot with $\alpha = 0.01$ is shown in supplementary figure 17. Curves show moving weighted averages using a Gaussian kernel with SD 3/10 of the X axis range.

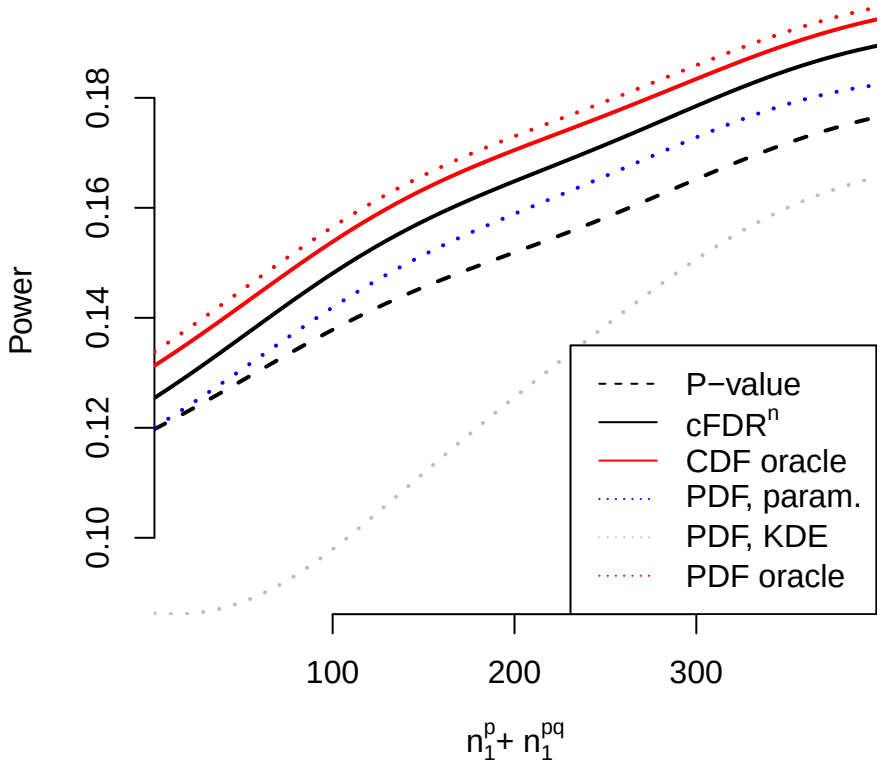

Figure 4: TDR of PDF-based methods against $n_1^p + n_1^{pq}$, the total number of variables associated with P (the primary study under consideration), restricting to simulations in which parametric assumptions were satisfied. A corresponding plot with $\alpha = 0.01$ is shown in supplementary figure 19. Curves show moving weighted averages using a Gaussian kernel with SD 3/10 of the X axis range.

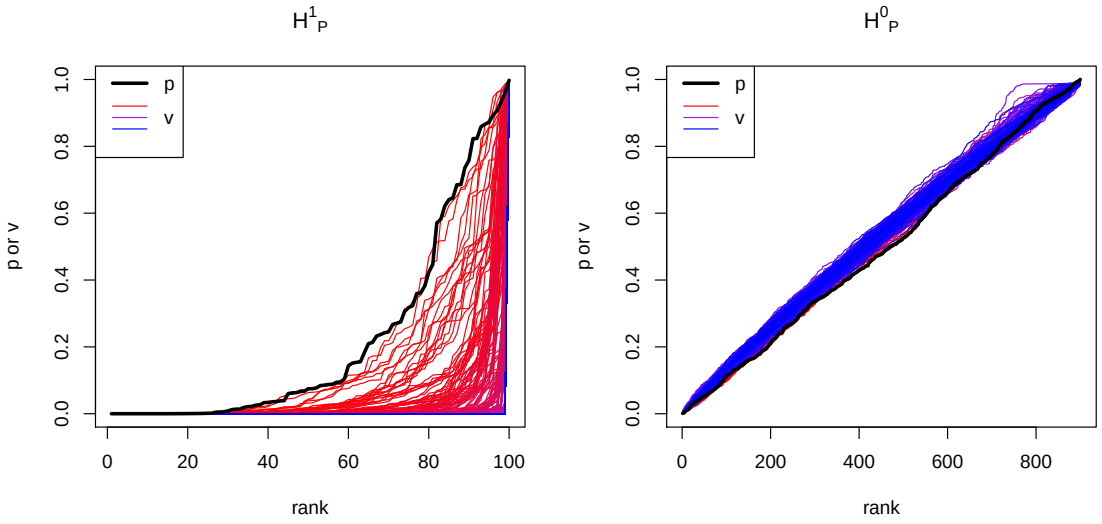

Figure 5: Plots of ordered p/v values for iterated cFDR. Summary statistics (p or v values) for non-null hypotheses (H_1^p) on the left and null hypotheses (H_0^p) on the right. The black line on the right corresponds to a uniform distribution. Colours change from red-purple-blue with further iteration. All v-values for non-null hypotheses tend toward 0, while v-values for null hypothesis remain in a uniform distribution.

Appendix

8.1 Optimal procedure

In this section, we show the following result. This is not original; it is shown in various forms in (at least) [1, 2, 3].

Theorem 1. *Let f_0 and f_1 be positive Lesbegue-integrable functions of (p, q) on some region Ω. Suppose a Lesbegue-measurable region R_0 satisfies:*

1. $R_0 = \{(p, q) : f_0(p, q)/f_1(p, q) \leq k, (p, q) \in \Omega\}$
2. $\int_{R_0} f_0(p, q)dpdq = \alpha$
3. $\int_{R_0} f_1(p, q)dpdq = 1 - \beta$

Then no other Lesbegue-measurable region $R \subset \Omega$ satisfies both

$$\int_R f_0(p, q)dpdq = \alpha \tag{21}$$

and

$$\int_R f_1(p, q)dpdq > 1 - \beta \tag{22}$$

Proof. Suppose such a region existed. Then given condition 21, we must have $f_0(p, q)/f_1(p, q) > k$ in $R \setminus R_0$, and since the integral of f_0 over R is equal to its integral over R_0,

$$\int_{R\setminus R_0} f_0(p, q)dpdq = \int_{R_0\setminus R} f_0(p, q)dpdq = \alpha - \int_{R_0\cap R} f_0(p, q)dpdq \tag{23}$$

Hence

$$\begin{aligned}
\int_R f_1(p, q)dpdq &= \int_{R\setminus R_0} f_1(p, q)dpdq + \int_{R\cap R_0} f_1(p, q)dpdq \\
&\leq k \int_{R\setminus R_0} f_0(p, q)dpdq + \int_{R\cap R_0} f_1(p, q)dpdq \\
&= k \int_{R_0\setminus R} f_0(p, q)dpdq + \int_{R\cap R_0} f_1(p, q)dpdq \\
&\leq \int_{R_0\setminus R} f_1(p, q)dpdq + \int_{R\cap R_0} f_1(p, q)dpdq \\
&= \int_{R_0} f_1(p, q)dpdq = 1 - \beta
\end{aligned} \tag{24}$$

a contradiction of 22. Regions $R \neq R_0$ can satisfy 21 and

$$\int_R f_1(p, q)dpdq = 1 - \beta \tag{25}$$

if and only if $R_0 \setminus R$ and $R \setminus R_0$ have Lesbegue measure 0. □

Corollary 1.1. *If $f_0(p,q) = f(P = p, Q = q|H_0^P)$ and $f_1(p,q) = f(P = p, Q = q|H_1^P)$ (where H_1^P is the alternative) then amongst all rejection regions R with fixed type-1 error rate $\alpha = \int_R f_0(p,q)dpdq$, power is maximised on a region inside a contour of f_0/f_1, if such a region exists.*

Denoting $f(p,q) = \pi_0 f_0(p,q) + (1-\pi_0)f_1(p,q)$, it is clear that a contour of f_0/f_1 is also a contour of f_0/f and of $P(H_0^p|P = p, Q = q)$, so an optimal rejection region is given by

$$\{(p,q) : P(H_0^p|P = p, Q = q) \leq k_\alpha\} \tag{26}$$

for some k_α.

8.2 Convergence results

In these appendices, we omit the X from $c\widehat{FDR}_X(p,q)$ and other functions when it is clear. We consider p, q to be in $(0,1)^2$.

Set $n = |X|$, $F_n(q) = \min(1, |\{i : q_i \leq q, (p_i, q_i) \in X\}|)$, $F_n(p,q) = \min(1, |\{i : p_i \leq p, q_i \leq q, (p_i, q_i \in X)\}|)$, $F(q) = P(Q \leq q)$, $F(p,q) = P(P \leq p, Q \leq q)$, and

$$C(p,q) = \frac{cFDR(p,q)}{P(H_0^p|Q \leq q)} = \frac{pF(q)}{F(p,q)} \tag{27}$$

We will assume $\partial F(p,q)/\partial p$ exists on $(0,1)^2$.

We define $c\widehat{FDR}t_X(p,q) = \min_{p' \geq p} c\widehat{FDR}_{X+(p',q)}(p',q)$, and $c\widehat{FDR}t_X^n(p,q)$ similarly for $c\widehat{FDR}_X^n$ ('t' for 'truncated'), so

$$L_X(\alpha) = \left\{(p,q) : c\widehat{FDR}t_X(p,q) \leq \alpha\right\} \tag{28}$$

The boundary of $L_X(\alpha)$ is continuous and piecewise-differentiable.

In this section, we show a series of results relating to convergence of cFDR estimates. We show results relating to convergence of $c\widehat{FDR}$ and $c\widehat{FDR}t$ on a line $q = q_0$, along with convergence of the co-ordinates of L-curves on such lines. We then show slightly weaker results regarding convergence across two-dimensional regions of the unit square.

Theorem 2. *Suppose that on a line segment $q = q_0$, $p_\gamma < p < 1$, we have $F(p,q) \geq \gamma > 0$ and $F(q_0) > 0$. Then on this segment, $c\widehat{FDR}(p,q)$ converges uniformly to $C(p,q)$ as $n \to \infty$. If additionally we have $\partial C(p,q)/\partial p \geq 0$, then $c\widehat{FDR}t(p,q)$ converges uniformly to $C(p,q)$ also.*

Proof. Condition on $q = q_0$, and (for the moment) $F_n(q) = m$. Set $\epsilon < \delta$ and let

$$g^-(p,\epsilon) = p\frac{\frac{m}{n}}{F(p,q) + \epsilon} \qquad g^+(p,\epsilon) = p\frac{\frac{m}{n}}{F(p,q) - \epsilon} \tag{29}$$

From the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality we have

$$\begin{aligned} Pr\left(F(p,q)-\epsilon \le \frac{F_n(p,q)}{m} \le F(p,q)+\epsilon \middle| q=q_0, F_n(q_0)=m\right) &\ge 1-e^{-2m\epsilon^2} \\ \Rightarrow Pr\left(g^-(p,\epsilon) \le \widehat{cFDR}(p,q)) \le g^+(p,\epsilon) \middle| q=q_0, F_n(q_0)=m\right) &\ge 1-e^{-2m\epsilon^2} \end{aligned} \tag{30}$$

If $\partial C(p,q)/\partial p \ge 0$, then (30) also holds for $\widehat{cFDRt}$. To see this, note $\widehat{cFDR}(p,q) \ge \widehat{cFDRt}(p,q)$, so if $\widehat{cFDRt}(p,q) \ge g^+(p,\epsilon)$ then $\widehat{cFDR}(p,q) \ge g^+(p,\epsilon)$ also. Now

$$\begin{aligned} \frac{\partial}{\partial p}C(p,q) \ge 0 &\Rightarrow F(p,q) \ge p\frac{\partial}{\partial p}F(p,q) \\ &\Rightarrow F(p,q)+\epsilon \ge p\frac{\partial}{\partial p}F(p,q) \\ &\Rightarrow g^-(p,\epsilon) > 0 \end{aligned}$$

Suppose that for some p we had $\widehat{cFDRt}(p,q_0) \le g^-(p)$. Then either $\widehat{cFDR}(p,q_0) = \widehat{cFDRt}(p,q_0)$ or $\widehat{cFDRt}(p,q_0) = \widehat{cFDR}(p',q_0)$ for some $p' > p$. In the first case $\widehat{cFDR}(p,q_0) \le g^-(p)$, and in the second, $\widehat{cFDR}(p',q_0) = \widehat{cFDRt}(p,q_0) \le g^-(p) \le g^-(p')$; in either case, $\widehat{cFDR}(p,q)$ escapes the bound $g^-(p)$ somewhere. Thus the probability on the LHS of (30) can only increase if $\widehat{cFDRt}$ replaces $\widehat{cFDR}$, and $\widehat{cFDRt}(p,q)$ is contained within the bounds $g^-(p)$, $g^+(p)$ with probability at least $1-\exp(-2m\epsilon^2)$.

We now move to remove the condition $F_n(q_0)=m$. Denote the events

$$\begin{aligned} A &: \left\{g^-(p,\epsilon) \le \widehat{cFDR}(p,q) \le g^+(p,\epsilon)\right\} \\ B &: \{q=q_0\} \end{aligned} \tag{31}$$

and, for some $\epsilon_2 < F(q_0)$

$$C : \left\{p\frac{F(q)-\epsilon_2}{F(p,q)+\epsilon} \le \widehat{cFDR}(p,q) \le p\frac{F(q)+\epsilon_2}{F(p,q)-\epsilon}\right\} \tag{32}$$

Denote by $S(\epsilon_2)$ the set of integers in $[n(F(q_0)-\epsilon_2), n(F(q_0)+\epsilon_2)]$ (and assume n is large enough that $S(\epsilon_2)$ is nonempty). If $m = F_n(q_0) \in S(\epsilon_2)$, the interval in event A is a

subinterval of that in event C. Thus

$$\begin{aligned} P(C|B) &= \sum_m P(C|B, F_n(q_0) = m)P(F_n(q_0) = m) \\ &\geq \sum_{m \in S(\epsilon_2)} P(C|B, F_n(q_0) = m)P(F_n(q_0) = m) \\ &\geq \sum_{m \in S(\epsilon_2)} P(A|B, F_n(q_0) = m)P(F_n(q_0) = m) \\ &\geq (1 - e^{-2\min\{S(\epsilon_2)\}\epsilon^2})P(m \in S(\epsilon_2)) \\ &\geq (1 - e^{-2n(F(q_0)-\epsilon_2)\epsilon^2})(1 - e^{-2n\epsilon_2^2}) \end{aligned} \tag{33}$$

where the last inequality comes from the DKW inequality on $F_n(q)$. Since $p \geq p_\epsilon$ and $F(p, q) \geq \gamma$ the widest part of the interval in event C can be made arbitrarily small on the interval $(p_\epsilon, 1)$ and $c\widehat{FDR}(p, q)$ converges uniformly to $C(p, q)$. If $\partial C(p, q)/\partial p \geq 0$, then so does $c\widehat{FDR}t(p, q)$.

□

Corollary 2.1. *Under the assumptions in theorem 2, $c\widehat{FDR}(p, q)$ and $c\widehat{FDR}t(p, q)$ are bound with fixed probability on the line segment $q = q_0$, $p_\gamma < p < 1$ in intervals of width $O(n^{-1/2})$*

Proof. In inequality 33, set

$$\epsilon = \frac{r}{\sqrt{F(q_0) - \epsilon_2}} \qquad \epsilon_2 = \frac{r_2}{\sqrt{n}} \tag{34}$$

Then the RHS is $(1 - \exp(-2r^2))(1 - \exp(-2r_2^2))$ which may be made arbitrarily small by varying r, r_2, and the difference between the upper and lower bounds in event $C|B$ is

$$p\frac{F(q_0) + \epsilon_2}{F(p, q_0) - \epsilon} - p\frac{F(q_0) - \epsilon_2}{F(p, q_0) + \epsilon} = 2p\frac{\sqrt{F(q_0)}r + F(p, q_0)r_2}{F(p, q_0)^2}\frac{1}{\sqrt{n}} + O\left(\frac{1}{n}\right) \tag{35}$$

□

Theorem 3. *Suppose that on a line segment $q = q_0$, $p_\gamma < p < 1$, we have $F(p, q) \geq \gamma > 0$, $F(q_0) > 0$, and $\partial C(p, q)/\partial p \geq \gamma_2 > 0$. Denote by $l(\alpha)$ the value of p at the intersection of the L-curve $L(\alpha)$ with the line $q = q_0$, so*

$$l(\alpha) = \sup\{p : c\widehat{FDR}t(p, q_0) \leq \alpha\} \tag{36}$$

and $c(\alpha)$ the value of p such that $C(p, q_0) = \alpha$ (unique if it exists). For any $\delta > 0$, the function $|l(\alpha) - c(\alpha)|$ converges uniformly to 0 for $\alpha \in [C(p_\epsilon, q_0) + \delta, 1]$.

Proof. Since $C(p, q_0)$ is continuous and increasing on $[p_\epsilon, 1]$, the value $c(\alpha)$ exists for $\alpha \in [C(p_\epsilon, q_0), C(1, q_0)] \supset [C(p_\epsilon, q_0) + \delta, 1]$ by the intermediate value theorem. The function $c\widehat{FDR}t(p, q)$ is continuous and nondecreasing on $[0, 1]$ and hence $l(\alpha)$ exists for $\alpha \in [c\widehat{FDR}t(0, q_0), c\widehat{FDR}t(1, q_0)] = [0, 1]$.

Given arbitrarily small positive ϵ_3, $\delta_2 < \delta$ choose n large enough that $c\widehat{FDR}t(p, q_0)$ is contained in $[C(p, q_0) - \delta_2, C(p, q_0) + \delta_2]$ for $p \in [p_\epsilon, 1]$ with probability at least $1 - \epsilon_3$. Then with probability $\geq 1 - \epsilon_3$, whenever the curve $c\widehat{FDR}t(p, q)$ is in the region bounded by the rectangle $p_\epsilon \leq p \leq 1$, $C(p_\epsilon, q_0) + \delta \leq q \leq 1$, it is bounded by the curves $C(p, q_0) - \delta_2)$, $C(p, q_0) + \delta_2$. The distance between the two curves in the q-direction is at most $2\gamma_2\delta_2$. Thus, if for some $\alpha \in [C(p_\epsilon, q_0) + \delta, 1]$, we have $|l(\alpha) - c(\alpha)| > \gamma_2\delta_2$, the curve $c\widehat{FDR}t(p, q_0)$ must escape the region bounded the curves $C(p, q_0) - \delta_2)$, $C(p, q_0) + \delta_2$.

So with probability at least $1 - \epsilon_3$ we have

$$\forall \alpha \in [C(p_\epsilon, q_0) + \delta, 1] : |l(\alpha) - c(\alpha)| \leq \gamma_2\delta_2 \tag{37}$$

which proves the statement. This is illustrated in figure 6.

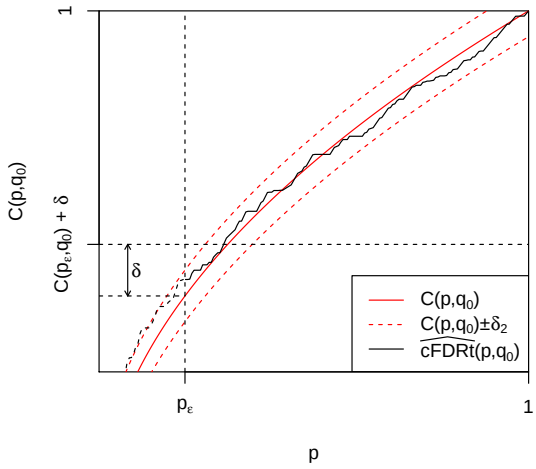

Figure 6: Convergence of intersections of L-curves with a line $q = q_0$. Functions $c\widehat{FDR}t(p, q_0)$, $C(p, q_0)$, and $C(p, q_0) \pm \delta_2$ are shown. The vertical distance between dashed red lines is $2\delta_2$, and since $\partial C(p, q0)/\partial p \geq \gamma_2$, the horizontal distance is at most $2\delta_2\gamma_2$. We must restrict the proof to $\alpha > C(p, q_0) + \delta$ because we cannot assert the behaviour of $c\widehat{FDR}t(p, q)$ left of the line $p = p_\epsilon$.

□

Theorem 4. *Let R be the region of the unit square for which $F(p,q) \geq \gamma > 0$ and $F(q) > 0$. Then on R, $c\widehat{FDR}(p,q)$ converges uniformly to $C(p,q)$, and if $\partial C(p,q)/\partial p \geq 0$, then so does $c\widehat{FDRt}(p,q)$*

Proof. We proceed very similarly to theorem 2. We employ a result from [47] that for any $\epsilon > 0$

$$Pr\left(\sup|F_n(p,q) - F(p,q)| \geq \frac{r}{\sqrt{n}}\right) \leq c(\epsilon)e^{(2-\epsilon)r^2} \tag{38}$$

from which, wherever $F(p,q) > f_{min} > \frac{r}{\sqrt{n}}$

$$Pr\left(\frac{1}{F(p,q)+\frac{r}{\sqrt{n}}} \leq \frac{1}{F_n(p,q)} \leq \frac{1}{F(p,q)-\frac{r}{\sqrt{n}}}\right) \geq 1 - c(\epsilon)e^{(2-\epsilon)r^2}$$
$$Pr\left(\frac{p\frac{F_n(q)}{n}}{F(p,q)+\frac{r}{\sqrt{n}}} \leq c\widehat{FDR}(p,q) \leq \frac{p\frac{F_n(q)}{n}}{F(p,q)-\frac{r}{\sqrt{n}}}\right) \geq 1 - c(\epsilon)e^{(2-\epsilon)r^2} \tag{39}$$

The values $F_n(p,q)$ and $F_n(q)$ are dependent. However, given some $r_2 > 0$, we have for all q (by the DKW inequality)

$$Pr\left(F(q) - \frac{r_2}{\sqrt{n}} \leq \frac{F_n(q)}{n} \leq F(q) + \frac{r_2}{\sqrt{n}}\right) \geq 1 - 2e^{-2r_2^2} \tag{40}$$

Denoting the event in (39) by A, the event in (40) by B, and C as:

$$\frac{p(F(q)-\frac{r_2}{\sqrt{n}})}{F(p,q)+\frac{r}{\sqrt{n}}} \leq c\widehat{FDR}(p,q) \leq \frac{p(F(q)+\frac{r_2}{\sqrt{n}})}{F(p,q)-\frac{r}{\sqrt{n}}} \tag{41}$$

we have, since the interval in A is a subinterval of that in C when conditioning on B:

$$\begin{aligned} P(C) &= P(C|B)P(B) + P(C|\neg B)P(\neg B) \\ &\geq P(C|B)(1-2e^{-2r_2^2}) \\ &\geq P(A|B)(1-2e^{-2r_2^2}) \\ &\geq (1-c(\epsilon)e^{(2-\epsilon)r^2})(1-2e^{-2r_2^2}) \end{aligned}$$

As before, this bound also holds for $c\widehat{FDRt}$ as long as $\partial C(p,q)/\partial p > 0$.

□

Corollary 4.1. *Under the assumptions in 4, $c\widehat{FDR}$ and $c\widehat{FDRt}$ are bound with fixed probability in R in intervals of width $O(n^{-1/2})$*

Proof. The difference between the upper and lower bounds in (41) is

$$\frac{2p(rF(q)+r_2F(p,q))}{F(p,q)^2}\frac{1}{\sqrt{n}}+O\left(\frac{1}{\sqrt{n}^3}\right) \tag{42}$$

□

Our final result describes errors on v-values. Given an L-region $L(\alpha)$, we define the M-region as the 'expected' L-region:

$$M(\alpha)=\{(p,q):C(p,q)\leq\alpha\} \tag{43}$$

and the 'error' on the v-value $v=\int_{L(\alpha)}f_0(p,q)dpdq$ as

$$|\Delta v|=\left|\int_{L(\alpha)}f_0(p,q)dpdq-\int_{M(\alpha)}f_0(p,q)dpdq\right| \tag{44}$$

We show the following:

Theorem 5. *Define R as in theorem 4, and further assume that on $f_0(p,q)=f(P=p,Q=q|H_0^P)$ is known and on R we have $\partial C(p,q)/\partial p\geq\gamma_2$. Write $R^c=[0,1]^2\setminus R$. Then the maximum error on any v-value is*

$$\int_{R^c}f_0(p,q)dpdq+O\left(\frac{1}{\sqrt{n}}\right) \tag{45}$$

Proof. Using theorem 4, bound $c\widehat{FDR}t(p,q)$ between $C(p,q)-\delta$, $C(p,q)+\delta$ with probability $\geq 1-\epsilon_3$, where $\delta=O\left(1/\sqrt{n}\right)$.

Since $F(p,q)$ is nondecreasing with p, we can describe $R=\{(p,q):F(p,q)\geq\gamma\}$ as the union of line segments $q=q_0$, $p_\epsilon(q_0)\leq p\leq 1$. We now define R_1 as the union of all line segments $q=q_0$, $p_\epsilon(q_0)+\delta\gamma_2\leq p\leq 1$; that is, R with the leftmost border shifted $\delta\gamma_2$ to the right.

We show the result by firstly noting that if an L-curve intersects a line segment $q=q_0$ at $l(\alpha)>p_\epsilon(q_0)+\delta\gamma_2$, and we have that $|l(\alpha)-c(\alpha)|>\delta\gamma_2$ (where $c(\alpha)$ is the intersection of the border of $M(\alpha)$ with $q=q_0$), then event C (equation 41) must have occurred in R, by the same argument as for theorem 3. Thus with probability at least $1-\epsilon_3$, every segment of a right-most border of an L-region $L(\alpha)$ in R_1 is at a horizontal distance from the corresponding rightmost-border of $M(\alpha)$ of at most $\delta\gamma_2$

We now write

$$\begin{aligned}\Delta v = &\left(\int_{L(\alpha)\cap R^c} f_0(p,q)dpdq - \int_{M(\alpha)\cap R^c} f_0(p,q)dpdq\right)\\ &+\left(\int_{L(\alpha)\cap(R\setminus R_1)} f_0(p,q)dpdq - \int_{M(\alpha)\cap(R\setminus R_1)} f_0(p,q)dpdq\right)\\ &+\left(\int_{L(\alpha)\cap R_1} f_0(p,q)dpdq - \int_{M(\alpha)\cap R_1} f_0(p,q)dpdq\right)\end{aligned} \tag{46}$$

The first term is at most $\int_{R^c} f_0(p,q)dpdq$. The region $R \setminus R_1$ has constant width $\delta\gamma_2$, and since f_0 only varies with q, hence the second term is at most $\int_{R\setminus R_1} f_0(p,q)dpdq = \delta\gamma_2 = O(n^{-1/2})$. Within R_1, if the horizontal separation between curves at the rightmost border of $L(\alpha)$ and $M(\alpha)$ is greater than $\delta\gamma_2$, then C has occurred, so this can happen with probability at most ϵ_3. Thus with probability $1-\epsilon_3$, the third term is also bounded by $\delta\gamma_2 = O(n^{-1/2})$, establishing the result.

□

8.3 Influence of a single point

Intuitively, adding a single point to a map defined by n other points should have a small effect on that map, and hence on the resultant v-values. We show the following:

Theorem 6. *Suppose we add a point (p^*, q^*) to a set of n points (p_i, q_i), considered as realisations of P, Q, and conditions are satisfied for convergence of v-values as above. Let Δv be the shift in a v-value corresponding to an L-curve through a randomly-chosen observation of P, Q after adding (p^*, q^*). Then*

$$E(|\Delta v|) = O\left(\frac{1}{n^2}\right) \tag{47}$$

where the expectation is over v-values, rather than (p^*, q^*)

Proof. Consider the profile of $c\widehat{FDR}t(p,q)$ on a line $q = q_0$, and how this changes with the addition of (p^*, q^*). The functions $F_n(q)$, $F_n(p,q)$ will be taken to be with respect to the n points (p_i, q_i) but not (p^*, q^*).

For $q_0 < q^*$, the addition of (p^*, q^*) changes neither $F_n(q_0)$ nor $F_n(p, q_0)$, so on lines $q = q_0 < q^*$ the profile of $c\widehat{FDR}t$ will remain the same.

Denote

$$c^+(p) = p\frac{F_n(q)+1}{F_n(p,q)} \qquad c^-(p) = p\frac{F_n(q)+1}{F_n(p,q)+1} \tag{48}$$

For $q_0 > q^*$, $p < p^*$, the value of $c\widehat{FDR}(p, q_0)$ will increase by

$$c^+(p) - p\frac{F_n(q)}{F_n(p,q)} = \frac{p}{F_n(p,q_0)} \tag{49}$$

and for $q_0 > q^*$, $p > p^*$, it will decrease by

$$p\frac{F_n(q)}{F_n(p,q)} - c^-(p) = p\frac{F_n(q_0) - F_n(p,q_0)}{F_n(p,q_0)(F_n(p,q_0)+1)} \tag{50}$$

In either case, $c\widehat{FDR}(p, q_0)$ changes by $O\left(\frac{1}{n^2}\right)$. The behaviour of $c\widehat{FDR}t$ is a little more complex. If we define $c_t^+(p)$ and $c_t^-(p)$ analogously to $c\widehat{FDR}t(p, q_0)$, then for $p > p^*$, $c\widehat{FDR}t(p, q_0)$ shifts to $c_t^-(p)$, and for $p < p^*$, it shifts to $\min(c_t^+(p), c_t^-(p^*)$ (see example in figure 7).

We can show that the absolute difference in $c\widehat{FDR}t(p, q_0)$ is always less than the absolute difference in $c\widehat{FDR}(p, q_0)$ after adding (p^*, q^*). Denote these differences $\Delta c\widehat{FDR}(p, q_0)$ and $\Delta c\widehat{FDR}t(p, q_0)$. Since $c\widehat{FDR}t(p, q_0)$ always shifts to between $c_t^-(p)$ and $c_t^+(p)$, it suffices to show that

$$c_t^+(p) - c\widehat{FDR}t(p, q_0) \leq \Delta c\widehat{FDR}(p, q_0) \tag{51}$$

$$c\widehat{FDR}t(p, q_0) - c_t^-(p) \leq \Delta c\widehat{FDR}(p, q_0) \tag{52}$$

Inequality (51) follows from the observation that $c^+(p) \propto c\widehat{FDR}(p, q_0)$, so order relations between $c\widehat{FDR}(p, q_0)$ and $c^+(p, q_0)$ are preserved. Thus

$$|\Delta c\widehat{FDR}t(p, q_0)| = \min_{p' \geq p} \Delta |c\widehat{FDR}(p', q_0)| \leq |\Delta c\widehat{FDR}(p', q_0)| \tag{53}$$

Order relations are not preserved between $c\widehat{FDR}(p, q_0)$ and $c^-(p)$, but the denominators increment at the same values of p. The functions $c\widehat{FDR}(p, q_0)$ and $c^-(p)$ both rise linearly in p between successive increment points p_a, p_d of $F_n(p, q_0)$, with $c\widehat{FDR}(p, q_0)$ having the higher gradient, since

$$\frac{F_n(q)}{F_n(p,q)} > \frac{F_n(q)+1}{F_n(p,q)+1} \tag{54}$$

At p_d, both functions are discontinuous and drop in value. On (p_a, p_d), the values of $c_t^-(p)$ and $c\widehat{FDR}t(p, q_0)$ are either equal to $c^-(p)$, $c\widehat{FDR}(p, q_0)$, or 'censored' at some values $c^-(p')$, $c\widehat{FDR}(p', q_0)$ with $p' > p$ (see the right-hand part of figure 7 for an example of this). We note that $c^-(p) > c^-(p'), p' > p \Rightarrow c\widehat{FDR}(p, q_0) > c\widehat{FDR}(p, q_0)$, so the first point at which $c_t^-(p)$ is censored on (p_1, p_2) is further right than the first point at which $c\widehat{FDR}t(p, q_0)$ is censored. Denote the leftmost point at which $c\widehat{FDR}t(p, q_0)$ is censored as p_b and the leftmost point at which $c_t^-(p)$ is censored as p_c, so $p_a \leq p_b \leq p_c \leq p_d$.

On (p_a, p_b), where neither are censored, $c\widehat{FDR}(p, q_0) - c^+(p) = c\widehat{FDRt}(p, q_0) - c_t^-(p)$ and $\Delta c\widehat{FDR}(p, q_0) = \Delta c\widehat{FDR}(p, q_0)$. On (p_b, p_c), when only $c\widehat{FDRt}(p, q_0)$ is censored, $c\widehat{FDRt}(p, q_0) - c_t^-(p) = c\widehat{FDR}(p_b, q_0) - c^-(p) < c\widehat{FDR}(p, q_0) - c^-(p)$, so $\Delta c\widehat{FDR}(p, q_0) \leq \Delta c\widehat{FDR}(p, q_0)$. On (p_c, p_d), we have $c\widehat{FDRt}(p, q_0) - c_t^-(p) = c\widehat{FDR}(p_b, q_0) - c^-(p_c) \leq c\widehat{FDR}(p_c, q_0) - c^-(p_c) \leq c\widehat{FDR}(p, q_0) - c^-(p)$, so again, $\Delta c\widehat{FDRt}(p, q_0) \leq \Delta c\widehat{FDR}(p, q_0)$. Thus, for all p,

$$|\Delta c\widehat{FDRt}(p, q_0)| \leq |\Delta c\widehat{FDR}(p, q_0)| = O\left(\frac{1}{n^2}\right) \tag{55}$$

where the multiplicative factor in $O(1/n^2)$ is independent of q_0. This inequality is demonstrated in the right panel of figure 7.

Denote by l_α the value of p at the intersection of an L-curve corresponding to $c\widehat{FDRt}(p, q) \leq \alpha$ with the line $q = q_0$. We have $l_\alpha = \max\{p : c\widehat{FDRt}(p, q_0) = \alpha\}$. The value l_α may shift substantially when adding p^*, q^*, as shown in figure 7 However, the effect is small on av-

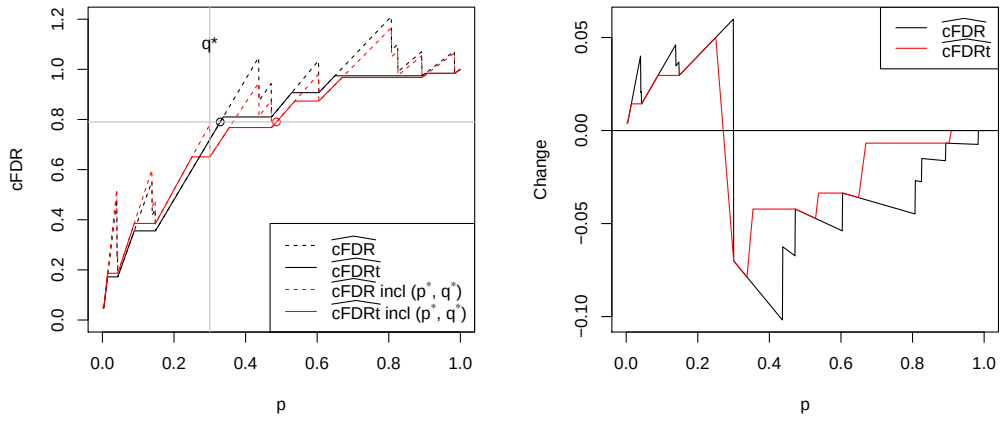

Figure 7: Behaviour of $c\widehat{FDR}$ and $c\widehat{FDRt}$ on a line $q = q_0$ after adding a point (p^*, q^*) to a set of n points (n is considerably smaller in this example than in figure 6). In the left panel, curves of $c\widehat{FDR}$ and $c\widehat{FDRt}$ before and after adding (p^*, q^*) are shown. Adding (p^*, q^*) may have a substantial impact on the intersection of an L-curve with $q = q_0$, such as that in the horizontal line: the black and red points show the intersection points of a curve before and after adding (p^*, q^*). However, the average effect across all curves is limited to the integral of the difference between the black and red lines, which is $O\left(1/n^2\right)$. The right panel demonstrates that $|\Delta c\widehat{FDRt}(p, q)| \leq |\Delta c\widehat{FDR}(p, q)|$.

erage. The plot of the function $l(\alpha)$ before and after adding (p^*, q^*) is identical to the plot of the function of p given by $c\widehat{FDRt}(p, q_0)$ before and after adding (p^*, q^*) rotated by

$\pi/2$. The average difference in movement of l_α is the integral of the difference in l_α with and without (p^*, q^*). However, this is simply the area between the two curves, which is invariant under rotating $\pi/2$. Hence

$$\int_0^1 \Delta l_\alpha d\alpha = \int_0^1 \Delta c\widehat{FDR}t(p, q_0)dp = O\left(\frac{1}{n^2}\right) \tag{56}$$

Denote the region $L(\alpha) : c\widehat{FDR}t(p, q) \leq \alpha$ and the co-ordinates of its rightmost border (L-curve) $(q, l_\alpha(q))$, $q \in (0, 1)$. Then, denoting the indicator function by I

$$\begin{aligned} v(L(\alpha)) &= \int_0^1 \int_0^1 I\left((p, q) \in L(\alpha)\right) f_0(p, q)dp\, dq \\ &= \int_0^1 \int_0^1 I\left((p, q) \in L(\alpha)\right) f_0^q(q)dp\, dq \\ &= \int_0^1 f_0^q(q) \int_0^1 I\left((p, q) \in L(\alpha)\right) dp\, dq \\ &= \int_0^1 f_0^q(q) l_\alpha(q)dq \end{aligned} \tag{57}$$

and the average error in v-values $v(L(\alpha))$ over $\alpha \sim U(0, 1)$ is

$$\begin{aligned} \int_0^1 \Delta v\left(L(\alpha)\right) d\alpha &= \int_0^1 \int_0^1 f_0^q(q)\Delta l_\alpha(q)dq\, d\alpha \\ &= \int_0^1 f_0^q(q) \int_0^1 \Delta l_\alpha(q)\, d\alpha dq \\ &= O\left(\frac{1}{n^2}\right) \int_0^1 f_0^q(q)dq \\ &= O\left(\frac{1}{n^2}\right) \end{aligned} \tag{58}$$

as required.

□

8.4 Asymptotic equivalence of PDF- and CDF- based L-regions

We show in this section that under a fairly common condition L-regions based on the PDF of p, q are similar to L-regions based on the CDF. In this section, we generally work on the Z-scale rather than the p-value scale for convenience.

Denote a 'fast-decreasing' function as a function g such that for each $\epsilon_1, \epsilon_2 > 0$, there exists δ such that for all X, Y of distance at least δ from the origin, we have

$$\iint\limits_{\substack{x\leq X, y\leq Y \\ (x-X)^2+(y-Y)^2\leq\epsilon_1^2}} g(x,y)dxdy > (1-\epsilon_2) \iint\limits_{x\leq X, y\leq Y} g(x,y)dxdy \tag{59}$$

so for $x \leq X, y \leq Y$, the function g falls off rapidly enough as x, y decrease that we can disregard its value except when it is close to X, Y.

We show the following:

Theorem 7. *Denote $c(x,y) = f_0(x,y)/f(x,y)$ and $C(x,y) = F_0(x,y)/F(x,y)$. Given a region of the $(-,-)$ quadrant $A_\epsilon = (-\infty, 0] \times (I_1 - \epsilon, I_2 + \epsilon)$ (where $\epsilon > 0$ is arbitrarily small), suppose that for $x, y \in A_\epsilon$ and for sufficiently small α we have*

1. *f_0 and f are fast-decreasing continuous positive functions*
2. *Along horizontal rays in A, $c(x,y)$ satisfies $\partial^2 \log(c(x,y))/\partial x^2 > 0$*
3. *The contour $c(x,y) = \alpha$ is a continuous and bounded function, and the rightmost bound increases to ∞ as $\alpha \to 0$*

Then for each $\epsilon_3 > 0$, there exists an ϵ_1 as above and an α_1 such that whenever $\alpha < \alpha_1$, there is a contour of $C(x,y)$ is never further than ϵ_3 from the contour $c(x,y) = \alpha$ in the region A_0.

Proof. Set R_3 as the region defined by the union of all circles of radius ϵ_3 with centres on points $y, l_\alpha(y)$. Choose $\epsilon_1 = \epsilon_3/2$ (supposing that $\epsilon_1 < \epsilon$), and define R_1 similarly to R_3 with radii ϵ_1. Let α^+ be the minimum value of f_0/f on the rightmost border of R_1, and α^- the maximum value on the leftmost border so $\alpha^+ > \alpha > \alpha^-$.

Condition 2 implies that for fixed y

$$\frac{d}{dx}\left(\frac{c(x+\epsilon_1, y)}{c(x,y)}\right) < 0 \tag{60}$$

Since the horizontal distance between the rightmost border of R_3 and the curve is at least $2\epsilon_1$ and similarly from the leftmost border of R_3, the values $\alpha^+ - \alpha$, $\alpha - \alpha^-$ must increase for fixed ϵ_1 as we move left. Thus, for some fixed $\epsilon_2 > 0$, choose δ_2 large enough that $\alpha^+/\alpha^- > 1/(1-\epsilon_2)^2$ and larger than the δ corresponding to ϵ_1, ϵ_2 by assumption, and α_1 large enough that the contour $c(x,y)$ is entirely left of the line $x = -\delta_2$.

Let X, Y be a point in A_0 to the right of R_3, so a circle of radius ϵ_1 centred at X, Y is in A_ϵ but does not intersect R_1. Thus across such a circle, the value of $c(x,y)$ is at least α^+. Similarly, across a circle of radius ϵ_1 centred to the left of R_3, the value of $c(x,y)$ is at most α^-

For x, y to the right of R_3, denote by H the circle of radius ϵ_1 centred at x, y. Now by the fast-decreasing property of f_0 and f, we have

$$F_0(x,y) > \int_H f_0(x,y)dxdy > \alpha^+ \int_H f(x,y)dxdy > \alpha^+(1-\epsilon_2)F(x,y) \tag{61}$$

so $C(x,y) > \alpha^+(1-\epsilon_2)$. Similarly for x, y to the left of R_3, we have $C(x,y) < \alpha^-/(1-\epsilon_2)$. By our choice of α_1, we have $\alpha^+(1-\epsilon_2) > \alpha^-/(1-\epsilon_2)$, so any contour of $C(x,y)$ at a level between these values must pass within R_3 through A_0. □

Contours of F_0/F correspond to contours of $cFDR$, and contours of f_0/f correspond to contours of $P(H_0^p|P=p,Q=q)$. Theorem 8.4 has obvious analogies in other quadrants, and for the p-value rather than z-score scale.

The conditions in the theorem may seem restrictive, but they are satisfied by many distributions; for instance, when f_0 and f are mixture Gaussian, and f dominates f_0 as $|x| \to \infty$. Figure 8 shows the similarity of a range of shapes of contours of C and c.

8.5 Failure of FDR control with $c\widehat{FDR} < \alpha$

As described in section 2, rejection procedure (9) is similar to the B-H procedure, and it may be naively thought that it also controls the FDR at α. This is not the case, and indeed the FDR of such a procedure (and the corresponding procedure with $c\widehat{FDR}^n$) may exceed α by an arbitrary factor depending on α and π_0.

This is most easily seen by considering the extreme case in which

$$P, Q|H_0^p \sim U(0,1)^2 \tag{62}$$
$$P, Q|H_1^p \sim (0,0) \tag{63}$$

where $\pi_0 = P(H_0^p)$ as usual. In this case we show:

Theorem 8. *Under the above distribution of P, Q, as $n \to \infty$, the FDR of rejection procedure* (9) *for $c\widehat{FDR}$ satisfies*

$$OR(FDR, \alpha) = \log\left(\frac{1-\alpha\pi_0}{1-\pi_0}\right) \tag{64}$$

and the corresponding procedure for $c\widehat{FDR}^n$ satisfies

$$\frac{FDR}{\alpha} = \frac{1-\log\left(\frac{\alpha}{1-\alpha}\frac{1-\pi_0}{\pi_0}\right)}{1-\alpha\log\left(\frac{\alpha}{1-\alpha}\frac{1-\pi_0}{\pi_0}\right)} \tag{65}$$

Corollary 8.1. *For $c\widehat{FDR}$, the relative error in FDR (relative to α) can grow arbitrarily large as $\pi_0 \to 1, \alpha \to 0$. For $c\widehat{FDR}^n$, the error can grow arbitrarily large as $\alpha \to 0$, regardless of π_0.*

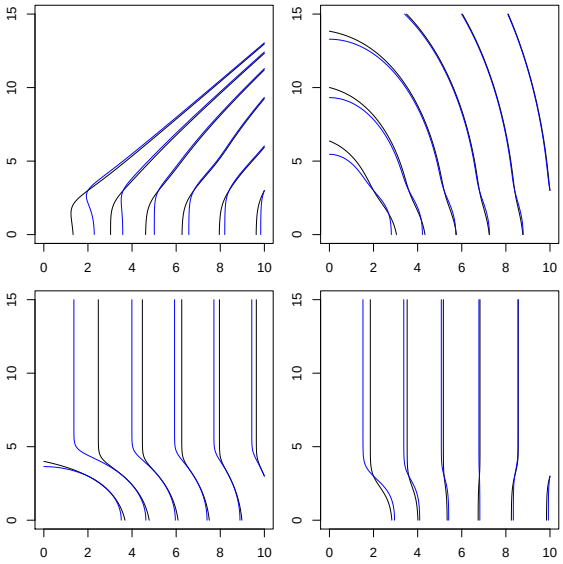

Figure 8: These plots show contours of CDF- based and PDF-based L-regions for a range of distributions of P, Q. Plots are on the Z-score scale (eg, rejection regions in terms of Z_P, Z_Q). The distributions are parametrised in terms of the mixture-Gaussian distribution detailed in supplementary material, section 9.4.1; parameters $((\pi_0,\ \pi_1,\ \pi_2,\ \tau_1,\ \tau_2,\ \sigma_1,\ \sigma_2))$ were (0.7,0.1,0.1,2,3,1.5,1.5), (0.7,0.1,0.1,2,2,3,3), (0.99,0.0005,0.0003,2,3,4,3), (0.7,0.1,0.05,3,2,2,2) respectively. Curves are generated passing through the points $Z_P, Z_Q = (3, 2..6)$, with curves further to the right corresponding to smaller α. As α gets smaller, contours $F_0(x, y)/F(x, y) = \alpha$ (black lines) become closer to contours of $f_0(x, y)/f(x, y) = \alpha$ (blue lines) under reasonably general circumstances

Proof. Suppose that we have a dataset $S = \{(p_i, q_i)\}$, $i \in 1..n$ of draws from P, Q under 62,63. Due to assumption 62 we have $P(P \leq p, Q \leq q|H_0^p) = pq$ and due to 63 we have $P(P \leq p, Q \leq q|H_1^p) = 1$. Now

$$\begin{aligned} cFDR(p, q) &= P(H_0^p|P \leq p, Q \leq q) \\ &= \frac{\pi_0 pq}{(1 - \pi_0) + \pi_0 pq} \end{aligned}$$

Now

$$\begin{aligned} P(H_0^p|Q \leq q) &= \frac{P(Q \leq q|H_0^p)P(H_0^p)}{P(Q \leq q|H_1^p)P(H_1^p) + P(Q \leq q|H_0^p)P(H_0^p)} \\ &= \frac{\pi_0 q}{(1 - \pi_0) + \pi_0 q} \end{aligned} \tag{66}$$

The estimate $\widehat{cFDR}(p,q)$ is proportional to a consistent estimator of

$$\frac{cFDR(p,q)}{P(H_0^p|Q \leq q)} = p\frac{(1-\pi_0)+\pi_0 q}{(1-\pi_0)+\pi_0 pq} \tag{67}$$

and since $P > 1/2 \Rightarrow H_0^p$, approximation 7 in the main paper is consistent, and $\widehat{cFDR}^n(p,q)$ is a generally consistent estimator of $cFDR(p,q)$.

The FDR of the rejection procedure $\widehat{cFDR}(p,q) \leq \alpha$ converges to the FDR of the rejection region $R_\alpha = \{(p,q) : cFDR(p,q)/P(H_0^p|Q) \leq q) < \alpha\}$ as $n \to \infty$ (see diagram in figure 9). Since this rejection region contains $(0,0)$, all $(1-\pi_0)n$ non-null hypotheses will be rejected, and the proportion of the total null hypotheses rejected will converge by the law of large numbers to

$$\begin{aligned}\int_{R_\alpha} f(P,Q|H_0^p)dpdq &= \int_{R_\alpha} dpdq \\ &= \frac{\alpha}{1-\alpha}\frac{1-\pi_0}{\pi_0}\log\left(\frac{1-\alpha\pi_0}{1-\pi_0}\right)\end{aligned} \tag{68}$$

and thus the FDR converges to

$$\begin{aligned}FDR &\to \frac{\text{number of null } (p_i,q_i) \text{ in } R_\alpha}{\text{total number of } (p_i,q_i) \text{ in } R_\alpha} \\ &= \frac{\pi_0 n \int_{R_\alpha} dpdq}{(1-\pi_0)n + \pi_0 n \int_{R_\alpha} dpdq} \\ &= \alpha\frac{\log\left(\frac{1-\alpha\pi_0}{1-\pi_0}\right)}{1-\alpha+\alpha\log\left(\frac{1-\alpha\pi_0}{1-\pi_0}\right)}\end{aligned} \tag{69}$$

which can be written as

$$\log\left(\frac{1-\alpha\pi_0}{1-\pi_0}\right) = \frac{FDR(1-\alpha)}{\alpha(1-FDR)} = OR(FDR,\alpha) \tag{70}$$

where OR is the odds ratio. Hence FDR can exceed α by an arbitrary degree when pi_0 is close to 1.

The second part of the proof can be shown similarly. The RHS of equation 65 rises as $-\log(\alpha)$ as $\alpha \to 0$, whatever the value of π_0.

□

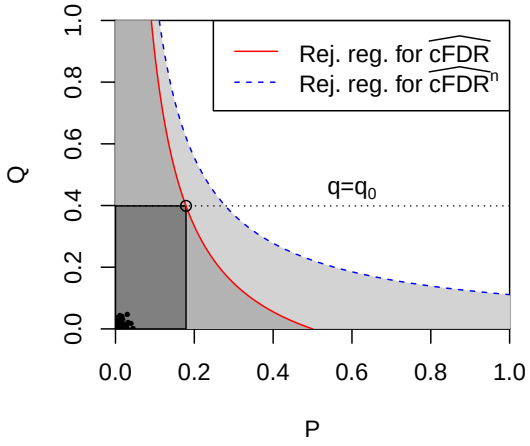

Figure 9: Rejection regions for rejection procedure (9) under assumptions in section 8.5. FDR is not controlled for either of the cFDR-based rejection regions. For reference, the B-H procedure applied to the set of (p, q) pairs with $q \leq q_0$ would reject everything in the dark gray rectangle, which includes all true positives, and this would control FDR at α. The cFDR based rejection regions reject the same number of true-positives, but far more false positives, so FDR control is lost.

9 Supplementary material

9.1 TWAS details

TWAS test association of gene expression (a biologically interpretable quantity) with some trait, when the trait and expression have not been measured on the same individuals, by using a reference expression-quantitative trait locus (eQTL) study and genome-wide association study (GWAS). A TWAS firstly uses eQTL data to learn rules to predict mRNA expression levels in a given tissue according to individual genotype, then applies these rules to predict expression for each individual in a GWAS, and finally compares these predicted expression levels across the GWAS trait of interest [40].

The results from a TWAS are a set of p-values corresponding to tissue-gene pairs. The object is to find which tissue-gene pairs are associated with the disease under consideration; that is, which p-values come from a distribution other than $U(0,1)$. In general, such tissue-gene pairs are a small proportion of all those considered. We consciously ignore any prior information which could be derived from tissues likely to be BRCA or OCA associated, for purposes of demonstration.

We considered TWAS datasets for breast cancer (BRCA, [41]) and ovarian cancer (OCA, [42]), containing tests for varying numbers of genes across 54 tissues. BRCA and OCA have considerable phenotypic overlap [43], and we may hope that summary statistics for one disease may be useful for leverage in association analyses of the other. We considered RNA-tissue pairs available in both datasets, restricting our analysis only to pairs in which RNA expression was predicted using data from the GTEx consortium, comprising a total of $n = 80222$ hypotheses.

Given the GWAS-scale dimensionality of testing, we chose a conservative FDR control level $\alpha = 1 \times 10^{-6}$. We used both $c\widehat{FDR}$ and $c\widehat{FDR}^n$ to generate v-values, and used the block-out method with blocks assigned according to genes, so expression levels for each gene were assigned a separate block (for 11327 folds in total).

9.2 Correlation of $P, Q|H_0^p$

Throughout this paper we have largely assumed that $P \perp\!\!\!\perp Q|H_0^p$, but methods can be easily adapted as long as the distribution $P \perp\!\!\!\perp Q|H_0^p$ is known (or we are happy to assume it is known). An example where this occurs is in the GWAS literature in which the studies giving rise to p_i and q_i share samples (usually control samples), which induces a known correlation between P and Q under H_0^p.

This can generally be managed by using the true f_0 (or an approximation allowing for dependence of P, Q under H_0^p) in equation 12 in the main paper and accounting for the true f_0 in the computation of $P(P \leq p|Q \leq q, H_0^p)$ necessary to estimate cFDR (equation 6 in main paper) where it is otherwise equal to p under assumption (1) in the main paper.

We demonstrate how to approximate $P(P \leq p|Q \leq q, H_0^p)$ and f_0 in the specific case of shared controls in a previous paper [23].

9.3 Simulation strategy

9.3.1 Simulations to establish FDR control and power

The distributions and variables we considered are shown in table 1. In general, our simulations took the form of: a fixed number n_1^{pq} of variables associated in both P and Q, a fixed number of variables n_1^p associated only with p, a fixed number of variables n_1^q associated only with Q and the remainder of variables associated with neither P nor Q. We will refer to these four classes of variables as C_1, C_2, C_3, C_4. We designated that within each class of variables, P and Q were iid.

Table 1: Variables used in simulations

Variable	Description	Point values	Sampling distribution for other values
n	Total number of variables	10^3, 10^4	$10^{U(3,4)}$ (rounded)
n_1^{pq}	Number of variables assoc. with P, Q	0, 10, 200	$U(0, 200)$
n_1^p	Number of variables associated with P	0,10,200	$U(0, 200)$ (rounded)
n_1^q	Number of variables associated with Q	0,10,200	$U(0, 200)$ (rounded)
s_p	Scale for distribution of P (see below)	$\frac{3}{2}$, 3	$U\left(\frac{3}{2}, 3\right)$
s_q	Scale for distribution of P_j	$\frac{3}{2}$, 3	$U\left(\frac{3}{2}, 3\right)$
d	Form of distributions	Normal, t (3df), Cauchy (eq. prob.)	Normal, t (3df), Cauchy (eq. prob.)

For variables in C_1, C_2, we set the distribution of P (determined by d, s_p) by first simulating Z scores:

d=1: $-\Phi^{-1}\left(\frac{P}{2}\right)\frac{1}{s_p} \sim N(0,1)$

d=2: $-\Phi^{-1}\left(\frac{P}{2}\right)\frac{1}{s_p} \sim t(\text{df} = 3, \text{ncp} = 0)$

d=3: $-\Phi^{-1}\left(\frac{P}{2}\right)\frac{1}{s_p} \sim \text{Cauchy}(\text{location} = 0, \text{scale} = 1)$

where $-\Phi^{-1}\left(\frac{P}{2}\right)$ can be considered a Z-score corresponding to P, and s_p a scaling factor for the distribution. We set the distribution of Q in C_1, C_3 similarly, with s_q in place of s_p. The values p_i, q_i for $i \in C_4$ were sampled from $U(0, 1)$.

Within each class $C_1 - C_4$ we simulated independent P, Q. We also considered s_p to be constant across C_1 and C_2, s_q to be constant across C_1 and C_3, and d to be constant in

each simulation. We deliberately oversampled parameter sets with $n_1^p + n_1^{pq} = 0$ given the importance of this part of the parameter space. We considerd $\alpha \in (0.01, 0.1)$.

9.3.2 Iterated cFDR

For our analysis of iterated cFDR, each set of $\{p_i\}$, $\{q_i^j\}$ was generated from 900 random samples from $U(0,1)$ and 100 random samples from $2\Phi(-|N(0,3^2)|)$, where Φ is the normal CDF.

Denoting H^p as an indicator variable for the hypothesis relating to $\{p_i\}$ and H_j^q as an indicator variable for the hypothesis relating to $\{q_i^j\}$, we designated $H^p \perp\!\!\!\perp H^q$ for odd iterations, and $P(H_i^q = 1|H^p = 1) = 15P(H_i^q = 1|H^p = 0)$ for even iterations. The value of 15 was chosen so that the number of overlaps beteen $H^p = 1$ and $H^q = 1$ would be approximately 50.

Starting with $v_0 = p$, we designated values $v_{i+1} = v(v_i, q^i)$. We used $c\widehat{FDR}^n$ as an estimator and used leave-one-out v-values.

9.4 Alternative cFDR estimators, and estimators of cfdr

In this section, we introduce new estimators of the cFDR $P(H_0^p|P \leq p, Q \leq q)$ and cfdr $P(H_0^p|P = p, Q = q)$, for use in simulations as detailed in section 5 of the main paper.

The main incentive for different estimators of cFDR is the tendency for the ECDF based estimator (equation (6) in the main paper) to have marked discontinuities at extremes of the unit square. This is illustrated in figure 10. The main incentive for estimators of cfdr is to allow comparison of PDF- and CDF- based estimators of the optimum rejection region detailed in section 2.2 in the main paper.

9.4.1 Parametric estimators for cFDR and cfdr

The estimate $c\widehat{FDR}$ is based on empirical quantities estimated directly by empirical CDFs of (P, Q). We consider here estimators based on approximating the joint distribution of P, Q using a bivariate mixture-normal parametrisation. This estimator enforces continuity of $c\widehat{FDR}$ on the open unit square, and is robust to small deviations in p-values, overcoming the effect detailed in figure 10. It is easiest to visualise parametrisations as distributions over the unsigned Z scores $(Z_p, Z_q) = \left(-\Phi^{-1}(P/2), -\Phi^{-1}(Q/2)\right)$ with $\Phi^{-1}(x)$ denoting the standard normal quantile function at x.

We use a parametrisation with seven parameters: (π_0, π_1, π_2, τ_1, τ_2, σ_1, σ_2), which parametrise a four-part bivariate mixture-Gaussian distribution over the $(+,+)$ quadrant

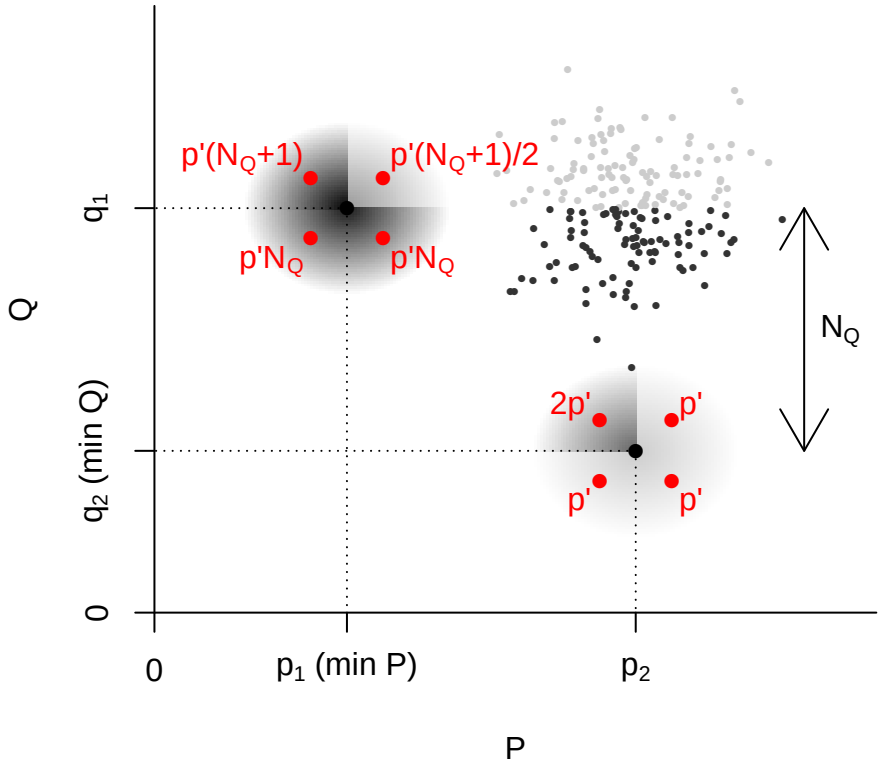

Figure 10: Dependence of $c\widehat{FDR}$ values on location of nearby points. In this example, we denote by (p_1, q_1), (p_2, q_2) the points at the (unique) left and lower extremes of the observed p-value distribution respectively; that is, $p_1 = \min(p_i)$, $q_2 = \min(q_i)$. We set N_Q as the number of points with $q_1 \leq q_i \leq q_2$ (small black points). If we add a test point (p', q') (shown in red) in a small neighbourhood of either (p_1, q_1) or (p_2, q_2), the estimated $c\widehat{FDR}_{S+(p',q')}(p', q')$ (shown in red next to the point) differs by a factor of 2 in different quadrants of the neighbourhood.

with PDF:

$$\begin{aligned} f^p(x,y) = & \ 4\pi_0 N_{\Sigma_0}(x,y) \\ & + 4\pi_1 N_{\Sigma_1}(x,y) \\ & + 4\pi_2 N_{\Sigma_2}(x,y) \\ & + 4(1-\pi_0-\pi_1-\pi_2) N_{\Sigma_3}(x,y) \end{aligned} \tag{71}$$

where $N_\Sigma(x,y)$ is the PDF of the bivariate normal distribution centred at the origin with variance Σ, the factor of 4 is due to to only unsigned Z-scores being used, and

$$\Sigma_0 = I_2 \qquad \Sigma_1 = \begin{pmatrix} \tau_1^2 & 0 \\ 0 & 1 \end{pmatrix} \qquad \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & \sigma_1^2 \end{pmatrix} \qquad \Sigma_3 = \begin{pmatrix} \tau_2^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \tag{72}$$

This model specifies a proportion π_1 of study variables to be associated only with the trait of interest P (with $SD(Z_P) = \tau_1$), a proportion π_2 to be associated only with the second trait Q (with $SD(Z_Q) = \sigma_1$), and a proportion $(1-\pi_0-\pi_1-\pi_2)$ to be associated with both ($Var(Z_P, Z_Q) = \Sigma_3$). We can now write:

$$\begin{aligned} f_0^p(x,y) &= f(Z_P = x, Z_Q = y | H_0^p) \\ &= 4\frac{\pi_0}{\pi_0+\pi_2} N_{\Sigma_0}(x,y) \\ &\quad + 4\frac{\pi_2}{\pi_0+\pi_2} N_{\Sigma_2}(x,y) \end{aligned} \tag{73}$$

We allow different values of σ_1, σ_2 and τ_1, τ_2 to allow for potentially different reasons for shared (both P and Q) and independent (P XOR Q) associations.

Maximum-likelihood estimates of parameters (π_0, π_1, π_2, τ_1, τ_2, σ_1, σ_2) can be obtained using an E-M algorithm [45]. Given these and corresponding estimates $\widehat{f^p}$, $\widehat{f_0^p}$ of f^p, f_0^p, we can then define corresponding estimates of cFDR, cfdr (implicitly conditioning on parametric assumptions):

$$\begin{aligned} \widehat{cFDR}^p(p,q) &= \frac{P(Z_P \le z_p, Z_Q \le z_q | H_0^p)}{P(Z_P \le z_p, Z_Q \le z_q)} P(H_0^p) \\ &\propto \frac{\int_{z_q}^{\infty}\int_{z_p}^{\infty} \widehat{f_0^p}(x,y)dxdy}{\int_{z_q}^{\infty}\int_{z_p}^{\infty} \widehat{f^p}(x,y)dxdy} \\ \widehat{cfdr}^p(p,q) &= \frac{f(Z_P = z_p, Z_Q = z_q | H_0^p)}{f(Z_P = z_p, Z_Q = z_q)} P(H_0^p) \\ &\propto \frac{\widehat{f_0^p}(z_P, z_Q)}{\widehat{f^p}(z_P, z_Q)} \end{aligned} \tag{74}$$

where X in this case is the set of points used in the estimation of parameters.

9.4.2 KDE-based estimators for cFDR and cfdr

To avoid distributional assumptions while maintaining a smooth form for the density of P, Q, a second estimator of $P(P < p|Q < q)$ can be derived from a two-dimensional kernel density estimator (KDE). We had no reason to prefer any kernel function over another, so opted to use a normal kernel with constant variance I_2. The PDF corresponding to Z_p, Z_q at x, y was modelled in the usual way as

$$f^k(x,y) = \frac{1}{n}\sum_i \frac{1}{\sigma_p\sigma_q}\phi\left(\sqrt{\left(\frac{x-\{-\Phi^{-1}(p_i/2)\}}{\sigma_p}\right)^2+\left(\frac{y-\{-\Phi^{-1}(q_i/2)\}}{\sigma_q}\right)^2}\right) \tag{75}$$

where $\phi(.)$ is the standard normal density. Values σ_p and σ_q are determined using a standard method based on the observations $p_i, q_i \in X$ [46].

Unlike the parametric estimate above, this does not intrinsically specify the density of $P, Q|H_0^p$. We thus incorporate the estimator $\widehat{Pr}_X(H_0^p|Q \le q)$ from equation 7 in the main paper, and write (implicitly conditioning on correctness of approximations)

$$\begin{aligned}
\widehat{cFDR}^k_X(p,q) &= \frac{P(P \le p|Q \le q, H_0^p)}{P(P \le p|Q \le q)} P(H_0^p|Q \le q) \\
&= \frac{pP(Q \le q)}{P(P \le p, Q \le q)} P(H_0^p|Q \le q) \\
&= \frac{p\int_{z_q}^{\infty}\int_0^{\infty} f^k(x,y)dxdy}{\int_{z_q}^{\infty}\int_{z_p}^{\infty} f^k(x,y)dxdy}\widehat{Pr}_X(H_0^p|Q \le q)
\end{aligned} \tag{76}$$

where X is the set of points used in the KDE in equation 75. We note that this estimator converges to $\widehat{cFDR}^n$ as $\sigma_p, \sigma_q \to 0$.

Estimating cfdr using KDEs requires estimation of $f(Z_P = z_p, Z_Q = z_q|H_0^p)$. We use assumption 1 from the main paper, and as for equation 7 in the main paper we assume that $Q|H_0^p \sim Q|P > 1/2$. We then fit a one-dimensional KDE to the values $z_{q_i}|p_i > 1/2$, and denote the resultant function of q as $\hat{f}_0^k(q)$. We then write (conditioning on assumptions)

$$\begin{aligned}
\hat{f}_0(p,q) &= p\hat{f}_0^k(q) \\
\widehat{cfdr}^k(p,q) &= \frac{f(P=p, Q=q|H_0^p}{f(P=p, Q=q)} \\
&= \frac{\hat{f}_0(p,q)}{f^k(z_p, z_q)}
\end{aligned} \tag{77}$$

9.5 Analysis of cFDR estimators

We required that all estimators be nonincreasing in p, so all were censored when generating L-curves or designing rejection procedures in the same way as in 9 and 11 for $\widehat{cFDR}$ in the main paper.

We show in figures 11, 12, 13 a series of plots at different values of n which indicate the behaviour of L-curves as n increases. In all cases, P, Q are sampled under the parametric assumptions in supplementary section 9.4.1, with (π_0, π_1, π_2, τ_1, τ_2, σ_1, σ_2) = $(0.7, 0.1, 0.15, 1.5, 2, 1.5, 2)$. Curves are drawn through $(0.1, 0.1)$, which would generally corresponds to a very high FDR level, so oracle PDF and oracle CDF curves are markedly different.

Importantly, $\widehat{cfdr}^p$ and $\widehat{cfdr}^k$ converge to the optimal rejection region (oracle PDF) while $c\widehat{FDR}$, $c\widehat{FDR}^p$ and $c\widehat{FDR}^k$ do not. However, the estimates of the latter are less noisy.

9.5.1 Parametric- and KDE- based cFDR estimators are less powerful than the ECDF-based estimator

We show in this subsection that under our simulated conditions, PDF- and KDE- based estimates of the cFDR lead to less powerful procedures than the ECDF-based cFDR when parametric assumptions are not satisfied. Details of the alternative estimators are given in supplementary material, section 9.4. Our parametric estimator assumed a distribution to $-\Phi^{-1}(P/2), -\Phi^{-1}(Q/2)$ given by a four-part bivariate mixture-Gaussian distribution, with seven degrees of freedom. We derived maximum-likelihood estimates of parameters using an E-M algorithm.

Our KDE estimator replaced each point p_i, q_i with a Gaussian density centred at p_i, q_i with covariance I_2. We note that the ECDF-based estimator is the limit of a KDE-based estimator as bandwidth tends to 0.

When parametric assumptions were satisfied (figure 14, left panel), the parametric-based estimator was more powerful than the ECDF-based estimator. When parametric assumptions were not satisfied (figure 14, right panel, in which f_1 had a bivariate t or Cauchy distribution), the ECDF estimator was more powerful than the parametric estimator. Both were more powerful than the KDE based estimator in both cases. The performance of an oracle CDF procedure (using exact contours of F_0/F as rejection regions) is shown for comparison.

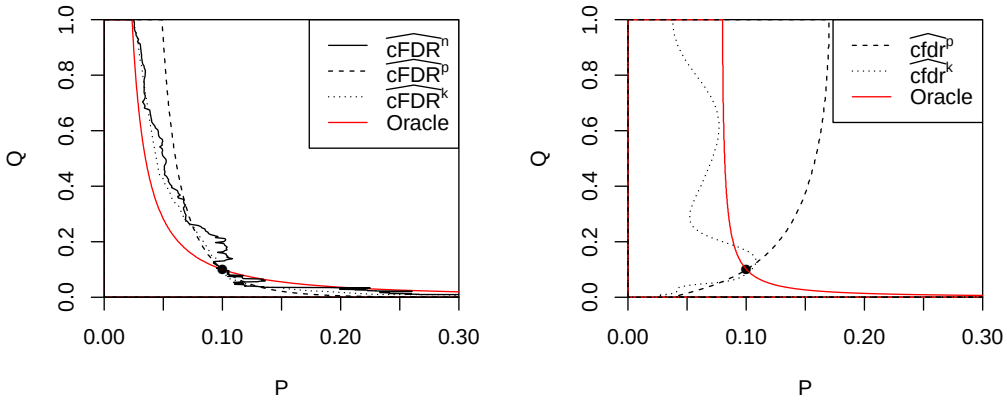

Figure 11: L-curves using various methods for cFDR estimation; n=1000. CDFs on left, PDFs on right.

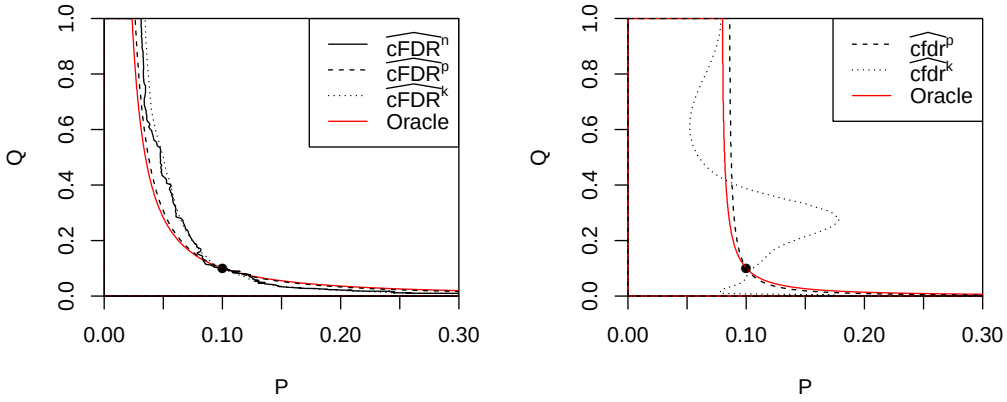

Figure 12: L-curves using various methods for cFDR estimation; n=10000. CDFs on left, PDFs on right.

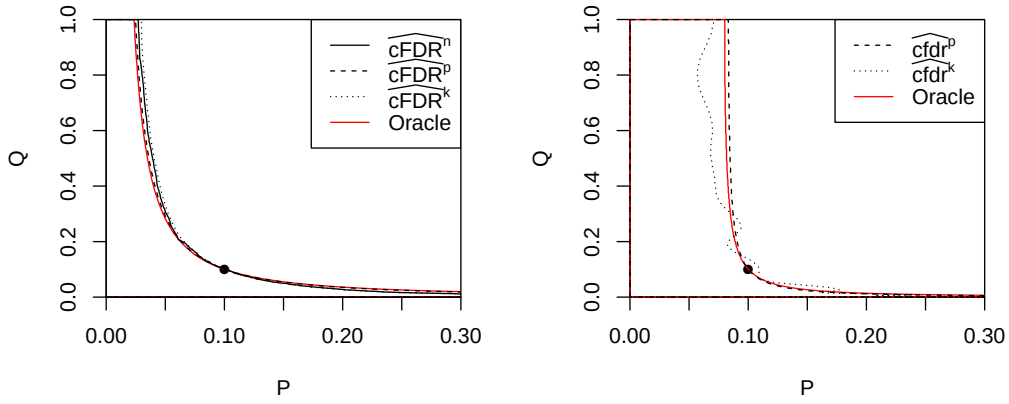

Figure 13: L-curves using various methods for cFDR estimation; n=100000. CDFs on left, PDFs on right.

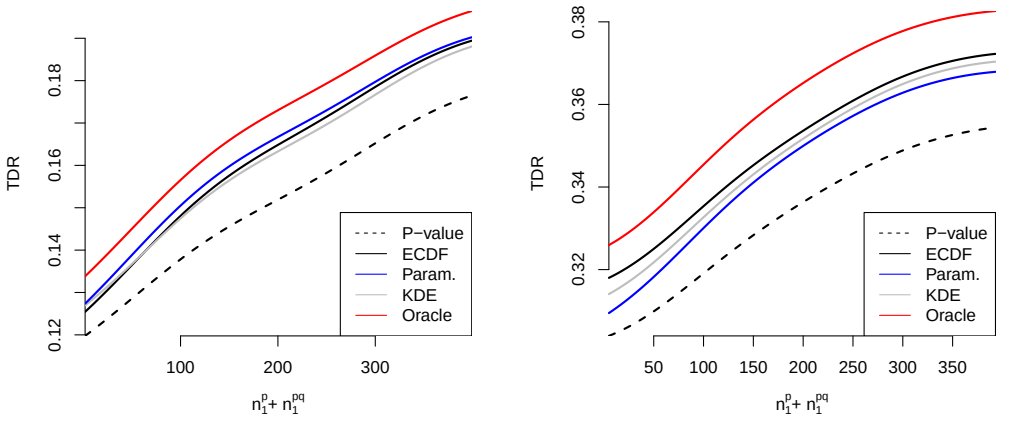

Figure 14: TDR of various methods against $n_1^p + n_1^{pq}$, the total number of variables associated with P (the primary study under consideration), restricting to simulations in which parametric assumptions were satisfied (left panel) or were not satisfied (right panel), at FDR control level $\alpha = 0.1$. A corresponding plot with $\alpha = 0.01$ is shown in supplementary figure 18. Curves show moving weighted averages using a Gaussian kernel with SD 3/10 of the X axis range.

10 Supplementary figures

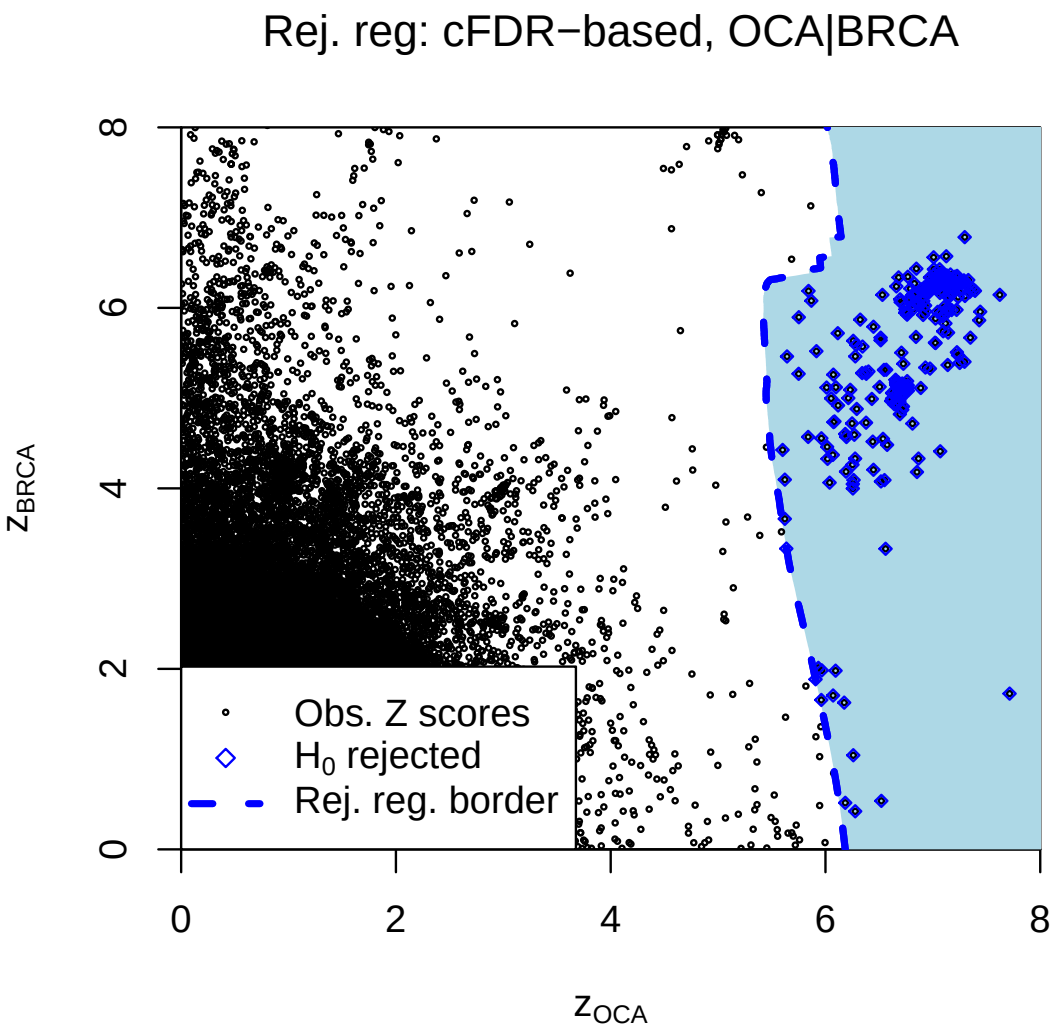

Figure 15: Association analysis of OCA using test statistics for BRCA as covariates. Variables and methods are similar to panel C in figure 1 in the main paper.

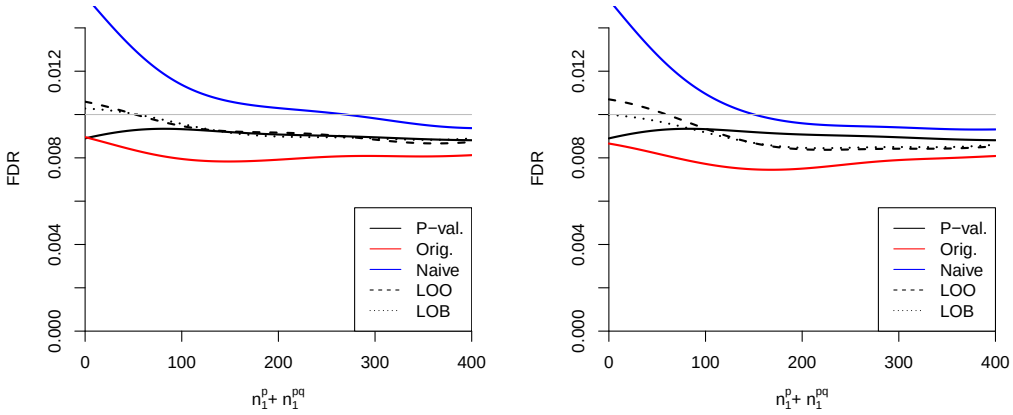

Figure 16: FDR control of various methods against $n_1^p + n_1^{pq}$, the total number of variables associated with P (the primary study under consideration). The horizontal line shows $\alpha = 0.01$, the desired FDR control level. Simulations in the left panel integrate L-regions over the the true distribution f_0; simulations in the right panel integrate over the estimated distribution as per equation (16) in the main paper. Curves show moving weighted averages using a Gaussian kernel with SD 3/10 of the X axis range.

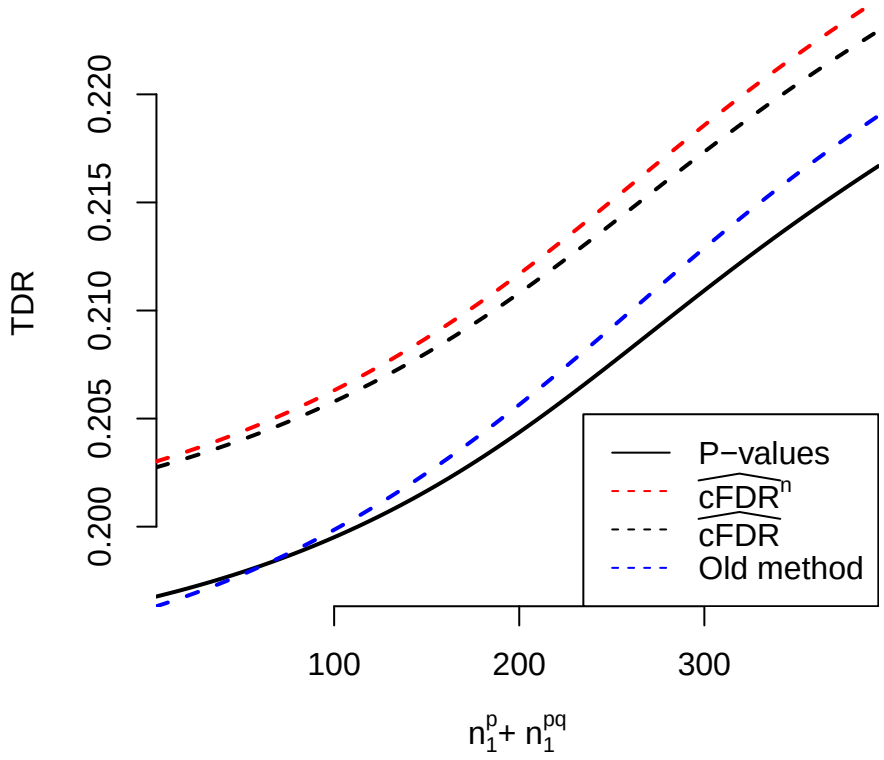

Figure 17: TDR of various methods against $n_1^p + n_1^{pq}$, the total number of variables associated with P (the primary study under consideration), at FDR control level $\alpha = 0.01$. Curves show moving weighted averages using a Gaussian kernel with SD 3/10 of the X axis range.

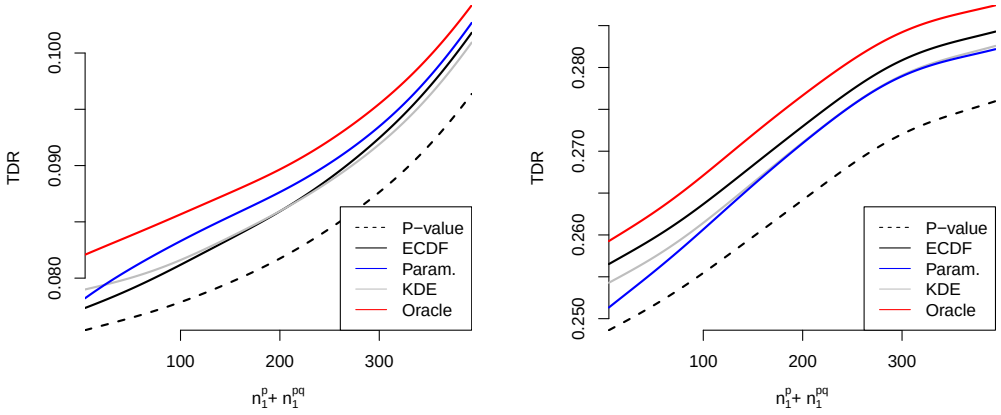

Figure 18: TDR of various methods against $n_1^p + n_1^{pq}$, the total number of variables associated with P (the primary study under consideration), restricting to simulations in which parametric assumptions were satisfied (left panel) or were not satisfied (right panel), at FDR control level $\alpha = 0.01$. Curves show moving weighted averages using a Gaussian kernel with SD 3/10 of the X axis range.

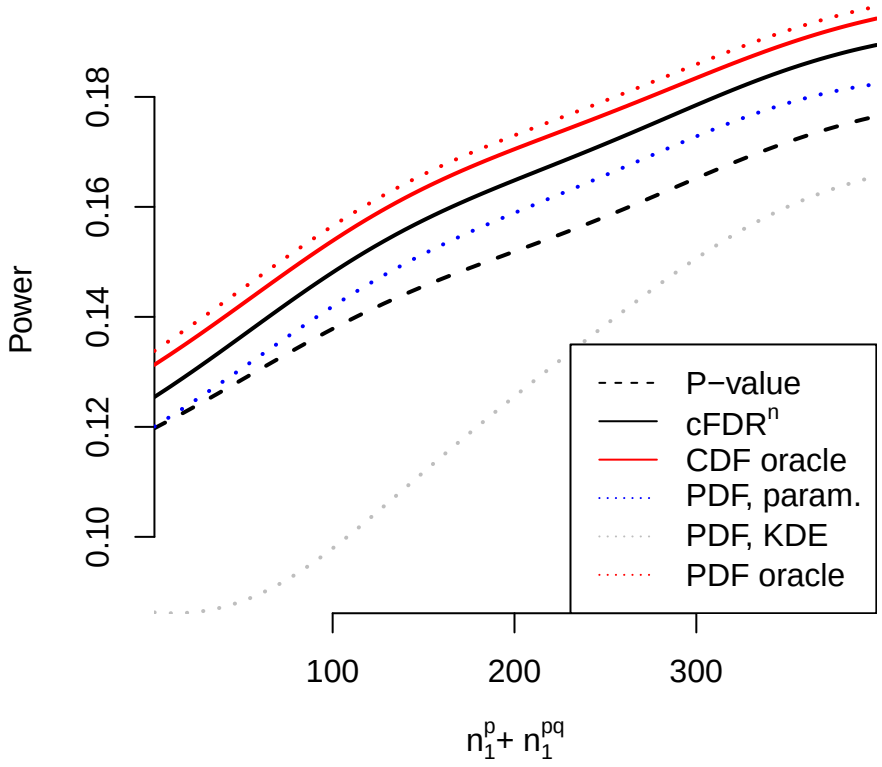

Figure 19: TDR of PDF-based methods against $n_1^p + n_1^{pq}$, the total number of variables associated with P (the primary study under consideration), restricting to simulations in which parametric assumptions were satisfied, at FDR control level $\alpha = 0.01$. Curves show moving weighted averages using a Gaussian kernel with SD 3/10 of the X axis range.

References

[1] Du L, Zhang C, et al. (2014) Single-index modulated multiple testing. The Annals of Statistics 42: 1262–1311.

[2] Alishahi K, Ehyaei AR, Shojaie A (2016) A generalized benjamini-hochberg procedure for multivariate hypothesis testing. arXiv preprint arXiv:160602386 .

[3] Lei L, Fithian W (2018) Adapt: an interactive procedure for multiple testing with side information. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80: 649–679.

[4] Bourgon R, Gentleman R, Huber W (2010) Independent filtering increases detection power for high-throughput experiments. Proceedings of the National Academy of Sciences 107: 9546–9551.

[5] Ignatiadis N, Klaus B, Zaugg JB, Huber W (2016) Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. Nature methods 13: 577.

[6] Genovese CR, Roeder K, Wasserman L (2006) False discovery control with p-value weighting. Biometrika 93: 509–524.

[7] Benjamini Y, Krieger AM, Yekutieli D (2006) Adaptive linear step-up procedures that control the false discovery rate. Biometrika 93: 491–507.

[8] Basu P, Cai TT, Das K, Sun W (2018) Weighted false discovery rate control in large-scale multiple testing. Journal of the American Statistical Association 113: 1172–1183.

[9] Cai TT, Sun W, Wang W (2016) Cars: Covariate assisted ranking and screening for large-scale two-sample inference. Technical Report .

[10] Scott JG, Kelly RC, Smith MA, Zhou P, Kass RE (2015) False discovery rate regression: an application to neural synchrony detection in primary visual cortex. Journal of the American Statistical Association 110: 459–471.

[11] Li A, Barber RF (2016) Multiple testing with the structure adaptive benjamini-hochberg algorithm. arXiv preprint arXiv:160607926 .

[12] Li A, Barber RF (2017) Accumulation tests for fdr control in ordered hypothesis testing. Journal of the American Statistical Association 112: 837–849.

[13] Ferkingstad E, Frigessi A, Rue H, Thorleifsson G, Kong A (2008) Unsupervised empirical Bayesian multiple testing with external covariates. Annals of Applied Statistics 2: 714-735.

[14] Andreassen OA, Thompson WK, Schork AJ, Ripke S, Mattingsdal M, et al. (2013) Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. PLOS Genetics 9(4).

[15] Liu JZ, Hov JR, Folseraas T, Ellinghaus E, Rushbrook SM, et al. (2013) Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis. Nature genetics 45: 670.

[16] Andreassen OA, Djurovic S, Thompson WK, Schork AJ, Kendler KS, et al. (2013) Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors. The American Journal of Human Genetics 92: 197–209.

[17] Andreassen OA, Thompson WK, Dale AM (2013) Boosting the power of schizophrenia genetics by leveraging new statistical tools. Schizophrenia bulletin 40: 13–17.

[18] Andreassen OA, McEvoy LK, Thompson WK, Wang Y, Reppe S, et al. (2014) Identifying common genetic variants in blood pressure due to polygenic pleiotropy with associated phenotypes. Hypertension 63: 819–826.

[19] Andreassen OA, Zuber V, Thompson WK, Schork AJ, Bettella F, et al. (2014) Shared common variants in prostate cancer and blood lipids. International journal of epidemiology 43: 1205–1214.

[20] Andreassen OA, Harbo HF, Wang Y, Thompson W, Schork A, et al. (2015) Genetic pleiotropy between multiple sclerosis and schizophrenia but not bipolar disorder: differential involvement of immune-related gene loci. Molecular psychiatry 20: 207.

[21] Desikan RS, Schork AJ, Wang Y, Witoelar A, Sharma M, et al. (2015) Genetic overlap between alzheimer's disease and parkinson's disease at the mapt locus. Molecular psychiatry 20: 1588.

[22] Desikan RS, Schork AJ, Wang Y, Thompson WK, Dehghan A, et al. (2015) Polygenic overlap between c-reactive protein, plasma lipids, and alzheimer disease. Circulation 131: 2061–2069.

[23] Liley J, Wallace C (2015) A pleiotropy-informed Bayesian false discovery rate adapted to a shared control design finds new disease associations from GWAS summary statistics. PLOS Genetics .

[24] Le Hellard S, Wang Y, Witoelar A, Zuber V, Bettella F, et al. (2016) Identification of gene loci that overlap between schizophrenia and educational attainment. Schizophrenia bulletin 43: 654–664.

[25] Schork AJ, Wang Y, Thompson WK, Dale AM, Andreassen OA (2016) New statistical approaches exploit the polygenic architecture of schizophrenia—implications for the underlying neurobiology. Current opinion in neurobiology 36: 89–98.

[26] Wang Y, Bos SD, Harbo HF, Thompson WK, Schork AJ, et al. (2016) Genetic overlap between multiple sclerosis and several cardiovascular disease risk factors. Multiple Sclerosis Journal 22: 1783–1793.

[27] Lv WQ, Zhang X, Zhang Q, He JY, Liu HM, et al. (2017) Novel common variants associated with body mass index and coronary artery disease detected using a pleiotropic cfdr method. Journal of molecular and cellular cardiology 112: 1–7.

[28] McLaughlin RL, Schijven D, Van Rheenen W, Van Eijk KR, O'Brien M, et al. (2017) Genetic correlation between amyotrophic lateral sclerosis and schizophrenia. Nature communications 8: 14774.

[29] Ferrari R, Wang Y, Vandrovcova J, Guelfi S, Witeolar A, et al. (2017) Genetic architecture of sporadic frontotemporal dementia and overlap with alzheimer's and parkinson's diseases. J Neurol Neurosurg Psychiatry 88: 152–164.

[30] Smeland OB, Frei O, Kauppi K, Hill WD, Li W, et al. (2017) Identification of genetic loci jointly influencing schizophrenia risk and the cognitive traits of verbal-numerical reasoning, reaction time, and general cognitive function. JAMA psychiatry 74: 1065–1075.

[31] Smeland OB, Wang Y, Frei O, Li W, Hibar DP, et al. (2017) Genetic overlap between schizophrenia and volumes of hippocampus, putamen, and intracranial volume indicates shared molecular genetic mechanisms. Schizophrenia bulletin 44: 854–864.

[32] Smeland OB, Wang Y, Lo MT, Li W, Frei O, et al. (2017) Identification of genetic loci shared between schizophrenia and the big five personality traits. Scientific reports 7: 2222.

[33] Witoelar A, Jansen IE, Wang Y, Desikan RS, Gibbs JR, et al. (2017) Genome-wide pleiotropy between parkinson disease and autoimmune diseases. JAMA neurology 74: 780–792.

[34] Broce I, Karch CM, Wen N, Fan CC, Wang Y, et al. (2018) Immune-related genetic enrichment in frontotemporal dementia: An analysis of genome-wide association studies. PLoS medicine 15: e1002487.

[35] Karch CM, Wen N, Fan CC, Yokoyama JS, Kouri N, et al. (2018) Selective genetic overlap between amyotrophic lateral sclerosis and diseases of the frontotemporal dementia spectrum. JAMA neurology 75: 860–875.

[36] Zuber V, Jönsson EG, Frei O, Witoelar A, Thompson WK, et al. (2018) Identification of shared genetic variants between schizophrenia and lung cancer. Scientific reports 8: 674.

[37] Shadrin AA, Smeland OB, Zayats T, Schork AJ, Frei O, et al. (2018) Novel loci associated with attention-deficit/hyperactivity disorder are revealed by leveraging polygenic overlap with educational attainment. Journal of the American Academy of Child & Adolescent Psychiatry 57: 86–95.

[38] Van der Meer D, Rokicki J, Kaufmann T, Córdova-Palomera A, Moberget T, et al. (2018) Brain scans from 21,297 individuals reveal the genetic architecture of hippocampal subfield volumes. Molecular Psychiatry : 1.

[39] Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, series B (methodological) 57: 289-300.

[40] Gusev A, Ko A, Shi H, Bhatia G, Chung W, et al. (2016) Integrative approaches for large-scale transcriptome-wide association studies. Nature genetics 48: 245.

[41] Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, et al. (2017) Association analysis identifies 65 new breast cancer risk loci. Nature 551: 92.

[42] Phelan CM, Kuchenbaecker KB, Tyrer JP, Kar SP, Lawrenson K, et al. (2017) Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. Nature genetics 49: 680.

[43] Greene MH, Clark JW, Blayney DW (1984) The epidemiology of ovarian cancer. In: Seminars in oncology. Elsevier, volume 11, pp. 209–226.

[44] Zablocki RW, Schork AJ, Levine RA, Andreassen OA, Dale AM, et al. (2014) Covariate-modulated local false discovery rate for genome-wide association studies. Bioinformatics 30: 2098–2104.

[45] Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, series B (methodological) 39: 1-38.

[46] Sheather SJ, Jones MC (1991) A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society Series B (Methodological) : 683–690.

[47] Kiefer J (1961) On large deviations of the empiric d.f. of vector chance variables and a law of the iterated logarithm. Pacific journal of mathematics 11: 649–660.