ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2019)

Post-selection estimation and testing following aggregate association tests

Ruth Heller,

Tel-Aviv University, Israel

Amit Meir

University of Washington, Seattle, USA

and Nilanjan Chatterjee

Johns Hopkins University, Baltimore, USA

[Received November 2017. Final revision February 2019]

Summary. The practice of pooling several individual test statistics to form aggregate tests is common in many statistical applications where individual tests may be underpowered. Although selection by aggregate tests can serve to increase power, the selection process invalidates inference based on the individual test statistics, making it difficult to identify those that drive the signal in follow-up inference. Here, we develop a general approach for valid inference following selection by aggregate testing. We present novel powerful post-selection tests for the individual null hypotheses which are exact for the normal model and asymptotically justified otherwise. Our approach relies on the ability to characterize the distribution of the individual test statistics after conditioning on the event of selection. We provide efficient algorithms for computation of the post-selection maximum likelihood estimates and suggest confidence intervals which rely on a novel switching regime for good coverage guarantees. We validate our methods via comprehensive simulation studies and apply them to data from the Dallas Heart Study, demonstrating that single-variant association discovery following selection by an aggregate test is indeed possible in practice.

Keywords: Aggregate tests; Conditional confidence interval; Conditional *p*-value; Multiple testing; Selective inference

1. Introduction

Many modern scientific investigations involve simultaneous testing of many thousands of hypotheses. Valid testing of large numbers of hypotheses requires strict multiple-testing adjustments, making it difficult to identify signals in the data if the signal is weak or sparse. One possible remedy is to pool groups of related test statistics into aggregate tests. This practice reduces the amount of multiplicity correction that needs to be applied and may assist in identifying weak signals that are spread over a number of test statistics. However, once an 'interesting' group of hypotheses has been identified, it is also often of interest to perform inference within the group to identify the individual units that contain the signal.

In many scientific fields, there is a natural predefined grouping of features of interest. In neuroscience, functional magnetic resonance imaging studies aim to identify the locations of

Address for correspondence: Ruth Heller, Department of Statistics and Operations Research, Tel-Aviv University, Ramat Aviv, Tel-Aviv 699 7801, Israel.
E-mail: ruheller@gmail.com

 1369–7412/19/81000

activation while a subject is involved in a cognitive task. The individual null hypotheses of no activation are at the voxel level, and regions of interest can be tested for activation by aggregating the measured signals at the voxel level (Penny and Friston, 2003; Benjamini and Heller, 2007). Following identification of the regions of interest, it is meaningful to localize the signal within the region. In microbiome research, the operational taxonomic units are grouped into taxonomic classifications such as species level and genus level. The data for the individual null hypotheses of no association between operational taxonomic unit and phenotype can be aggregated to test the null hypotheses of no association at the species or genus level (Bogomolov *et al.*, 2017). Here as well, following identification of the family that is associated with the phenotype, it is of interest to identify the operational taxonomic units within the family that drive the association. In genomewide association studies, the disease being analysed may have multiple subtypes of interest. The standard analysis aim is to identify the single-nucleotide polymorphisms that are associated with the overall disease, but another important aim is to identify associations with specific subtypes of the disease (Bhattacharjee *et al.*, 2012). In genetic association studies, there is also a natural grouping of the genome, since genes are comprised of single variants. The test statistics of single variants within a gene can be aggregated into a test statistic for powerful identification of associations at the gene level (Wu *et al.*, 2011; Bhattacharjee *et al.*, 2012; Derkach *et al.*, 2014; Yoo *et al.*, 2016). Following identification at the gene level, it may be of interest to identify the single variants within the gene that drive the association.

For a single group of features, let $\vec{\hat{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_m)'$ be the estimator for the vector of parameters of interest in the group, $\vec{\beta}$. Much research has focused on developing powerful aggregate tests for selecting the groups of interest, i.e. for testing at the group level the null hypothesis that $\vec{\beta} = \vec{0}$. When $\vec{\hat{\beta}}$ has (approximately) a known covariance and a normal distribution, classical test statistics are the score, Wald and likelihood ratio statistics, all of which have an asymptotic χ^2_m-distribution. Recent examples are the works by Loftus and Taylor (2014) and Reid *et al.* (2018), which include novel tests that improve on classical tests. Other examples come from the field of statistical genetics, where many gene level tests have been recently proposed based on weighted linear or quadratic combinations of score statistics for analysing genomic studies of rare variants; see Derkach *et al.* (2014) for a review.

In this work we seek to develop methods for conducting inference on the co-ordinates of $\vec{\beta}$ following selection by an aggregate test. Failure to account for data-driven selection of any kind can lead to biased inference. For example, in linear regression, where the relationship of the predictors with a response is assumed to be linear for a group, selection by an aggregate test constrains the response vector to values for which the aggregate test p-value is below a threshold, and the post-selection distribution of the data is no longer multivariate normal but a truncated multivariate normal distribution. Generally speaking, ignoring the selection will result in biased inference if there is dependence between the selection event and the individual test statistic: if the individual test statistic contributes to the selection by the aggregate test, then conditionally on being selected the distribution of the individual test statistic is changed.

Inference following selection is an emerging field of research, which is of great interest both in the statistics community and in application fields. In the multiple-testing literature, Benjamini and Bogomolov (2014) presented a novel approach for the problem of inference within families of hypotheses following selection of the families by any selection rule. Marginal confidence intervals following selection are addressed in generality in Benjamini and Yekutieli (2005), from a Bayesian perspective in Yekutieli (2012) and for a specific selection rule (i.e. that the test statistic is larger than a certain threshold) in Weinstein *et al.* (2013). Significant progress has also been made in the regression context, where variables are first selected into the model, and inference on the selected variables follows. Failing to account for the data-driven variable-selection process

invalidates the inference (Pötscher, 1991; Berk *et al*., 2013; Fithian *et al*., 2014). Recent valid post-model selection procedures can be found in Berk *et al*. (2013), Fithian *et al*. (2014), Lee and Taylor (2014), Lee *et al*. (2016), Taylor and Tibshirani (2018), Tian and Taylor (2015) and Meir and Drton (2017). Loftus and Taylor (2014) addressed post-model-selection at the group level: first, insert groups of variables into the regression model by a model selection procedure; second, test each of the groups of variables in the model selected.

In this paper, we concentrate on identifying the single variables that drive the signal in each group, which is natural in the motivating applications. Rather than starting by model building at the level of groups of predictors, as in Loftus and Taylor (2014), we select groups by marginally testing for association of each group with the outcome. Following group selection, we provide post-selection inference at the level of individual variables. Since we consider one group at a time, we assume that the sample size exceeds the number of variables in each group. Recently, Heller *et al*. (2018) have addressed the problem of post-selection inference following group level testing, in the setting where the individual test statistics are independent. Our contributions in this work are as follows.

(a) We leverage the polyhedral lemma of Lee and Taylor (2014) and the work of Fithian *et al*. (2014) on optimal testing following model selection to generalize the work of Heller *et al*. (2018) to allow for (approximately) known dependence across the individual test statistics. In particular, this enables valid testing of predictors in a generalized linear model that was selected via an aggregate test.
(b) We propose a novel approach for post-selection inference with theoretical guarantees, which is based on conducting inference under a conservative parameterization. For example, for aggregate testing with a quadratic Wald test we show that the p-values that are computed under the global null are conservative. We propose to combine p-values that are computed under the global null with p-values computed by using the polyhedral lemma into *hybrid* p-values that have improved power when the signal in the data is weak or sparse.
(c) We develop methods for computing selection-adjusted point estimates based on the conditional likelihood of the data. In particular, we show that, in the special case of aggregate testing with a Wald test or a linear aggregate test, the problem of computing the multivariate conditional maximum likelihood estimator (MLE) can be cast as a simple line search problem.
(d) We discuss computation of post-selection confidence intervals which are based on the inversion of the post-selection tests, and we propose novel regime switching post-selection confidence intervals that adapt to the unknown underlying distribution of the data.

The paper is organized as follows. In Section 2 we formally introduce our inference framework and goals. We develop theory for post-selection testing and estimation in Section 3 and Section 4 respectively. We conduct empirical evaluation of our test statistics and post-selection estimates in Section 5. In Section 6, we apply our methods to a genomic application. Finally, Section 7 concludes.

2. The set-up and the inferential goals

Suppose that $\hat{\vec{\beta}}$ is generated from a multivariate normal distribution with mean $\vec{\beta}$ and known covariance $\mathbf{\Sigma}$:

$$\hat{\vec{\beta}} \sim N_m(\vec{\beta}, \mathbf{\Sigma}).$$

Our goal is to infer on $\vec{\beta} \in \mathcal{R}^m$ following selection by an aggregate test for the global null hypothesis that $\vec{\beta} = \vec{0}$.

The global null hypothesis is tested with a quadratic test of the form $S = \hat{\vec{\beta}}' \mathbf{K} \hat{\vec{\beta}} > S_{1-t_1}$, where $\mathbf{K}$ is a semipositive definite matrix and S_{1-t_1} is the $(1-t_1)$-quantile of the null distribution of S. Setting $\mathbf{K} = \boldsymbol{\Sigma}^{(-1)}$ results in the well-known Wald test statistic. Setting $\mathbf{K} = \boldsymbol{\Sigma}^{(-1)} \mathbf{W} \boldsymbol{\Sigma}^{(-1)}$, where $\mathbf{W}$ is a diagonal matrix of weights, results in the sequence kernel association test SKAT (Wu *et al.*, 2011) that is commonly used for rare variant testing. The developments when group selection is by a linear aggregate test $S = a' \hat{\vec{\beta}}$ are similar and detailed in the on-line supplementary appendix I.

The value of t_1 comes from the analysis at the group level. In genetic association studies, the individual units for analysis are the single variants, and they are naturally grouped into the same functional unit. For example, a gene coding region can be extended upstream and downstream to incorporate additional functional elements and the regulatory region (Morris and Zeggini, 2010). When the group is the gene, then typically $t_1 \approx \alpha/20000$. This is because the Bonferroni procedure is commonly used for identifying genes that are associated with phenotypes using aggregate tests, so the familywise error rate FWER on the family of the order of 20000 genes is controlled at level α.

Given that an aggregate test has been rejected at a level t_1, our aim is to make inference on the parameters $\beta_1, \ldots, \beta_m$. For $j \in \{1, \ldots, m\}$, let

$$H_j : \beta_j = 0.$$

Our first aim is to test the family of hypotheses $\{H_j : j = 1, \ldots, m\}$ with FWER- or false discovery rate FDR-control following group level selection, i.e. following the event that the global null p-value for the group was t_1 or less. Let V and R be the number of false and total rejections in the selected group respectively. The conditional FWER and FDR are respectively $E[I\{V > 0\} | S > S_{1-t_1}]$ and $E[V/\max\{R, 1\} | S > S_{1-t_1}]$, where '$I\{\cdot\}$' is the indicator function. We provide procedures for conditional FWER- or FDR-control in Section 3.

Our second aim is to estimate the magnitude of the regression coefficients $\beta_1, \ldots, \beta_m$ given selection. Denoting the likelihood for $\vec{\beta}$ by $\mathcal{L}(\vec{\beta})$, the conditional likelihood can be written as

$$\mathcal{L}(\vec{\beta} | S > S_{1-t_1}) = \frac{\mathcal{L}(\vec{\beta})}{P_{\vec{\beta}}(S > S_{1-t_1})} I\{S > S_{1-t_1}\}.$$

We propose to use the maximizer of the conditional likelihood as a point estimate, and we show how to obtain it, as well as associated confidence intervals, in Section 4.

In the following example, as a specific illustration, we demonstrate how the above general set-up can be applied to genetic association studies.

2.1. Example 1: generalized linear models

Suppose that we observe a response vector $\vec{y} = (y_1, \ldots, y_n)' \in \mathcal{R}^n$, and m predictors of interest in a group (e.g. the single variants in a gene), $\vec{X}_j$, $j = 1, \ldots, m$. Let $\vec{V}_j$, $j = 1, \ldots, k$, be a set of additional covariates to be accounted for in the model (e.g. environmental factors or ancestry variables in genomewide association studies). Suppose that we are interested in modelling the relationship of the predictors in a group with the response vector by using a generalized linear model. So, we assume that $y_i \sim f_{\theta_i}$, an exponential family distribution with canonical parameter $\theta = g^{-1}(\eta_i) \in \Theta$ for some continuous link function $g : \Theta \to \mathcal{R}$ and

$$\vec{\eta} = \alpha_0 + \sum_{l=1}^{k} \vec{V}_l \alpha_l + \sum_{j=1}^{m} \vec{X}_j \beta_j. \tag{2.1}$$

In the case of linear regression, $g(\vec{\eta}) = \vec{\eta}$ is the identity function and $\vec{y} \sim N(\vec{\eta}, \sigma^2\mathbf{I})$. In genetic studies it is reasonable to assume that variants within a single gene, i.e. $\vec{X}_1, \ldots, \vec{X}_m$, explain little of the variance in $\vec{y}$, and σ^2 can be well estimated by the empirical variance of the residuals from the linear model with $\vec{\beta} = 0$. If the variance in $\vec{y}$ explained by the group of predictors is non-negligible, we suggest a novel variance estimator that takes selection into account in procedure 2 in Section 4.3.1.

When $\vec{y}$ is not normal, the MLE for the regression coefficients has an asymptotic normal distribution $\sqrt{n}(\hat{\vec{\beta}} - \vec{\beta}) \to^{\mathrm{D}} N\{0, \mathbf{I}^{-1}(\vec{\alpha}, \vec{\beta})\}$ and an asymptotic truncated normal distribution post selection. Although $\mathbf{I}^{-1}(\vec{\alpha}, \vec{\beta})$ depends on $\vec{\beta}$ and therefore cannot be assumed to be known in general, if $\vec{X}_1, \ldots, \vec{X}_m$ explain little of the variance in $\vec{y}$ it is reasonable to estimate the variance of $\vec{\beta}$ under the assumption that $\vec{\beta} = 0$.

3. Testing following selection

In the absence of selection, we can test for $H_j : \beta_j = 0$ by using the p-value of the test statistic $\hat{\beta}_j/\mathrm{SE}_j$: $p_j = 2\{1 - \Phi(|\hat{\beta}_j/\mathrm{SE}_j|)\}$ where $\mathrm{SE}_j = \sqrt{(\vec{e}_j' \boldsymbol{\Sigma} \vec{e}_j)}$ and $\vec{e}_j$ is the $m \times 1$ unit vector with a single entry of 1 in position $j \in \{1, \ldots, m\}$. However, conditionally on selection, P_j will often have a distribution that is stochastically smaller than uniform, meaning that its realization p_j will no longer be a valid p-value for testing H_j.

To correct for selection, it appears necessary to evaluate the probability that $S \geqslant S_{1-t_1}$. However, this probability depends on the unknown $\vec{\beta}$, and hence it cannot be evaluated when H_j is true unless we assume that all other entries in $\vec{\beta}$ are 0. In the special case of $\vec{\beta} = \vec{0}$ the distribution of $\hat{\beta}_j/\mathrm{SE}_j$, conditionally on $S \geqslant S_{1-t_1}$, is known. Of course, in practice we do not know which of the entries of $\vec{\beta}$ are non-zero.

In Section 3.1 we suggest a way around this problem, by computing valid conditional p-values by using the polyhedral lemma that was first introduced by Lee *et al.* (2016). In practice, we find that the power of the statistical tests based on the polyhedral lemma depends on the sparsity of the signal. In Section 3.2 we suggest an inference method that automatically adapts to the sparsity level of $\vec{\beta}$. In Section 3.3 we discuss applying multiple-testing procedures to the valid conditional p-values.

3.1. The conditional p-values based on the polyhedral lemma

Inferring on $\vec{\beta}$ after selection is difficult because the post-selection distribution of $\hat{\vec{\beta}}$ is not location invariant and depends heavily on the unknown parameter value. Suppose that we are interested in performing inference on an arbitrary linear function of the parameter vector $\vec{\eta}'\vec{\beta}$. Lee *et al.* (2016) showed that, by conditioning on additional information beyond the selection event, the post-selection distribution of $\vec{\eta}'\hat{\vec{\beta}}$ can be reduced to a univariate truncated normal distribution which depends only on the single unknown parameter $\vec{\eta}'\vec{\beta}$. Furthermore, Fithian *et al.* (2014) showed that such conditioning yields a (unique) family of uniformly most powerful unbiased (UMPU) selective tests on linear contrasts of the regression coefficients, $\vec{\eta}'\vec{\beta}$ (referred to in Fithian *et al.* (2014) as inference under the saturated model).

Denote by $\mathrm{TN}(\beta, \sigma^2, \mathcal{A})$ the truncated normal distribution constrained to $\mathcal{A} \subseteq \Re$, i.e. the conditional distribution of an $N(\beta, \sigma^2)$ random variable conditionally on its being in $\mathcal{A}$. Let $F^{\mathcal{A}}_{\beta,\sigma^2}$ be the cumulative distribution function of $\mathrm{TN}(\beta, \sigma^2, \mathcal{A})$. The following theorem, which is an application of the polyhedral lemma of Lee *et al.* (2016), provides us with a conditional distribution of any linear contrast of $\hat{\vec{\beta}}$ that we can use for post-selection inference.

Theorem 1. Let $\vec{\eta}'\hat{\vec{\beta}}$ be a linear combination of $\hat{\vec{\beta}}$, and $t_1 \in (0, 1]$ a fixed selection threshold. Let $\vec{W} = (\mathbf{I}_m - \vec{c}\vec{\eta}')\hat{\vec{\beta}}$, where $\vec{c} = (\vec{\eta}'\mathbf{\Sigma}\vec{\eta})^{-1}\mathbf{\Sigma}\vec{\eta}$ and $\mathbf{I}_m$ is the $m \times m$ identity matrix. Then

$$\vec{\eta}'\hat{\vec{\beta}} | S \geqslant S_{1-t_1}, \vec{W} \sim \mathrm{TN}\{\vec{\eta}'\vec{\beta}, \vec{\eta}'\mathbf{\Sigma}\vec{\eta}, \mathcal{A}(\vec{W})\}, \tag{3.1}$$

where $\mathcal{A}(\vec{W})$ is defined in lemma A.1.

See the on-line supplementary appendix A for a proof.

Since the only unknown parameter in the truncated distribution of result (3.1) is $\vec{\eta}'\vec{\beta}$, it is straightforward to compute a p-value under the null hypothesis and to construct confidence intervals via test inversion. Let $F^{\mathcal{A}}_{\vec{\eta}'\vec{\beta}, \vec{\eta}'\mathbf{\Sigma}\vec{\eta}}$ be the cumulative distribution function of $\vec{\eta}'\vec{\beta} | S \geqslant S_{1-t_1}, \vec{W}$.

Corollary 1. For testing the null hypothesis $\vec{\eta}'\vec{\beta} = 0$:

(a) for a right alternative ($\vec{\eta}'\vec{\beta} > 0$), the conditional p-value is

$$P' = 1 - F^{\mathcal{A}}_{0, \vec{\eta}'\mathbf{\Sigma}\vec{\eta}}(\vec{\eta}'\hat{\vec{\beta}}); \tag{3.2}$$

(b) for testing a two-sided alternative ($\vec{\eta}'\vec{\beta} \neq 0$) we use the UMPU selective test of Fithian *et al.* (2014) which we describe next. The level α conditional UMPU test rejects the null hypothesis when the following binary decision function is 1:

$$\phi_\alpha(t, \vec{w}) = I\{t < c_1(\alpha, \vec{w}) \text{ or } c_2(\alpha, \vec{w}) > t\},$$

where $c_1(\alpha, \vec{w})$ and $c_2(\alpha, \vec{w})$ satisfy

$$E_{\vec{\eta}'\vec{\beta}=0}[\phi_\alpha(\vec{\eta}'\hat{\vec{\beta}}, \vec{W}) | S > S_{1-t_1}, \vec{W} = \vec{w}] = \alpha,$$

$$E_{\vec{\eta}'\vec{\beta}=0}[\vec{\eta}'\hat{\vec{\beta}}\phi_\alpha(\vec{\eta}'\hat{\vec{\beta}}, \vec{W}) | S > S_{1-t_1}, \vec{W} = \vec{w}] = \alpha E_{\vec{\eta}'\vec{\beta}=0}[\vec{\eta}'\hat{\vec{\beta}} | S > S_{1-t_1}, \vec{W} = \vec{w}].$$

The conditional two-sided p-value is

$$P' = \arg\inf\{\alpha : \phi_\alpha(\vec{\eta}'\hat{\vec{\beta}}, \vec{W}) = 1\}. \tag{3.3}$$

If $\vec{\eta}'\vec{\beta} = 0$, then for P' computed by equations (3.2) or (3.3)

$$P' | S > S_{1-t_1}, \vec{W} \sim U(0, 1).$$

For details regarding the derivation of the UMPU test see Fithian *et al.* (2014). If the null conditional distribution of $\eta'\vec{\beta}$ is symmetric about zero given $S > S_{1-t_1}$ and $\vec{W}$, which is the case when $S = \hat{\vec{\beta}}'\mathbf{\Sigma}^{(-1)}\hat{\vec{\beta}}$, then the two-sided UMPU conditional p-value is

$$P' = 2\min\{1 - F^{\mathcal{A}}_{0, \vec{\eta}'\mathbf{\Sigma}\vec{\eta}}(\eta'\hat{\vec{\beta}}), F^{\mathcal{A}}_{0, \vec{\eta}'\mathbf{\Sigma}\vec{\eta}}(\eta'\hat{\vec{\beta}})\}.$$

Confidence regions inverting UMPU selective tests were termed uniformly most accurate unbiased by Fithian *et al.* (2014).

3.1.1. Example 2: effect of conditioning on $\vec{W}$ for an orthogonal design matrix

Let $\hat{\vec{\beta}} \sim N(\vec{\beta}, \mathbf{I}_m)$ and suppose that we are interested in testing $H_1 : \beta_1 = 0$ after rejecting the global null hypothesis that $\vec{\beta} = 0$. In this case, the relevant contrast is $\vec{\eta} = \vec{e}_1$ and the orthogonal projection is $\vec{W}_1 = (0, \hat{\beta}_2, \ldots, \hat{\beta}_m)$. It is clear that $\hat{\beta}_1$ is independent of $\vec{W}_1$ and therefore, conditionally on selection, the only relevant information that is contained in $\vec{W}_1$ is that $\hat{\beta}_1^2 > S_{1-t_1} - \Sigma_{j=2}^m \hat{\beta}_j^2$ so $\mathcal{A} = \{b : b^2 - S_{1-t_1} + \Sigma_{j=2}^m \hat{\beta}_j^2 > 0\}$. If we do not condition on $\vec{W}_1$ then $\hat{\beta}_1 | S > S_{1-t_1}$ has support

$\Re$, but its null distribution depends on the unknown parameters $(\beta_2, \ldots, \beta_m)$. By conditioning on $\vec{W}_1$, the support of $\hat{\beta}_1$ is still $\Re$ if the signal in the other co-ordinates is strong, i.e. if $\Sigma_{j=2}^m \hat{\beta}_j^2 > S_{1-t_1}$, and then the conditional p-value for $\hat{\beta}_1$, p_1', is identical to its original p-value p_1. However, if $\Sigma_{j=2}^m \hat{\beta}_j^2 < S_{1-t_1}$ then the support of $\hat{\beta}_1$ is $\mathcal{A} \subset \Re$ and $p_1' > p_1$. The inflation $p_1' - p_1$ will be larger the smaller $\Sigma_{j=2}^m \hat{\beta}_j^2$ (and hence $\mathcal{A}$) is. So, if $\hat{\beta}_1$ is the only variant driving the selection of the region, we may have low power to discover it following selection because p_1' is likely to be much larger than p_1.

3.2. A hybrid conditional p-value

Our empirical investigation in Section 5 suggests that p-values that are computed on the basis of the polyhedral lemma tend to have good power when $\vec{\beta}$ is not sparse or has a large magnitude. However, when only a single entry in $\vec{\beta}$ is non-zero, p-values based on the polyhedral lemma (which are valid for any configuration of the unknown $\vec{\beta}$) tend to be considerably less powerful than p-values that are computed on the basis of the distribution of $\vec{\beta}$ under the global null distribution ($\vec{\beta} = \vec{0}$). Therefore, we would like to consider a test that adapts to the unknown sparsity of the signal, by combining the two approaches for computing p-values into a single test of H_j, allowing for powerful identification of the non-null coefficients. The combined test will be useful in applications where multiple groups are analysed together, some of which have sparse signals and some of which have non-sparse signals.

Sampling from the truncated multivariate normal distribution is a well-studied problem; see for example Pakman and Paninski (2014). Specifically, under the global null, i.e. $\vec{\beta} = \vec{0}$, we can use samples from the truncated distribution to assess the likelihood of the observed regression coefficients. Let

$$p'_{j,\mathrm{GN}} = \mathrm{Pr}_{\vec{\beta}=\vec{0}}(P_j \leqslant p_j | S > S_{1-t_1}) = \frac{1}{t_1} \mathrm{Pr}_{\vec{\beta}=\vec{0}}(P_j \leqslant p_j, S > S_{1-t_1}), \tag{3.4}$$

$j = 1, \ldots, m$. When $\vec{\beta} = \vec{0}$, both P'_j (computed in expression (3.2) with $\vec{\eta} = \vec{e}_j$) and $P'_{j,\mathrm{GN}}$ have a uniform distribution. However, their distribution differs when $\vec{\beta} \neq 0$. If the only non-zero predictor in the model is the jth predictor, the test based on $P'_{j,\mathrm{GN}}$ can be expected to be more powerful than the test based on P'_j. In contrast, when more than one of the co-ordinates of $\vec{\beta}$ are non-zero, $P'_{j,\mathrm{GN}}$ will often be substantially larger than the original p-value, p_j; for example, if $\hat{\beta}_j^2/\mathrm{SE}_j^2 \geqslant S_{1-t_1}$, then $p'_{j,\mathrm{GN}} = p_j/t_1$; P'_j will not suffer any loss of power due to selection if the aggregate test passes the selection threshold t_1 regardless of the value of p_j, i.e. $p'_j = p_j$, and it will clearly be smaller than $p'_{j,\mathrm{GN}}$.

Since the preference for using $p'_{j,\mathrm{GN}}$ instead of p'_j depends on the (unknown) $\vec{\beta}$, we suggest the following test that combines the two valid post-selection p-values:

$$p'_{j,\mathrm{hybrid}} = 2\min(p'_j, p'_{j,\mathrm{GN}}). \tag{3.5}$$

Clearly, $p'_{j,\mathrm{hybrid}}$ would be a valid p-value, i.e. with a null distribution that is either uniform or stochastically larger than uniform, if both p'_j and $p_{j,\mathrm{GN}}$ are valid p-values. In the previous section we indeed showed that p'_j is a valid p-value. But by the definition in equation (3.4) it is only clear that $p'_{j,\mathrm{GN}}$ is valid when $\vec{\beta} = 0$. Intuitively, for $\vec{\beta} \neq 0$, it may be reasonable to assume that $p'_{j,\mathrm{GN}}$ is conservative (i.e. has a null distribution that is stochastically larger than uniform). We shall now provide a rigorous justification.

We start with the special case that the quadratic aggregate test for selection is Wald's test. Following selection by Wald's test, $P'_{j,\,\mathrm{hybrid}}$ is a valid p-value for testing $H_j : \beta_j = 0$. This fol-

lows by showing that the marginal null distribution of $P'_{j,\mathrm{GN}}$ is at least stochastically as large as the uniform distribution, so the test based on the global null distribution where $\vec{\beta} = \vec{0}$ is conservative.

Theorem 2. If $S = \hat{\vec{\beta}}' \boldsymbol{\Sigma}^{(-1)} \hat{\vec{\beta}}$, and $\hat{\vec{\beta}} \sim N_m(\vec{\beta}, \boldsymbol{\Sigma})$, then

$$\sup_{\vec{\beta}:\beta_j=0} \Pr_{\vec{\beta}}(P'_{j,\mathrm{GN}} < x | S > S_{1-t_1}) \leqslant x, \qquad \forall\, x \in [0,1].$$

See the on-line supplementary appendix B for a proof. See the on-line supplementary appendix I for a similar result regarding inference following linear aggregate tests.

More generally, when selection is based on $S = \hat{\vec{\beta}}\, \mathbf{K} \hat{\vec{\beta}} > S_{1-t_1}$, where $\mathbf{K}$ is any positive definite symmetric matrix, we can still justify the use of $p'_{j,\,\mathrm{hybrid}}$ for testing $H_j : \beta_j = 0$ for a sufficiently large sample size. This follows since the conditional p-values under the global null are necessarily larger than the original p-values, as formally stated in the following lemma.

Lemma 1. If $\mathbf{K}$ is a positive definite matrix, $S = \hat{\vec{\beta}}'\, \mathbf{K} \hat{\vec{\beta}}$, and $\hat{\vec{\beta}} \sim N_m(\vec{\beta}, \boldsymbol{\Sigma})$, then

$$\Pr_{\vec{\beta}=\vec{0}}(\hat{\beta}_j^2 > b | S > s) \geqslant \Pr_{\beta_j=0}(\hat{\beta}_j^2 > b) \tag{3.6}$$

for arbitrary fixed $b, s > 0$.

See the on-line supplementary appendix C for a proof. Setting b to be the realized test statistic and $s = S_{1-t_1}$, $p'_{j,\mathrm{GN}} = \Pr_{\vec{\beta}=\vec{0}}(P_j \leqslant p_j \mid S > S_{1-t_1})$ is the left-hand side of inequality (3.6) and $p_j = \Pr(\chi_1^2 \geqslant \hat{\beta}_j^2 / \mathrm{SE}_j^2)$ is the right-hand side. It thus follows that

$$p'_{j,\mathrm{GN}} \geqslant p_j.$$

Asymptotically (in the number of observations) the selection event $S > S_{1-t_1}$ occurs with probability 1 regardless of the true value of β_j if $\beta_k \neq 0$ for at least one $k \neq j$. So p_j is an asymptotically valid p-value if $\beta_k \neq 0$ for at least one $k \neq j$. Since $p'_{j,\mathrm{GN}} \geqslant p_j$, it follows that $p'_{j,\mathrm{GN}}$ and $p'_{j,\mathrm{hybrid}}$ are asymptotically valid p-values for any $\vec{\beta}$.

3.3. Controlling the conditional error rate

To identify the non-null entries in $\vec{\beta}$, we can apply a valid multiple-testing procedure on the conditional p-values computed as in Section 3.1 or Section 3.2. We can then achieve conditional error control.

The Bonferroni–Holm procedure will control the conditional FWER, since the conditional p-values are valid p-values and the procedure is valid under any dependence structure among the test statistics.

A conservative procedure that will control the conditional FDR is the Benjamini–Yekutieli procedure for general dependence, which was introduced in Benjamini and Yekutieli (2001). The theoretical guarantee follows since the conditional p-values are valid p-values and the procedure is valid under any dependence structure among the test statistics.

We recommend use of the Benjamini–Hochberg (BH) procedure for conditional FDR-control. Although the BH procedure does not have proven FDR-control for general dependence among the p-values, it has been conjectured that the BH procedure controls FDR for dependences encountered in practice, involving both one-sided and two-sided hypotheses (Reiner-Benaim, 2007; Benjamini, 2008). We similarly conjecture, on the basis of extensive simulations, that the BH procedure controls FDR when applied to conditional p-values following selection. Theoretical guarantees in specific settings are given next.

If the individual test statistics are independent, as occurs when the design matrix $\mathbf{X}$ is orthogonal in the linear model, and the aggregate test statistic is monotone increasing in the absolute

value of each test statistic (keeping all others fixed), then we have a theoretical guarantee that the BH procedure on $p'_1, \ldots, p'_m$ controls the conditional FDR, even though these conditional p-values are dependent. This is a direct result of theorem 3.1 in Heller *et al.* (2018). We extend this result in the following theorem.

Theorem 3. Assume that $\hat{\vec{\beta}} \sim N(\vec{\beta}, \boldsymbol{\Sigma})$ where, for each entry with a null expectation, the covariance with all other entries is positive, i.e. if $\beta_i = 0$ then, for each $j \neq i$, $\boldsymbol{\Sigma}_{ij} \geqslant 0$. If $S = \hat{\vec{\beta}}' \boldsymbol{\Sigma}^{(-1)} \hat{\vec{\beta}}$, then the BH procedure at level α on $p'_1, \ldots, p'_m$ controls the conditional FDR at level $(m_0/m)\alpha$, where m_0 is the number of null coefficients in $\vec{\beta}$, if

(a) the p-values are one sided or
(b) $\boldsymbol{\Sigma}$ is a diagonal matrix.

See the on-line supplementary appendix D for a proof. In theorem 3 we made a positivity assumption regarding the entries of the covariance matrix corresponding to the null variables. This assumption implies that, before selection, $\hat{\vec{\beta}}$ satisfies the positive regression property of Benjamini and Yekutieli (2001). Thus, without selection, the FDR is controlled when applying the BH procedure to the original one-sided p-values (theorem 1.2 of Benjamini and Yekutieli (2001)). Theorem 3 shows that, following selection, the BH procedure on the conditional p-values has the same guarantee.

4. Estimation following selection

So far we have focused on valid testing after selection by an aggregate test. In addition to testing, it is often also desirable to assess the absolute magnitude of parameters of interest. In Section 4.1 we discuss the computation of estimators based on the conditional maximum likelihood. Beyond point estimates, valid post-selection confidence intervals can be constructed by inverting the post-selection tests that were described in Section 3. These, however, may be either underpowered in the case of confidence intervals based on the polyhedral lemma or too conservative in the case of the hybrid confidence intervals. Thus, in Section 4.2 we propose novel regime switching confidence intervals that maintain the validity and power of the hybrid method intervals while ensuring the desired level of confidence asymptotically. In Section 4.3 we discuss a heuristic for estimating the residual variance when a non-negligible percentage of the variance in the response is explained by the m predictors.

4.1. Conditional maximum likelihood estimation

Let $l(\vec{\beta})$ be the log-likelihood for $\vec{\beta}$, and $l(\vec{\beta}|S > S_{1-t_1})$ the corresponding conditional log-likelihood. Define the conditional MLE as the maximizer of the conditional likelihood:

$$\tilde{\vec{\beta}} = \arg\max_{\vec{\beta}} l(\vec{\beta}) - \log\{\Pr_{\vec{\beta}}(S > S_{1-t_1})\}. \tag{4.1}$$

For notational convenience, we suppress the dependence of $\tilde{\vec{\beta}}$ on the selection threshold t_1. Although it is difficult to compute in many practical cases, computing the conditional MLE following selection by aggregate testing is a relatively simple task. For the special case where $\mathbf{K} = \boldsymbol{\Sigma}^{(-1)}$, we can show that the MLE is given by the solution to a simple line search problem.

Theorem 4. Under the conditions of theorem 2, the conditional MLE is given by

$$\tilde{\vec{\beta}} = \arg\max_{\vec{\beta}} l(\vec{\beta}) - \log\{\Pr_{\vec{\beta}}(S > S_{1-t_1})\}$$

$$= \arg \max_{\lambda \in [0,1]} l(\lambda \hat{\vec{\beta}}) - \log\{\Pr_{\lambda \hat{\vec{\beta}}}(S > S_{1-t_1})\}$$

where $\hat{\vec{\beta}}$ is the observed value.

See the on-line supplementary appendix E for a proof.

Theorem 4 shows that the maximum likelihood estimation is reduced to maximizing the likelihood only with respect to a scalar factor. This follows when $\mathbf{K} = \boldsymbol{\Sigma}^{(-1)}$ because the distribution of S is governed by one unknown parameter. In the general case, the distribution of S is a sum of χ^2 random variables which depends on rank($\mathbf{K}$) parameters, making the optimization problem slightly more involved. So, for $\mathbf{K} \neq \boldsymbol{\Sigma}^{-1}$ we use the stochastic optimization approach of Meir and Drton (2017) to maximize the likelihood. Let

$$\vec{z}(\vec{\beta}) \sim f_{\vec{\beta}}(\hat{\vec{\beta}} | S > S_{1-t_1})$$

be a sample from the post-selection distribution of $\hat{\vec{\beta}}$ for a mean parameter value $\vec{\beta}$. Then, taking gradient steps of the form

$$\tilde{\vec{\beta}}^{t+1} = \tilde{\vec{\beta}}^{t} + \gamma_t \boldsymbol{\Sigma}^{(-1)} \{\hat{\vec{\beta}} - \vec{z}(\tilde{\vec{\beta}}^{t})\} \tag{4.2}$$

will lead to convergence to the conditional MLE as long as

$$\sum_{t=1}^{\infty} \gamma_t = \infty \quad \text{and} \quad \sum_{t=1}^{\infty} \gamma_t^2 < \infty.$$

Theorem 5. Suppose that $\hat{\vec{\beta}} \sim N(\vec{\beta}, \boldsymbol{\Sigma})$ and that inference is conducted only if $S > S_{1-t_1}$ with $S = \hat{\vec{\beta}}' \mathbf{K} \hat{\vec{\beta}}$. Then, the algorithm that is defined by equation (4.2) converges to the conditional MLE for the post-aggregate testing problem which satisfies

$$\lim_{t \to \infty} \hat{\vec{\beta}} - E_{\vec{\beta}^t}[\hat{\vec{\beta}} | S > S_{1-t_1}] = 0.$$

Proof. The result follows from the fact that the variance of the post-selection distribution of $\hat{\vec{\beta}}$ can be uniformly bounded from above by $\boldsymbol{\Sigma}/t_1$; see Meir and Drton (2017) for details. □

We discuss the topic of conditional maximum likelihood estimation in generalized linear models and the related problem of estimation after aggregate testing with a linear test in the on-line supplementary appendices H and I.

The conditional MLE is consistent assuming the following result. Suppose that we observe a sequence of regression coefficient estimates $\vec{\beta}_1, \ldots, \vec{\beta}_n, \ldots$ such that

$$\hat{\vec{\beta}}_n \sim N(\vec{\beta}, \boldsymbol{\Sigma}_n), \qquad n\boldsymbol{\Sigma}_n \text{ converges in probability.} \tag{4.3}$$

Furthermore, suppose that we perform inference on the individual co-ordinates of $\hat{\vec{\beta}}_n$ if and only if

$$S_n > S_{1-t_1}, \qquad S_n = \hat{\vec{\beta}}_n' \mathbf{K}_{\mathbf{n}} \hat{\vec{\beta}}_n. \tag{4.4}$$

The good behaviour of the conditional MLE hinges on the probability of passing the selection by the aggregate test. The lower bound on this probability is given trivially by t_1 and therefore the conditional MLE is consistent.

Theorem 6. Assume that results (4.3) and (4.4) hold. Then, the conditional MLE is consistent for $\vec{\beta}$, satisfying

$$\lim_{n\to\infty} \Pr(\|\hat{\vec{\beta}}_n - \vec{\beta}\|_\infty > \varepsilon | S_n > S_{1-t_1}) = 0, \qquad \forall\, \varepsilon > 0.$$

Proof. The result follows from the theory that was developed in the work of Meir and Drton (2017) for selective inference in exponential families and the fact that

$$\inf_{\vec{\beta}} \Pr_{\vec{\beta}}(S_n > S_{1-t_1}) = t_1, \qquad \forall n.$$

4.2. Confidence intervals following selection by an aggregate test

From theorem 1 it is clear that the truncated normal distribution can be used to construct confidence intervals post selection in a straightforward manner. However, the extra conditioning (on $\vec{W}$) may lead to wide confidence intervals relative to confidence intervals that are based on the sampled distributions, as pointed out by Tian and Taylor (2015). As an alternative, it is possible to invert a global-type test (specifically, the test with null hypothesis $\vec{\beta} = \vec{e}_j b$ for testing that $\beta_j = b$) and to construct a hybrid-type confidence interval to obtain a confidence interval with more power to determine the sign of the regression coefficients (Weinstein *et al.*, 2013).

For constructing a confidence interval at a $(1-\alpha)$-level, let $L'_j(\alpha)$ and $U'_j(\alpha)$ be the lower and upper bounds of the uniformly most accurate unbiased confidence interval for the jth variable computed on the basis of inverting the UMPU tests of Fithian *et al.* (2014) (corollary 1 with $\vec{e}'_j\vec{\beta} = b$ instead of $\vec{e}'_j\vec{\beta} = 0$). Similarly, let $L'_{\mathrm{GN},j}(\alpha)$ and $U'_{\mathrm{GN},j}(\alpha)$ be the lower and upper limit of the confidence interval for the jth variable assuming that all other entries in $\vec{\beta}$ are 0:

$$\{b\colon \alpha/2 \leqslant F_{\vec{\beta}=\vec{e}_j b}(\hat{\beta}_j | S > S_{1-t_1}) \leqslant 1-\alpha/2\},$$

where $F_{\vec{\beta}=\vec{e}_j b}(\hat{\beta}_j | S > S_{1-t_1})$ is the cumulative distribution function of $\vec{e}'_j\hat{\vec{\beta}}$ given selection, for the parameter vector $\vec{\beta} = \vec{e}_j b$. We use the Robbins–Monro process to find $L'_{\mathrm{GN},j}$ and $U'_{\mathrm{GN},j}$ (Garthwaite and Buckland, 1992). As in testing, the polyhedral confidence interval tends to be shorter and more efficient if there are several variables in the model that are highly correlated with the response variable and the global null confidence intervals tend to be more powerful when the model is sparse or if the global null hypothesis holds (approximately). As we have done in Section 3.2, we propose a hybrid method for constructing a confidence interval, as defined by the lower and upper bounds:

$$L_{\mathrm{hybrid},j}(\alpha) = \max\{L'_j(\alpha/2), L'_{\mathrm{GN},j}(\alpha/2)\},$$

$$U_{\mathrm{hybrid},j}(\alpha) = \min\{U'_j(\alpha/2), U'_{\mathrm{GN}_j}(\alpha/2)\}.$$

The hybrid confidence intervals, although having a good degree of power to determine the sign regardless of the true underlying model, tend to be inefficient when there is strong signal in the data. To see why, consider the case of a regression model where $\beta_1, \beta_2 > 0$. Then, for a sufficiently large sample size the polyhedral confidence interval will apply no correction and the hybrid confidence interval will be conservative, with an asymptotic confidence level of $1-\alpha/2$. As a remedy, we propose a regime switching scheme for constructing confidence intervals in which we first determine whether $\|\vec{\beta}\| \approx 0$ or $\|\vec{\beta}\| \gg 0$ and then construct confidence intervals accordingly.

4.2.1. Procedure 1: post-selection level $1-\alpha$ confidence interval for β_j, with switching regime at level $t_2 < \alpha \times t_1$ (with default value $t_2 = \alpha^2 \times t_1$)

Step 1: compute $S_{1-t_2} > S_{1-t_1}$.

Step 2: if $S < S_{1-t_2}$, i.e. the aggregate test does not pass the more stringent threshold t_2, then compute the hybrid conditional confidence interval at level $1-\alpha^* = 1-(\alpha - t_2/t_1)$.
Step 3: if $S \geqslant S_{1-t_2}$, compute the unconditional confidence interval, at level $1-\alpha^* = 1-\alpha$.

Theorem 7. Post-selection confidence intervals constructed with procedure 1 have a confidence level of at least $1-\alpha$ if $\vec{\beta} = \vec{0}$, and an asymptotic level $1-\alpha$ if $\vec{\beta} \neq 0$.

See the on-line supplementary appendix F for a proof.

Remark 1. Ours is not the first regime switching procedure proposed for inference in the presence of data-driven variable selection; see for example the works of Chatterjee and Lahiri (2011) and McKeague and Qian (2015). In both these cases, one must determine whether some (or all) of the parameters are 0 and construct a test in an appropriate manner. The usual prescription for selecting tuning parameters in such procedures is to scale the tuning parameter of the test (in our case, t_2) in such a way that the correct regime is selected with probability approaching 1 as the sample size grows. In our case, this would amount to setting, for sample size n, $t_{2,n}$ in such a way that $\lim_{n\to\infty} t_{2,n} = 0$ and $S_{1-t_{2,n}} = o(n)$. However, in practice it is necessary to select a single value for t_2 and so we chose to fix t_2 at a small value to maintain a good degree of power when there is only a limited amount of signal in the data and to modify our procedure in such a way as to ensure some finite sample coverage guarantees.

4.2.2. *Example 3*

Fig. 1 shows point estimates and confidence intervals for the normal means vector which was selected via a quadratic aggregate test. Fig. 1 was generated by sampling a single data set, $\hat{\vec{\beta}} \sim N_8(\vec{\beta}, \boldsymbol{\Sigma})$ with $\Sigma_{ij} = 0.3$ for $i \neq j$ and $\Sigma_{ii} = 1$ for $i, j = 1, \ldots, 8$, $\beta_1 = -2.5$, $\beta_2 = 0.5$ and $\beta_3 = \ldots = \beta_8 = 0$. The aggregate test applied was a Wald test at an $\alpha = 0.001$-level. The naive and conditional estimates are plotted along with naive, polyhedral and hybrid 95% confidence intervals. The conditional MLE applies the same multiplicative shrinkage of 0.86 to all the co-ordinates of $\hat{\vec{\beta}}$ and so the shrinkage is more visible for the larger observed values. Because the selection is driven by $\hat{\beta}_1$ corresponding to the large negative co-ordinate β_1, the polyhedral confidence intervals for the other co-ordinates of β are similar in size to the naive confidence intervals. The naive confidence intervals overestimate the magnitude of β_1, the polyhedral confidence intervals cover the true parameter value but fail to determine its sign and the regime

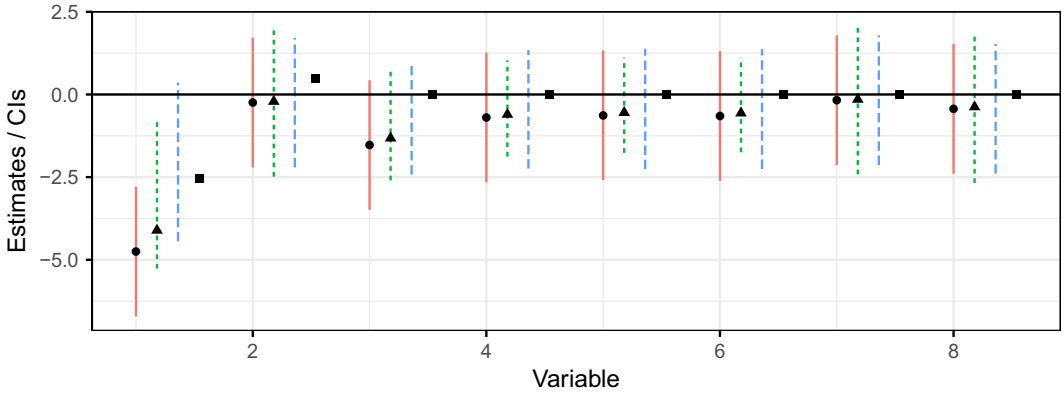

Fig. 1. Point estimates and confidence intervals for the artificial data example described in example 1: ●, naive point estimate; ▲, conditional MLE; ■, true value of the parameter; |, naive confidence intervals; :, regime switching confidence interval; ⋮, polyhedral confidence interval

switching confidence intervals both cover the true parameter value and succeed in determining the sign.

4.3. Variance estimation

So far we have assumed that the residual variance is known. This assumption is justified in settings such as rare variant testing, where $n \gg m$, n is in the thousands and the percentage of variance explained by each group of hypotheses is expected to be small and is thus well estimated by the observed variance of $\vec{y}$:

$$\hat{\sigma}^2_{\text{null}} = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

In other applications, however, it may be desirable to obtain a less conservative estimate of the residual variance. To do so, we propose a regime switching estimation method that is similar to procedure 1. Namely, we propose to conduct a secondary post-selection test; if the secondary test is rejected then we impute the naive variance estimate for the residual variance

$$\hat{\sigma}^2_{\text{naive}} = \frac{1}{n-m} \|\vec{y} - X\hat{\vec{\beta}}\|_2^2,$$

and otherwise we use the null variance $\hat{\sigma}^2_{\text{null}}$ as our estimate.

4.3.1. Procedure 2: regime switching post-selection variance estimation method with parameter t_2 (with default value $t_2 = \alpha^2 \times t_1$)

Step 1: compute $S_{1-t_2} > S_{1-t_1}$.
Step 2: if $S < S_{1-t_2}$ i.e. the aggregate test does not pass the more stringent threshold t_2, then set $\hat{\sigma}^2 = \hat{\sigma}^2_{\text{null}}$.
Step 3: if $S \geqslant S_{1-t_2}$, set $\hat{\sigma}^2 = \hat{\sigma}^2_{\text{naive}}$.

Although the naive estimator following selection will often typically be biased downwards because of selection, by accounting for selection via regime switching it is straightforward to show that, if we let $t_2 \to 0$ as the sample size grows at an appropriate rate, then using the regime switching variance estimate as a part of a valid conditional inference procedure will yield a consistent post-selection inference procedure (see remark 1).

5. Simulations

In this section we conduct a simulation study where we assess the methods that are proposed in this work and verify our theoretical findings. We conduct three types of simulation. In Section 5.1 we conduct experiments with simulated data sets aimed at mimicking genetic association studies of rare variants and analyse the behaviour of our procedures conditionally on selection at the single gene level. Then, in Section 5.2 we conduct simulations, where we generate large data sets with 5000 genes each and compare the performance of our conditional inference procedures with the Benjamini–Bogmolov (BB) procedure (Benjamini and Bogomolov, 2014) which controls the average FDR and to the BH procedure applied to all single variants without selection. Finally, in Section 5.3 we relax our assumption regarding the covariance of $\hat{\vec{\beta}}$ being (approximately) known and conduct an experiment with normal design where the covariance must be estimated from the same data as used for inference.

5.1. Single-gene experiments

In this section we aim to investigate the performance of our methods at the single-gene (group)

level conditionally on selection. In Section 5.1.1 we assess the post-selection tests that were proposed in Section 3 with respect to their ability to control FDR. In Section 5.1.2 we compare the various testing methods with respect to their power to detect true signal in the data. In Section 5.1.3 we compare the conditional MLE and the unadjusted MLE with respect to their estimation error. Finally, in Section 5.1.4 we assess the coverage rates of the polyhedral and regime switching confidence intervals.

In all of our simulations we generate data in a similar manner. We first generate a design matrix in a manner that is meant to approximate association studies of rare variants. We sample marginal expression proportions for our variants from $g_1, \ldots, g_m \sim \text{gamma}(1, 300)$ constrained to $[2 \times 10^{-4}, 0.1]$ and for each subject we sample two multivariate normal vectors $r_{1,i,.}, r_{2,i,.} \sim N_m(0, U)$ with $U_{ij} = 0.8^{|i-j|}$. We then set $X_{ij} = \Sigma_{k=1}^{2} I\{\Phi(r_{k,i,j}) \leqslant g_j\}$ to obtain a design matrix with dependent columns and a marginal distribution $X_{ij} \sim \text{Bin}(2, g_j)$. We generate a sparse regression coefficients vector with $m - s$ zero co-ordinates and s co-ordinates which are sampled from the Laplace(1) distribution. We normalize the values of the regression coefficients such that the percentage of explained variance

$$R^2 = 1 - \frac{n}{n + \vec{\beta}' \mathbf{X}' \mathbf{X} \vec{\beta}} \tag{5.1}$$

equals some prespecified value. Finally, we generate a response variable $\vec{y} = \mathbf{X}\vec{\beta} + \vec{\varepsilon}$ with $\vec{\varepsilon} \sim (0, \mathbf{I})$. In all of our simulations we use a Wald aggregate test with a significance level of $t_1 = 0.001$.

5.1.1. *Assessment of false discovery rate control*

We assess how well the testing procedures proposed control FDR under the model assumed as well as under model misspecification. We generate data sets with $m = 50$, $n = 10^4$, $s = 3$, $R^2 \in \{0, 0.001\}$ and three types of distribution for the model residuals, all of which have a variance of 1:

$$\varepsilon_i^{(1)} \sim N(0, 1),$$

$$\varepsilon_i^{(2)} \sim \text{Laplace}(\sqrt{2}),$$

$$\varepsilon_i^{(3)} \sim \text{Unif}(-\sqrt{12}/2, \sqrt{12}/2).$$

We compare four testing procedures: the BH procedure on naive p-values which are not adjusted for selection, the BH procedure on the polyhedral p-values as described in equation (3.2), the BH procedure on the p-values based on the global null distribution as described in equation (3.3) and the BH procedure on the hybrid p-values as described in equation (3.4).

We plot the empirical FDR against the nominal FDR in Fig. 2. When there is no signal in the data and the noise is not heavy tailed, all selection-adjusted methods obtain close to nominal FDR-levels (the top left-hand and right-hand panels). When the noise is heavy tailed (Laplace), the hybrid method, which is based on the global null normal distribution, has higher-than-nominal FDR-rates, whereas the p-values that are computed with the polyhedral method exhibit a more robust behaviour (the top centre panel). When there is some signal in the data, all selection-adjusted p-values control FDR at the nominal or conservative rate (the bottom row). The naive p-values do not control FDR in any of the simulation settings. Thus, we conclude that the polyhedral p-values may be preferable to the hybrid p-values if the distribution of the data is heavy tailed. However, as we show next, the hybrid method tends to have more power

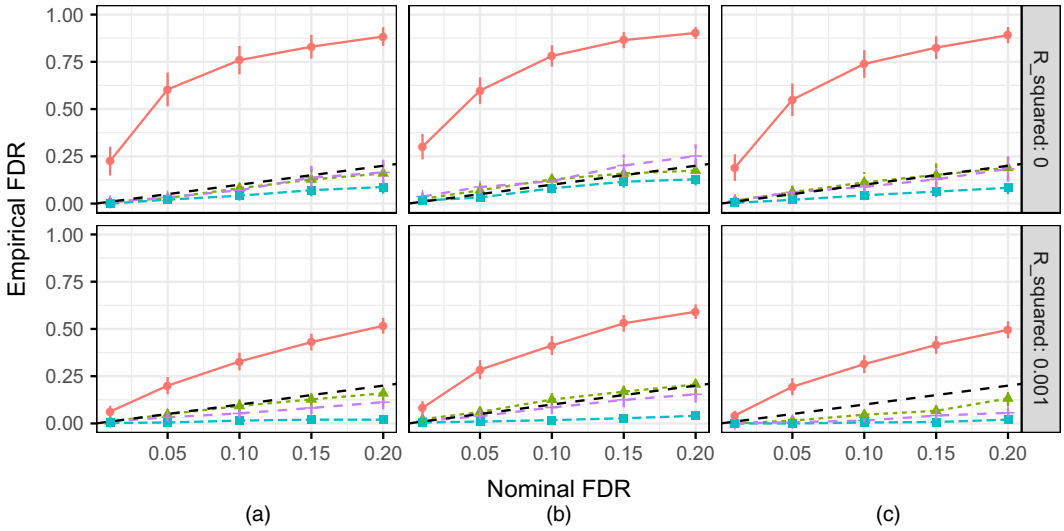

Fig. 2. False discovery rates after aggregate testing (empirical FDR *versus* nominal FDR for the unadjusted naive *p*-values (●), the polyhedral *p*-values (▲), global null *p*-values (■) and the hybrid *p*-values (+) (– – –, diagonal line) (the figure is separated according to the distribution of the noise and the percentage of variance explained by the gene, as defined in equation (5.1); details regarding the data generation process are in Section 5.1)): (a) Gaussian noise; (b) Laplace noise; (c) uniform noise

compared with the polyhedral method in sparse settings and may thus be preferable when the residual distribution is well behaved.

5.1.2. Assessment of power to detect true signal

To compare the power to detect the signal of the testing procedures proposed, we generate data sets with $m = 50$, $n = 10^4$, $s \in \{1, 2, 4, 8\}$, $R^2 \in \{0.001, 0.004, 0.016, 0.062\}$ and $\varepsilon_i \sim N(0, 1)$. The BH procedure is applied at the nominal 0.05 FDR level. We compare the same testing procedures as in Section 5.1.1.

We plot the results of the simulation in Fig. 3. In all of the simulations the naive unadjusted *p*-values have the most power, at the cost of an inflated FDR. The method based on the global null distribution is the most powerful when the signal is sparse and low. The polyhedral method has more power when the signal is not too sparse or low. The hybrid method seems to adapt to the sparsity and signal strength well, exhibiting comparatively good power in all settings.

5.1.3. Assessment of estimation error

In this section, we compare the conditional MLE with the naive, unadjusted point estimate $\hat{\vec{\beta}}$ itself. We set $n \in \{5000, 10000, 15000, 20000\}$, $m \in \{5, 10, 20\}$, $s = 2$, $R^2 \in \{0, 0.000003\}$ and sample model residuals from the normal distribution with a standard deviation of 1.

We plot the results of the simulation in Fig. 4. When the dimension of $\vec{\beta}$ is small, the conditional MLE estimates the vector of regression coefficients better than the unadjusted MLE. The gap between the conditional and naive estimator is roughly constant across the different sample sizes when $\vec{\beta} = \vec{0}$ because the probability of selection remains constant for all sample sizes. However, when there is some signal in the data the probability of passing the aggregate test increases in the sample size and the gap between the estimators shrinks. The difference between the conditional MLE and the naive MLE decreases in the size of $\vec{\beta}$, to the extent that for $m = 20$ the two estimators are indistinguishable from one another.

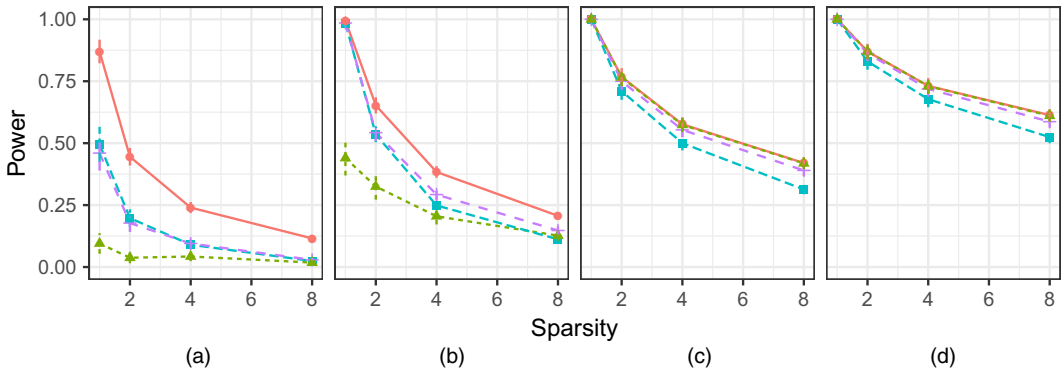

Fig. 3. Power to detect true signals after aggregate testing (we plot the power of the various inference methods as a function of the number of non-zero regression coefficients for the unadjusted naive p-values (●), the polyhedral p-values (▲), p-values based on the exact post-selection null distribution (■) and the hybrid method (+) (the figure is separated according to the strength of the signal as defined in equation (5.1); details regarding the data generation process are in Section 5.1.2)): (a) $R^2 = 0.001$; (b) $R^2 = 0.004$; (c) $R^2 = 0.016$; (d) $R^2 = 0.062$

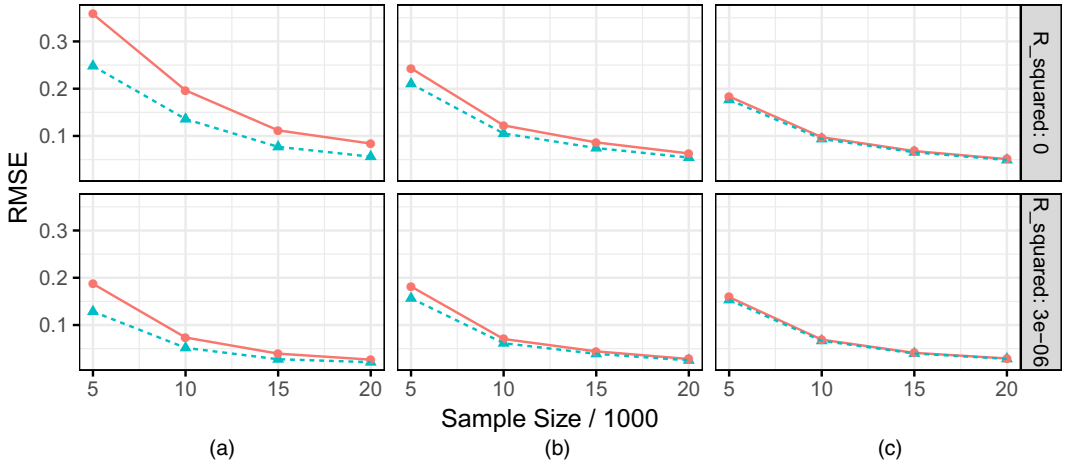

Fig. 4. Root-mean-squared error RMSE for estimation after aggregate testing (we plot RMSE for estimating the vector of regression coefficients $\vec{\beta}$ with the naive unadjusted estimator $\hat{\vec{\beta}}$ (●) and the conditional MLE $\tilde{\vec{\beta}}$ (▲) (the figure is separated according to the signal level as defined in equation (5.1) and the size of $\vec{\beta}$, m)): (a) dimension 5; (b) dimension 10; (c) dimension 20

To see why this occurs, consider the following example. Let $\vec{y} \sim N_m(0, \mathbf{I})$, suppose that we perform selection using a Wald test at a fixed level t_1 and consider the conditional log-likelihood function

$$l(\vec{y}) \propto -\frac{1}{2}\sum_{i=1}^{m}(y_i - \beta_i)^2 - \log\{\Pr_{\vec{\beta}}(S > S_{1-t_1})\}$$

As we let the dimension m grow, the decrease in the value of the (unconditional) Gaussian log-likelihood due to a possible shrinkage of β grows linearly in m. At the same time, the additional penalty term $-\log\{\Pr_{\beta}(S > S_{1-t_1})\}$ remains bounded below by $-\log(t_1)$ regardless of the dimension of the problem.

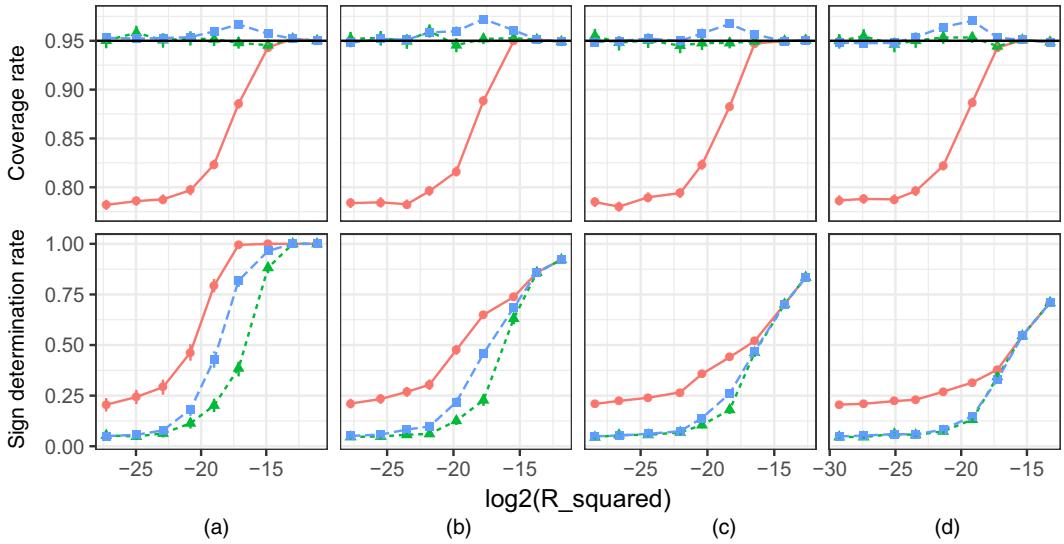

Fig. 5. Coverage rates (top row) and power to determine the sign (bottom row) of confidence intervals constructed after aggregate testing for the naive unadjusted confidence intervals (●), polyhedral confidence intervals (▲) and regime switching intervals with $t_2 = t_1\alpha^2$ (■): the number of non-null coefficients ranges from (a) 1, (b) 2 and (c) 4 to (d) 8

5.1.4. Assessment of confidence interval coverage rates

Here we evaluate the regime switching and polyhedral confidence intervals with respect to their coverage rates and power to determine the sign of the non-zero coefficients. We set the parameters of the simulation to $m = 20$, $n = 10^4$, $s \in \{1, 2, 4, 8\}$ and $R^2 \in [6 \times 10^{-9}, 0.0001]$, and sample the residuals from the normal distribution.

We plot the results of the simulation in Fig. 5. The naive confidence intervals have a coverage rate that is far below nominal when R^2 is very small. As could be expected, the polyhedral method achieves the correct coverage rates up to Monte Carlo error in all the simulation settings. When there is very little signal in the data the regime switching confidence intervals have close to nominal coverage. When the percentage of explained variance is moderate, the regime switching confidence intervals are conservative because the polyhedral confidence intervals are superior to those based on the global null assumption with high probability whereas the probability that S exceeds S_{1-t_2} is still not overwhelmingly large. When R^2 is larger the regime switching confidence intervals are mostly identical to the naive intervals, because the selection occurs with probability close to 1, and so they have the correct coverage rate. Despite being more conservative than the polyhedral confidence intervals, the regime switching confidence intervals can have better power to determine the sign. Specifically, the regime switching confidence intervals tend to have more power when the true model is sparse and the signal-to-noise ratio is low or moderate.

5.2. Simulations mimicking genomewide association studies

So far we have studied post-selection inference for a single selected group. We now turn to the practical case of applying our methods to data sets with many groups of hypotheses. Specifically, we generate data sets with 5000 genes where each gene is generated according to the process that was described in Section 5.1.

We vary the following parameters. We set the percentage of null genes to 0.99 or 0.995 and the percentage of variance explained by all genes to $R^2 \in \{0.1, 0.2, 0.3\}$, and we vary the percentage

of non-null variants within each non-null gene over 0.05, 0.1 and 0.2. In all simulation settings, we sample the size of each gene from $\{25, 55, 100\}$ with equal probabilities.

We consider procedures that control two types of error. The first type of error is the overall FDR as evaluated over the entire genome. Let $g \in \{1, \ldots, G\}$ index the individual genes, let V_g be the number of false rejections in the gth gene and let R_g be the total number of rejections in the gth gene. The overall FDR is given by

$$\text{overall FDR} = E\left[\frac{\sum_{g=1}^{G} V_g}{\max\left(\sum_{g=1}^{G} R_g, 1\right)}\right].$$

We compare three methods for controlling the overall FDR. The first two methods have two steps, and they differ only in the second step. In step 1, the Wald test p-value is computed for each gene, and the BH procedure is applied at the 0.05-level on the family of Wald test p-values. In step 2, the conditional p-values are computed for all variants within genes discovered in step 1, and the BH procedure is applied at the 0.05-level on the family of conditional p-values.

We compute conditional p-values using either the polyhedral method or the hybrid method. We compare the conditional methods with the BH procedure applied to all genes (without selection by aggregate testing). To do so, we first fit a linear regression to each gene separately, and apply the BH procedures to all individual p-values. As we have discussed in Section 3.3, we expect the BH procedure to control the FDR despite the dependence across the test statistics, since the nominal FDR-level of BH is maintained in a wide range of dependent settings. If a theoretical guarantee of FDR-control is desired, then it is possible to use the Benjamini–Yekutieli procedure in place of the BH procedure, but the power may be much lower.

The second type of error that we propose to control is the average FDR that was first defined by Benjamini and Bogomolov (2014). The average FDR is given by

$$\text{average FDR} = E\left[\frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}\frac{V_g}{\max(R_g, 1)}\right],$$

where $\mathcal{G}$ is the set of indices of selected genes. Benjamini and Bogomolov proposed a two-stage procedure for controlling the average FDR at a level α: first, conduct aggregate tests at the gene level to select $\mathcal{G}$; next, apply the BH procedure to the naive p-values within each selected gene at an $\alpha|\mathcal{G}|/G$-level. As an alternative to the BB procedure, we propose to apply the BH procedure to the conditional p-values within each selected gene at an α-level. The gene selection for the BB and for our approach is by step 1 above. We expect our proposed procedure to control the average FDR since the conditional FDR is controlled; see theorem 3 in Heller *et al.* (2018).

All methods were compared in terms of average power, i.e. the average (over simulated data sets) number of true rejections divided by the total number of non-null coefficients.

5.2.1. Average power of FDR-controlling procedures

In our experiments, all methods controlled FDR as expected. A detailed description of our results regarding FDR-control can be found in the on-line supplementary appendix G. We plot our results regarding power to detect signal when controlling for the overall FDR in Fig. 6.

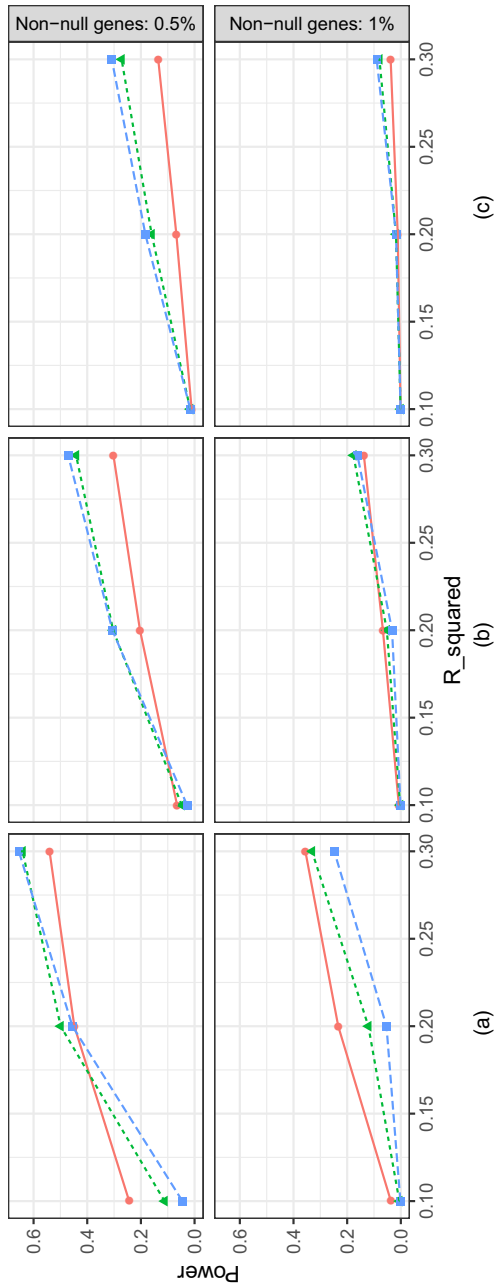

Fig. 6. Average power to detect single variants when controlling FDR *versus* R^2 (we compare three FDR-controlling methods; the first is based on applying the BH procedure to all single variants (●); the two others are based on applying the BH procedure to conditional *p*-values computed by using the polyhedral method (■) or the hybrid method (▲) in genes selected via aggregate testing; the figure is separated according to the fraction of non-zero variants within each gene and the fraction of non-zero genes): (a) non-zero variants 5%; (b) non-zero variants 10% (c) non-zero variants 20%

Applying the BH procedure to all variants yields better power to detect single variants when the signal is concentrated in a small number of variants within each gene, rendering the aggregate testing step inefficient at detecting genes that are correlated with the response. When many variants within each gene are correlated with the response our two-step inference procedures have improved power to detect signal compared with the BH procedure applied to all the variants because the aggregate testing step efficiently narrows down our attention only to genes that are correlated with the response, reducing the number of hypotheses that are tested at the second stage. All of the methods have less power when there are more non-null genes because when there are many non-null genes the same amount of signal is spread over a larger number of genes.

We note that our two-step inference procedures have the added benefit that both the gene level and the individual variant level analyses preserve their desired nominal error rate, and moreover the single-variant discoveries are always within the gene level discoveries. This is particularly useful if gene level inference is of primary interest, and single-variant discovery is of secondary importance.

5.2.2. *Average power of average-FDR-controlling procedures*

All of the average-FDR-controlling procedures have maintained proper error control in our experiments; we refer the reader to on-line supplementary appendix G for detailed results. We compare our conditional inference procedures with the BB procedure in Fig. 7. We see that using the hybrid p-values for error control obtains comparable or better power than the BB procedure in all simulation settings. The gap in power is smaller when the fraction of non-null genes is larger (more generally, as the fraction of selected genes increases, the power of the BB procedure increases, so this behaviour is expected). As we have seen in the previous section, the polyhedral p-values have more power than the hybrid method when there is a large amount of signal in the data spread over many single variants, but has lower power when the signal is weak or sparse.

5.3. *Inference with estimated covariance*

In our final set of simulations we relax our assumption of known $\mathbf{\Sigma}$. To demonstrate the sensitivity of our methods to this assumption, we work with relatively small sample sizes setting the number of covariates to $m = 20$ and varying the sample size $n \in \{50, 100, 200\}$. We sample the rows of the design matrix from the multivariate normal distribution $\vec{X}_i \sim N_{20}(0, U)$ with an auto-regressive covariance $U_{st} = 0.5^{|s-t|}$. We set the number of non-zero regression coefficients to 3 and test a large number of explained variance values R^2 ranging from 0 to 0.75.

To investigate how well our inference procedures control the type I error rate conditionally on selection, we selected the group if the Wald test p-value was at most 0.01, and we applied the BH procedure to the polyhedral or hybrid conditional p-values at a nominal FDR-level of 0.05. We note that for small sample sizes ($n = 50$) the empirical type I error rate of the Wald aggregate test was considerably higher than the desired nominal level (which was about 0.1 instead of 0.01 for $R^2 = 0$). For conditional inference, we estimate the residual variance σ^2 by using the adaptive estimation method described in Section 4.3. For comparison, we also compute the conditional p-values based on the true design matrix that was used to generate the data (so $\mathbf{\Sigma}$ is known up to σ^2).

We plot the simulation results in Fig 8. As we have seen before, when we use our knowledge regarding the true variance, both the polyhedral and the hybrid methods control the conditional FDR at the desired level. However, when we estimate the covariance from the data, the polyhedral method has a slightly inflated FDR (approximately 0.06), especially when the sample

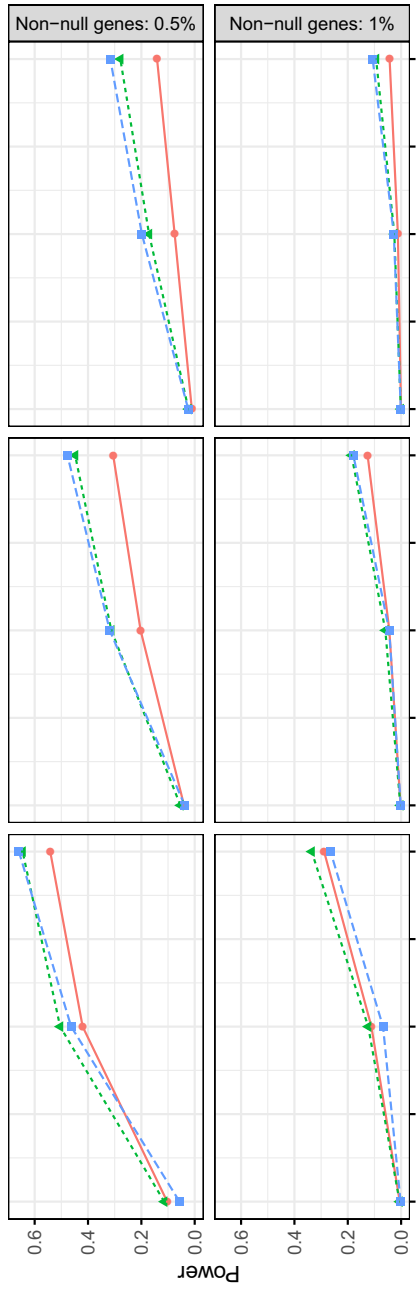

Fig. 7. Average power to detect single variants when controlling the average FDR *versus* R^2 (we compare three FDR-controlling methods; the BB method (●), the BH procedure applied to polyhedral conditional *p*-values (■) and the BH procedure applied to hybrid conditional *p*-values (▲) (the figure is separated according to the fraction of non-zero variants within each gene and the fraction of non-zero genes)): (a) non-zero variants 5%; (b) non-zero variants 10%; (c) non-zero variants 20%

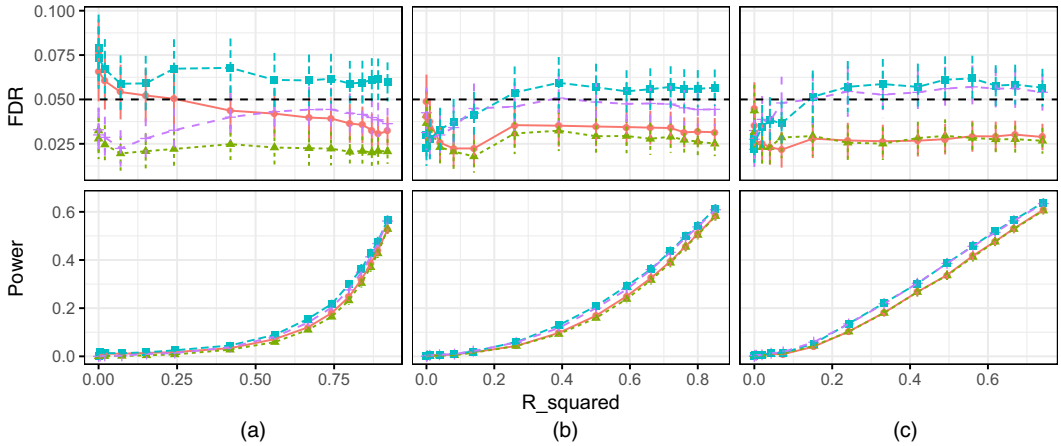

Fig. 8. FDR and power of conditional inference procedures when the covariance is estimated as a function of R^2 (we compare FDR and power of the polyhedral and hybrid methods when the covariance is known (●, polyhedral; ▲, hybrid) and when the covariance is estimated from the data (■, polyhedral; +, hybrid), the hypothesis groups were selected on the basis of a level 0.01 Wald test, and the conditional FDR was controlled at a nominal level of 0.05; the figure is separated according to the power or FDR-control, and the sample size) (|, |, |, |, 2-standard-deviation confidence intervals for the Monte Carlo estimates of the error rates): (a) sample size 50; (b) sample size 100 (c) sample size 200

size is small. The need to estimate the covariance has only a small effect on the power of our methods. For very small sample sizes, using the empirical covariance leads to slightly higher power at the cost of slightly elevated FDR-levels, but when $n = 200$ there are no noticeable differences in power.

6. Application to variant selection following gene level testing

Large genomewide association studies of uncommon and rare variants are now becoming increasingly feasible with the advent of newer genotyping chips, cheaper sequencing technologies and sophisticated algorithms that allow imputation of low frequency variants based on combinations of common variants that are already genotyped in large genomewide association studies. Thus, association studies of rare variants are a very active area of research and some of the early studies have already begun to report their findings, e.g. UK10K Consortium and University College London–Edinburgh–Bristol Consortium (2015) and Fuchsberger *et al.* (2016).

As the statistical power for testing association of traits with individual rare variants may be low, it has been suggested that tests for genetic associations be performed at an aggregate level by combining signals across multiple variants within genomic regions such as those defined by functional units of genes (Goeman *et al.*, 2006; Madsen and Browning, 2009; Morris and Zeggini, 2010; Neale *et al.*, 2011; Wu *et al.*, 2011; Lee *et al.*, 2012; Sun *et al.*, 2013). There is, however, currently a lack of rigorous methods for variant selection following gene level association testing.

The Dallas Heart Study (Romeo *et al.*, 2007) considered four genes of potential interest, genotyped in 3549 individuals (601 Hispanic, 1830 non-Hispanic black, 1043 non-Hispanic white and 75 other ethnicities). We focus on the 32 variants in the ANGPTL4 gene, which includes both rare and common variants. The second column of Table 1 shows the number of subjects with rare variants.

Table 1. For the 32 variants in ANGPTL4, the number of subjects with rare variants (second column), the estimated effect size (third column), the hybrid and polyhedral conditional p-value (fourth and fifth columns), the original p-value (sixth column), the default beta density weight in SKAT (seventh column) and the contribution of the variant to the SKAT statistic $\Sigma_{j=1}^{32} w_{m,j} U_j^2$ (eighth column)†

Variant	*Number of rare variants*	$\hat{\beta}$	*Hybrid p-value*	*Polyhedral p-value*	*Naive p-value*	*SKAT weight*	$w_{m,j}U_j^2$
M1T	1.0000	0.8967	0.6872	0.3434	0.3434	4.9831	19.9657
P5L	2.0000	0.8588	0.3999	0.1995	0.1995	4.9663	74.7223
E40K	50.0000	−0.4490	0.0020	0.0010	0.0010	4.2172	7687.4667
M41I	28.0000	0.1403	0.8860	0.4333	0.4333	4.5466	288.0531
S67R	2.0000	0.3681	1.000	0.5827	0.5827	4.9663	15.9910
R72L	3.0000	−0.3468	1.000	0.5262	0.5262	4.9495	22.0274
G77R	1.0000	−1.0913	0.4928	0.2489	0.2489	4.9831	29.5739
E167K	1.0000	−1.2002	0.4072	0.2049	0.2049	4.9831	34.4125
P174S	1.0000	0.1040	1.000	0.9125	0.9125	4.9831	0.4010
E190Q	32.0000	0.2140	0.4108	0.2054	0.2054	4.4850	1199.4289
E196K	1.0000	−0.3991	1.000	0.6733	0.6733	4.9831	3.5122
K217X	1.0000	−0.7300	0.8704	0.4406	0.4406	4.9831	12.4116
G223R	1.0000	−1.3242	0.3239	0.1621	0.1621	4.9831	40.5690
R230C	1.0000	−0.5784	1.000	0.5412	0.5412	4.9831	7.6586
F237V	1.0000	−0.2453	1.000	0.7956	0.7956	4.9831	1.2266
K245fs	1.0000	0.5157	1.000	0.5858	0.5858	4.9831	6.6047
P251T	1.0000	1.3860	0.2854	0.1432	0.1432	4.9831	49.3013
T266M	1887.0000	0.0230	0.8619	0.2454	0.2454	0.0006	0.0002
R278Q	207.0000	−0.1945	0.0055	0.0103	0.0023	2.4309	10667.1021
V291M	1.0000	−0.6203	1.000	0.5123	0.5123	4.9831	9.5533
L293M	1.0000	0.4021	1.000	0.6717	0.6717	4.9831	1.3213
E296V	1.0000	−0.0308	1.000	0.9741	0.9741	4.9831	0.0235
S302fs	1.0000	−0.7089	0.9012	0.4539	0.4539	4.9831	12.4794
P307S	1.0000	−0.0793	1.000	0.9333	0.9333	4.9831	0.0274
V308M	3.0000	1.4719	0.0143	0.0071	0.0071	4.9495	477.6972
R336C	7.0000	−0.0469	1.000	0.8957	0.8957	4.8829	2.5696
D338E	1.0000	0.2314	1.000	0.8073	0.8073	4.9831	0.0339
W349C	1.0000	−0.6120	1.000	0.5179	0.5179	4.9831	9.3008
G361R	2.0000	0.4744	0.9598	0.4784	0.4784	4.9663	23.2973
R371Q	1.0000	0.4724	1.000	0.6177	0.6177	4.9831	5.5410
R384W	1.0000	−0.7610	0.8420	0.4225	0.4225	4.9831	22.6686
G361S	1.0000	−1.0389	0.5388	0.2724	0.2724	4.9831	26.7995

†Variant E40K has conditional p-value below $0.05/32 = 0.0016$ and is therefore discovered by a Bonferroni test, with a guarantee of conditional FWER-control at the 0.05-level. The contribution of variant R278Q is by far the largest towards the SKAT statistic, and therefore the polyhedral conditional p-value is larger than the naive p-value for R278Q. For all other variants in this gene, the polyhedral conditional p-values coincide with the naive p-values. This is expected when, by conditioning on the test statistics of all the other variants, the SKAT test significance is guaranteed regardless of the single-variant test statistic value.

To detect associations with triglyceride TG, which is a metabolism trait, we applied the variance component test SKAT of Wu *et al.* (2011), with outcome TG on a logarithmic scale, while adjusting for the covariates race, sex and age on a logarithmic scale. ANGPTL4 is one of the four genes in the ANGPTL family (Romeo *et al.*, 2009). Using a Bonferroni correction for testing the genes in the family, ANGPTL4 is selected for post-selection inference if the SKAT p-value is at most 0.05/4. To identify the potentially susceptible variants, we proceeded as suggested in Section 3.

The SKAT p-value for ANGPTL4 was 7.5×10^{-5} and therefore the gene was selected. The seventh column of Table 1 lists the weights that were assigned to each variant in the SKAT test.

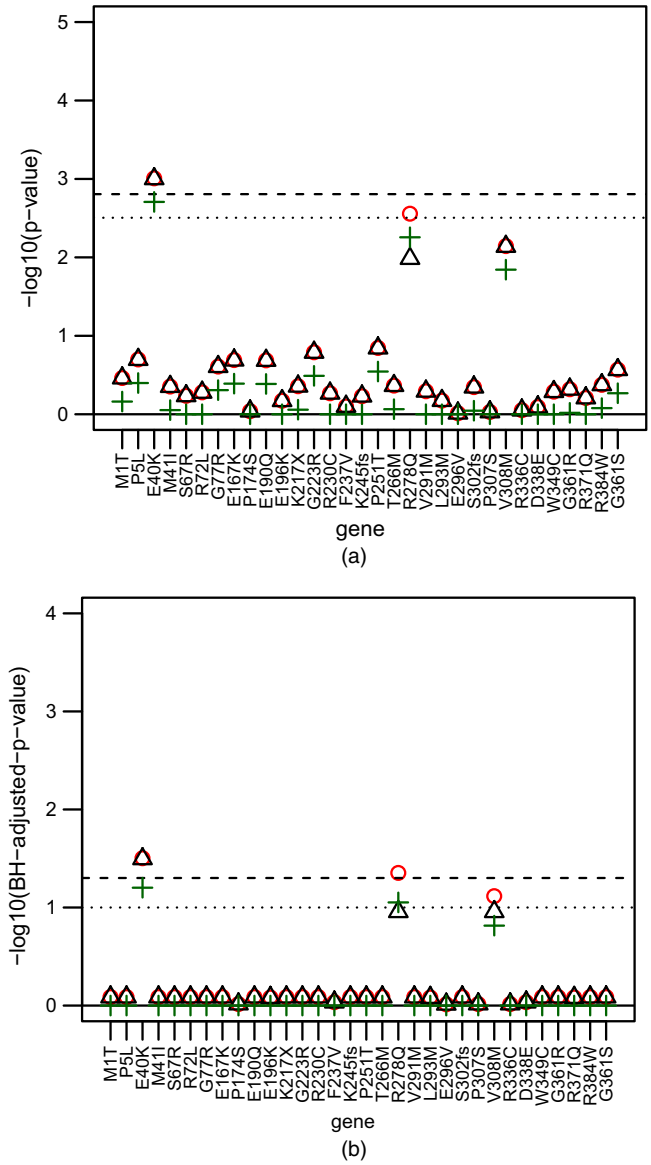

Fig. 9. (a) The naive and conditional p-values on a negative common log-scale and (b) FDR-adjusted p-values on a negative common log-scale for the 32 variants: the p-values plotted are the naive p-values (○), polyhedral conditional p-values (△) and the hybrid conditional p-values (+); · · · · · , multiplicity-adjusted threshold of 0.1 (horizontal line at $-\log(0.1/32)$) in (a) and $-\log(0.1)$ in (b); – – –, multiplicity-adjusted threshold of 0.05

These weights were obtained by using the default settings of the publicly available R library `SKAT` (Lee *et al.*, 2017). Fig. 9 and Table 1 provide respectively a graphical display and the actual numbers for the naive (i.e. unconditional, not corrected for selection) and conditional p-values. When using the polyhedral method that was described in Section 3.1, one variant, E40K, passes the Bonferroni threshold for FWER-control at the 0.05-level. When using the hybrid method that was described in Section 3.2, two variants, E40K and R278Q, are identified

at an FDR-level of 0.1. This example demonstrates that it is possible to make further discoveries in a follow-up analysis after aggregate testing, to identify which underlying variants drive the signal. The variant E40K is indeed associated with TG, as validated by external studies (Dewey *et al*., 2016).

7. Discussion

In this work, we provided valid inference for linear contrasts of estimated parameters, after the aggregate test passed a predefined threshold. For the post-selection inference that we suggest in this paper, we need only the summary statistics for the selected group of interest and knowledge of the selection threshold t_1. The selection threshold does not have to be fixed. For example, a data-dependent threshold will be valid if the groups are independent and selected by using the BH procedure, or any other simple selection rule (as defined in Benjamini and Bogomolov (2014)).

In many applications, e.g. whole genome studies, multiple groups are examined simultaneously. We considered two error measures that can be controlled across multiple groups of hypotheses: the overall FDR and the average FDR. For overall FDR-control, we demonstrated that our hierarchical approach of first selecting groups and then testing the individual hypotheses within the selected groups can be more powerful than testing all individual hypotheses when the signal is not too sparse or weak (i.e. settings where aggregate tests are appropriate). The selection rule for choosing the groups can have a great effect on the power; the optimal choice of t_1 to maximize the chance of discovery for individual hypotheses is an open problem, and data-adaptive methods for choosing t_1 may invalidate the post-selection inference. We are currently investigating potential approaches, but they are outside the scope of this paper.

For average FDR-control, the numerical experiments that we carried out suggest that our hierarchical approach has the same or better power than the hierarchical approach of Benjamini and Bogomolov (2014). Moreover, our approach has the added advantage of conditional FDR-control within each group. It will be interesting to compare the approaches in real whole genome data sets, which may contain genes with different sparsity levels.

Our methods can be extended to tree-structured hypothesis tests in a straightforward manner. See Bogomolov *et al*. (2017) and the references within for state of the art work on hierarchical testing when there are more than two layers. An interesting genomic application is as follows. Within a selected gene, the tests may be further divided naturally into subgroups, e.g. clusters of single-nucleotide polymorphisms within a gene (Yoo *et al*., 2016). It may be of interest to develop a multilevel analysis, where following selection we first examine the subgroups, and only then the individual effects.

In this work we suggested switching regimes to adapt to the different unknown sparsity of the estimated effects. We observed that, by combining a powerful method for the sparse setting with a powerful method for the non-sparse setting, we obtain a method that has overall good performance. Such an approach can be very useful in genomic applications, where the signal is expected to be sparse in some groups but non-sparse in others. The switching regime approach may benefit other post-selection settings as well, e.g. confidence intervals for the selected parameters in a regression model.

8. Supplementary material

An R implementation of the methods in this paper is available from `https://github.com/`

ammeir2/PSAT. The on-line supplement contains all proofs, the developed inference following selection by a linear aggregate test and additional simulation results.

Acknowledgements

RH and AM contributed equally to this paper. The authors thank Andriy Derkach for providing his R implementation of SKAT, and for helpful conversations on the rare variant testing applications, and Mathias Drton for helpful discussions regarding the properties of the conditional MLE. The authors are also grateful to Bin Zhu for helpful discussions of the Dallas Heart Study example. We thank the Joint Editor, the Associate Editor and two reviewers for their helpful comments which helped to improve the paper significantly. This research was supported by Israeli Science Foundation grant 1049/16 (RH).

References

Benjamini, Y. (2008) Microarrays, empirical Bayes and the two-group model. *Statist. Sci.*, **23**, 23–28.

Benjamini, Y. and Bogomolov, M. (2014) Selective inference on multiple families of hypotheses. *J. R. Statist. Soc.* B, **76**, 297–318.

Benjamini, Y. and Heller, R. (2007) False discovery rate for spatial signals. *J. Am. Statist. Ass.*, **102**, 1272–1281.

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

Benjamini, Y. and Yekutieli, D. (2005) False discovery rate-adjusted multiple confidence intervals for selected parameters. *J. Am. Statist. Ass.*, **100**, 71–81.

Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013) Valid post-selection inference. *Ann. Statist.*, **41**, 802–837.

Bhattacharjee, S., Rajaraman, P., Jacobs, K., Wheeler, W., William, A., Melin, B., Hartge, P., Yeager, M., Chung, C., Chanock, S. and Chatterjee, N. A. (2012) A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *Am. J. Hum. Genet.*, **90**, 821–835.

Bogomolov, M., Peterson, C., Benjamini, Y. and Sabatti, C. (2017) Testing hypotheses on a tree: new error rates and controlling strategies. *Preprint arXiv:1705.07529*.

Chatterjee, A. and Lahiri, S. N. (2011) Bootstrapping lasso estimators. *J. Am. Statist. Ass.*, **106**, 608–625.

Derkach, A., Lawless, J. F. and Sun, L. (2014) Pooled association tests for rare genetic variants: a review and some new results. *Statist. Sci.*, **29**, 302–321.

Dewey, F. E., Gusarova, V., O'Dushlaine, C., Gottesman, O., Trejos, J., Hunt, C., Van Hout, C. V., Habegger, L., Buckler, D., Lai, K.-M. V., Leader, J. B., Murray, M. F., Ritchie, M. D., Kirchner, H. L., Ledbetter, D. H., Penn, J., Lopez, A., Borecki, I. B., Overton, J. D., Reid, J., Carey, D. J., Murphy, A. J., Yancopoulos, G. D., Baras, A., Gromada, J. and Shuldiner, A. R. (2016) Inactivating variants in ANGPTL4 and risk of coronary artery disease. *New Engl. J. Med.*, **374**, 1123–1133.

Fithian, W., Sun, D. and Taylor, J. (2014) Optimal inference after model selection. *Preprint arXiv:1410.2597*. Department of Statistics, University of California at Berkeley, Berkeley.

Fuchsberger, C., Flannick, J., Teslovich, T. M., Mahajan, A., Agarwala, V., Gaulton, K. J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D. J., *et al.* (2016) The genetic architecture of type 2 diabetes. *Nature*, **536**, 41–47.

Garthwaite, P. H. and Buckland, S. T. (1992) Generating Monte Carlo confidence intervals by the Robbins-Monro process. *Appl. Statist.*, **41**, 159–171.

Goeman, J. J., de Geer, S. A. and van Houwelingen, H. C. (2006) Testing against a high dimensional alternative. *J. R. Statist. Soc.* B, **68**, 477–493.

Heller, R., Chatterjee, N., Krieger, A. and Shi, J. (2018) Post-selection inference following aggregate level hypothesis testing in large scale genomic data. *J. Am. Statist. Ass.*, **113**, 1770–1783.

Lee, S., Miropolsky, L. and Wu, M. (2017) SKAT: SNP-set (sequence) kernel association. *R Package Version 1.3.2.1*.

Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2016) Exact post-selection inference, with application to the lasso. *Ann. Statist.*, **44**, 907–927.

Lee, J. D. and Taylor, J. E. (2014) Exact post model selection inference for marginal screening. In *Proc. 27th Int. Conf. Neural Information Processing Systems, Montreal*, pp. 136–144. Cambridge: MIT Press.

Lee, S., Wu, M. C. and Lin, X. (2012) Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, **13**, 762–775.

Loftus, J. and Taylor, J. (2014) Selective inference in regression models with groups of variables. *Preprint arXiv:1511.01478v1*.

Madsen, B. E. and Browning, S. R. (2009) A groupwise association test for rare mutations using a weighted sum statistic. *PLOS Genet.*, **5**, no. 2, article e1000384.
McKeague, I. W. and Qian, M. (2015) An adaptive resampling test for detecting the presence of significant predictors. *J. Am. Statist. Ass.*, **110**, 1422–1433.
Meir, A. and Drton, M. (2017) Tractable post-selection maximum likelihood inference for the lasso. *Preprint arXiv:1705.09417*. Department of Statistics, University of Washington, Seattle.
Morris, A. P. and Zeggini, E. (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidem.*, **34**, 188–193.
Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S. M., Roeder, K. and Daly, M. J. (2011) Testing for an unusual distribution of rare variants. *PLOS Genet.*, **7**, no. 3, article e1001322.
Pakman, A. and Paninski, L. (2014) Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. *J. Computnl Graph. Statist.*, **23**, 518–542.
Penny, W. and Friston, K. (2003) Mixtures of general linear models for functional neuroimaging. *IEEE Trans. Med. Imgng*, **22**, 504–514.
Pötscher, B. M. (1991) Effects of model selection on inference. *Econometr. Theory*, **7**, 163–185.
Reid, S., Taylor, J. and Tibshirani, R. (2018) A general framework for estimation and inference from clusters of features. *J. Am. Statist. Ass.*, **113**, 280–293.
Reiner-Benaim, A. (2007) FDR control by the BH procedure for two-sided correlated tests with implications to gene expression data analysis. *Biometr. J.*, **49**, 107–126.
Romeo, S., Pennacchio, L. A., Fu, Y., Boerwinkle, E., Tybjaerg-Hansen, A., Hobbs, H. H. and Cohen, J. C. (2007) Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nat. Genet.*, **39**, no. 4, 513–516.
Romeo, S., Yin, W., Kozlitina, J., Pennacchio, L. A., Boerwinkle, E., Hobbs, H. H. and Cohen, J. C. (2009) Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. *J. Clin. Investign*, **119**, no. 1, 70–79.
Sun, J., Zheng, Y. and Hsu, L. (2013) A unified mixed-effects model for rare-variant association in sequencing studies. *Genet. Epidem.*, **37**, 334–344.
Taylor, J. and Tibshirani, R. (2018) Post-selection inference for l1-penalized likelihood models. *Can. J. Statist.*, **46**, 41–61.
Tian, X. and Taylor, J. E. (2015) Selective inference with a randomized response. *Preprint arXiv:1507.06739*.
UK10K Consortium and University College London–Edinburgh–Bristol Consortium (2015) The UK10K project identifies rare variants in health and disease. *Nature*, **526**, 82.
Weinstein, A., Fithian, W. and Benjamini, Y. (2013) Selection adjusted confidence intervals with more power to determine the sign. *J. Am. Statist. Ass.*, **108**, 165–176.
Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. and Lin, X. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.
Yekutieli, D. (2012) Adjusted Bayesian inference for selected parameters. *J. R. Statist. Soc.* B, **74**, 515–541.
Yoo, Y., Sun, L., Poirier, J., Paterson, A. and Bull, S. (2016) Multiple linear combination (MLC) regression tests for common variants adapted to linkage disequilibrium structure. *Genet. Epidem.*, **41**, 108–121.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material for: Post-selection estimation and testing following aggregate association tests'.