Statistics in Medicine

Saddlepoint approximations in logistic regression for analysing genome-wide association studies

Journal:	*Statistics in Medicine*
Manuscript ID	SIM-22-0239
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	25-Mar-2022
Complete List of Authors:	Johnsen, Pål; SINTEF, DIGITAL; Norwegian University of Science and Technology, Mathematical Sciences Bakke, Øyvind; Norwegian University of Science and Technology, Mathematical Sciences Bjørnland, Thea; Norwegian University of Science and Technology, Department of Mathematical Sciences DeWan, Andrew; Yale University School of Public Health, Chronic Disease Epidemiology and Center for Perinatal, Pediatric and Environmental Epidemiology Langaas, Mette; Norwegian University of Science and Technology, Mathematical Sciences
Keywords:	GWAS, imbalanced binary response, logistic regression, score test statistic, saddlepoint approximation

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Saddlepoint approximations in logistic regression for analysing genome-wide association studies

Pål V. Johnsen[1,2] | Øyvind Bakke[2] | Thea Bjørnland[2] | Andrew Thomas DeWan[3] | Mette Langaas[2]

[1]SINTEF DIGITAL, Oslo, Norway
[2]Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway
[3]Department of Chronic Disease Epidemiology and Center for Perinatal, Pediatric and Environmental Epidemiology, Yale School of Public Health, New Haven, Connecticut, USA

Correspondence
Pål V. Johnsen,
Email: pal.johnsen@sintef.no

Summary

We investigate saddlepoint approximations applied to the score test statistic in logistic regression for genome-wide association studies. The inaccuracy in the normal approximation of the score test statistic increases with increasing sample imbalance and with decreasing minor allele count. Applying saddlepoint approximations to the score test statistic distribution greatly improve the accuracy, even far out in the tail of the distribution. By using exact results for an intercept model and a binary covariate model, as well as simulations for models with nuisance parameters, we emphasize the need for continuity corrections in order to achieve valid p-values. The performance of the saddlepoint approximations is evaluated by overall and conditional type I error rate on simulated data. We investigate the methods further on data from UK Biobank with skin and soft tissue infections as phenotype, using both common and rare variants. The analysis confirms that continuity correction is important particularly for rare variants, and that the normal approximation gives a highly inflated type I error rate for logistic regression with case imbalance.

KEYWORDS:
GWAS, imbalanced binary response, logistic regression, score test statistic, saddlepoint approximation

1 | INTRODUCTION

We consider score tests for logistic regression models in which the response is imbalanced and the covariate of interest is discrete and skewed. This typically occurs in a genome-wide association study (GWAS) with binary phenotypes, henceforth denoted binary GWAS.

In a GWAS each single nucleotide polymorphism (SNP) is tested individually for association with a particular phenotype. In a modern biobank including several hundred thousands SNPs, rejection of the null hypothesis needs to be evaluated with a very low p-value threshold, typically equal to $5 \cdot 10^{-8}$, in order to control the family-wise error rate (FWER). In a binary GWAS with imbalanced response, new challenges arise.

As an example, we consider a follow-up study on skin and soft tissue infection (SSTI) using UK-biobank data, motivated by Rogne et al.[1] Using data on unrelated white European individuals with no prior history of SSTI at recruitment, we obtain 6.5 years of follow-up data on approximately 300 000 individuals, out of which approximately 0.7% were diagnosed with SSTI during follow-up, and classified as cases. The overall sample size may be large, but if there are few cases or controls with a certain genotype, relying on asymptotic normality of the score test statistic may yield spurious results. In fact, the score test applied under asymptotic theory yields invalid p-values if the case proportion is too small. In addition, the severity in this flaw increases with decreasing minor allele frequencies (MAF). Both Ma et al.[2] and Dey et al.[3] have illustrated this issue for sample

sizes of up to 20 000 individuals of which between 1% and 10% were cases. Motivated by the UK-Biobank SSTI data set, we show that the normal approximation can be flawed even when the total sample size is in the order of several hundred thousands. A solution proposed by Ma et al.[2] is to apply the Firth[4] bias-corrected logistic regression test. The test gives valid p-values when the imbalance is not too severe, and it is at the same time less conservative than the likelihood ratio test. As Firth's test is computationally inefficient for genome-wide testing, a test based on a saddlepoint approximation to the score statistic was proposed by Dey et al.[3] This so-called SPA-test showed good properties yielding both valid or close to valid p-values even when Firth's test failed to do so, as well as being as powerful as Firth's test.

Our theoretical contribution to the ongoing development of valid score tests for genome-wide association studies with imbalanced binary phenotypes is twofold. First, we establish the discrete and bounded nature of the score, and derive the exact conditional distribution of the score test statistic for two particular examples of logistic regression models, namely models with intercept and genetic variant only, as well a models with an additional binary nuisance covariate (covariates associated with regression nuisance parameters). Second, we propose continuity-corrected saddlepoint approximations to the conditional distribution of the score statistic. We compare our proposed method against exact results as well as the approach introduced in Dey et al.[3] We study the validity of tests both conditionally and unconditionally.

We show that a score test derived from the efficient score, or equivalently a null-orthogonal reparameterization of the logistic regression model, coincides with the SPA-test by Dey et al.[3], thus providing a novel interpretation of the SPA-test as a two-step approximation to the conditional distribution of the score statistic.

We study our proposed continuity-corrected saddlepoint approximations as well as other existing methods, using the follow-up study of SSTIs as explained above, and on simulated data.

2 | THE SCORE TEST STATISTIC FOR LOGISTIC REGRESSION MODELS IN GWAS

2.1 | Notation, statistical model and hypotheses

We consider tests for genotype–phenotype associations in large cohorts or populations. We assume that binary phenotypes, Y_i, non-genetic covariates $\boldsymbol{x}_i$ and allele counts g_i for a single variant, $i = 1, \ldots, n$, have been collected from n individuals. We consider directly biallelic allele counts in which $g_i = 0, 1$ or 2. We model the relationship between the response and the covariates in a logistic regression model in which the Y_i are independent and Bernoulli distributed with success probability μ_i and

$$\text{logit}\, \mu_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \gamma g_i, \tag{1}$$

$i = 1, \ldots, n$. Here, $\boldsymbol{x}_i$ is a vector of dimension d containing 1 (corresponding to an intercept) and $d - 1$ covariates, $\boldsymbol{\beta}$ a d-dimensional vector of nuisance parameters and γ the parameter of interest. Our aim is to perform the hypothesis test

$$H_0 : \gamma = 0 \quad \text{against} \quad H_1 : \gamma \neq 0. \tag{2}$$

In a GWAS, the test is performed multiple times, for different genetic variants. To control the FWER at a 5% level in GWAS involving common variants, a significance level of $5 \cdot 10^{-8}$ is commonly used for each test[5].

2.2 | The score test statistic

The score vector is the gradient of the log-likelihood function with respect to the parameters, which for the logistic regression model (1) is

$$\boldsymbol{U} = \begin{pmatrix} \boldsymbol{U}_\beta \\ U_\gamma \end{pmatrix} = \begin{pmatrix} X^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{\mu}) \\ \boldsymbol{g}^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{\mu}) \end{pmatrix}, \tag{3}$$

where $\boldsymbol{Y}$ and $\boldsymbol{g}$ are column vectors of length n with Y_i and g_i as elements respectively, $\boldsymbol{\mu} = E\boldsymbol{Y}$, and X is an $n \times d$ matrix with $\boldsymbol{x}_i^{\mathrm{T}}$ as rows. We have partitioned the score vector according to the parameter of interest, γ, and the nuisance parameters, $\boldsymbol{\beta}$. The score vector has mean $\boldsymbol{0}$ and covariance matrix, by definition referred to as the expected Fisher information

$$F = \begin{pmatrix} F_{\beta\beta} & \boldsymbol{F}_{\gamma\beta}^{\mathrm{T}} \\ \boldsymbol{F}_{\gamma\beta} & F_{\gamma\gamma} \end{pmatrix} = \begin{pmatrix} X^{\mathrm{T}}WX & X^{\mathrm{T}}W\boldsymbol{g} \\ \boldsymbol{g}^{\mathrm{T}}WX & \boldsymbol{g}^{\mathrm{T}}W\boldsymbol{g} \end{pmatrix}, \tag{4}$$

where W is a diagonal matrix with $\mu_i(1 - \mu_i)$ as the ii entry.

Using the score test, the null hypothesis of (2) is rejected if there is sufficient distance between the null value $\gamma = 0$ and the maximum likelihood estimate of γ. To judge this distance, without actually calculating the estimate, one uses the partial

derivative U_γ of the log-likelihood with respect to γ at $\gamma = 0$, along with the probability distribution of U_γ under the null. The proof of the following observation is given in Appendix A.

Observation 1. When $g_i \in \{0, 1, 2\}$, the score U_γ with respect to γ is a bounded lattice random variable with support on $-\boldsymbol{g}^{\mathrm{T}}\boldsymbol{\mu}$, $1 - \boldsymbol{g}^{\mathrm{T}}\boldsymbol{\mu}$, $2 - \boldsymbol{g}^{\mathrm{T}}\boldsymbol{\mu}$, …, $\boldsymbol{g}^{\mathrm{T}}\mathbf{1} - \boldsymbol{g}^{\mathrm{T}}\boldsymbol{\mu}$.

Importantly, the score is – as in our situation – often a function of unknown nuisance parameters. Then, one may consider the *conditional* null distribution of the score for the parameter of interest, U_γ, given that the components of the score vector corresponding to the nuisance parameters are equal to zero, $\boldsymbol{U}_{\boldsymbol{\beta}} = \mathbf{0}$ (see e.g. Smyth[6]). In this conditional framework, the unknown nuisance parameters are equal to the corresponding maximum likelihood estimates calculated under the null hypothesis $\gamma = 0$, so that $U_\gamma = \boldsymbol{g}^{\mathrm{T}}(\boldsymbol{Y} - \hat{\boldsymbol{\mu}})$, where $\hat{\boldsymbol{\mu}}$ consists of the fitted values of the null model. However, this conditional score test statistic will still be a lattice random variable, yet with a narrower support than described in Observation 1. See Appendix B.

In many applications, one may approximate the distribution of the score vector $\boldsymbol{U}$ by a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix F. The conditional distribution of U_γ given $\boldsymbol{U}_{\boldsymbol{\beta}} = \mathbf{0}$ under the null ($\gamma = 0$) is then asymptotically a normal distribution with mean 0 and variance

$$\tilde{F}_{\gamma\gamma} = \boldsymbol{g}^{\mathrm{T}}W\boldsymbol{g} - \boldsymbol{g}^{\mathrm{T}}WX(X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}W\boldsymbol{g}. \tag{5}$$

As outlined in the Introduction, the normal approximation to the score vector may lead to spurious results for genotype–phenotype associations when the phenotype is a binary variable. For example, even if the the overall sample size is large, the normal approximation may be inaccurate if the sample contains few individuals with response $y_i = 1$ (e.g., having the disease under study) and genotype $g_i > 0$ (carrying the minor allele).

In the next section, we present a score test for (2) based on a double saddlepoint approximation to the conditional null distribution of the score statistic U_γ for the logistic regression model (1), given $\boldsymbol{U}_{\boldsymbol{\beta}} = \mathbf{0}$. Here, we first state two observations that give the *exact* conditional null distribution for two special cases of the regression model (1). Proofs are given in Appendix A.

Observation 2. Consider a logistic regression model as in (1), but with logit $\mu_i = \beta + \gamma g_i$, henceforth denoted the *intercept model*. Let n_j be the number of individuals with genotype $g_i = j$, $j = 0, 1, 2$, and let logit $\mu = \beta$ under the null. Then, the null distribution of U_γ given $U_\beta = 0$ is a sum of trivariate hypergeometric point probabilities,

$$P(U_\gamma = u \mid U_\beta = 0) = \sum_{(v_0,v_1,v_2)\in S} \frac{\binom{n_0}{v_0}\binom{n_1}{v_1}\binom{n_2}{v_2}}{\binom{n}{n\mu}} = \sum_{k=\max(\lceil (u^*-n_1)/2\rceil,0)}^{\min(\lfloor u^*/2\rfloor,n_2)} \frac{\binom{n_0}{n\mu-u^*+k}\binom{n_1}{u^*-2k}\binom{n_2}{k}}{\binom{n}{n\mu}},$$

where the sum is taken over all triples (v_0, v_1, v_2) of integers in the set S defined by $0 \le v_j \le n_j$ for $j = 0, 1, 2$, $v_0+v_1+v_2 = n\mu$ and $v_1 + 2v_2 = u^*$, and $u^* = u + (n_1 + 2n_2)\mu$. The function outputs $\lceil x \rceil$ and $\lfloor x \rfloor$ denote the least integer greater than or equal to x (ceiling), and the largest integer less than or equal to x (floor) respectively.

Observation 3. Consider a logistic regression model as in (1), where logit $\mu_i = \beta_0 + \beta_1 x_i + \gamma g_i$, and x_i is a binary covariate taking value 0 or 1 (model with intercept and one binary non-genetic covariate). Let l_j be the number of individuals with $x_i = 0$ and genotype $g_i = j$, $j = 0, 1, 2$, and let $l = l_0 + l_1 + l_2$. Define similar counts m_j and m for individuals with $x_i = 1$. Let logit $\mu_0 = \beta_0$, and logit $\mu_1 = \beta_0 + \beta_1$. Then, under the null hypothesis,

$$P(U_\gamma = u \mid \boldsymbol{U}_{\boldsymbol{\beta}} = \mathbf{0}) = \sum_{s\in S} \frac{\binom{l_0}{v_0}\binom{l_1}{v_1}\binom{l_2}{v_2}}{\binom{l}{l\mu_0}} \frac{\binom{m_0}{w_0}\binom{m_1}{w_1}\binom{m_2}{w_2}}{\binom{m}{m\mu_1}},$$

where the sum is taken over all sextuples $\boldsymbol{s} = (v_0, v_1, v_2, w_0, w_1, w_2)$ of integers in the set S defined by $0 \le v_j \le l_j$, $0 \le w_j \le m_j$ for $j = 0, 1, 2$, $v_0 + v_1 + v_1 = l\mu_0$, $w_0 + w_1 + w_2 = m\mu_1$ and $v_1 + 2v_2 - (l_1 + 2l_2)\mu_0 + w_1 + 2w_2 - (m_1 + 2m_2)\mu_1 = u$.

From Observations 2 and 3, it follows that an *exact* p-value for the hypothesis test (2) can be computed for these two special cases of the logistic regression model (1). An extension of Observation 3 can also be derived for regression models with more categorical covariates. However, for more complex covariate patterns, this approach becomes computationally infeasible, and even intractable when continuous covariates are included. The next section introduces a method of computing p-values using double saddlepoint approximation.

3 | DOUBLE SADDLEPOINT APPROXIMATION

Tail probabilities $P(U_\gamma \geq u \mid \boldsymbol{U}_\beta = \boldsymbol{0})$ may be estimated by *double saddlepoint approximation*[7]. This will require the *cumulant generating function* of $\boldsymbol{U} = \begin{pmatrix} \boldsymbol{U}_\beta^{\mathrm{T}} & U_\gamma \end{pmatrix}^{\mathrm{T}} = \begin{pmatrix} X & \boldsymbol{g} \end{pmatrix}^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{\mu})$ (Section 2.2) and of $\boldsymbol{U}_\beta$.

3.1 | Cumulant generating function

The joint cumulant generating function of $\boldsymbol{U}$ is defined by $K(\boldsymbol{t}) = \ln E\left(e^{\boldsymbol{t}^{\mathrm{T}}\boldsymbol{U}}\right)$, were $\boldsymbol{t}$ is a vector of dimension $d+1$. By using the fact that Y_i is Bernoulli distributed with parameter μ_i (Section 2.1), we obtain

$$K(\boldsymbol{t}) = \sum_{i=1}^{n}\Big(\ln\big(1 - \mu_i + \mu_i e^{\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}_i}\big) - \mu_i \boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}_i\Big), \tag{6}$$

$$\nabla K(\boldsymbol{t}) = \sum_{i=1}^{n}\mu_i\left(\frac{1}{(1-\mu_i)e^{-\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}_i} + \mu_i} - 1\right)\boldsymbol{z}_i, \quad \text{and} \tag{7}$$

$$H(\boldsymbol{t}) = \sum_{i=1}^{n}\frac{\mu_i(1-\mu_i)e^{-\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}_i}}{\big((1-\mu_i)e^{-\boldsymbol{t}^{\mathrm{T}}\boldsymbol{z}_i} + \mu_i\big)^2}\boldsymbol{z}_i\boldsymbol{z}_i^{\mathrm{T}}, \tag{8}$$

where ∇K and H denote the gradient and the Hessian of K, respectively, and $\boldsymbol{z}_i = \begin{pmatrix} \boldsymbol{x}_i^{\mathrm{T}} & g_i \end{pmatrix}^{\mathrm{T}}$. The cumulant generating function of $\boldsymbol{U}_\beta$, its gradient and Hessian, K_β, ∇K_β and H_β, respectively, are obtained by replacing $\boldsymbol{z}_i$ by $\boldsymbol{x}_i$ and letting $\boldsymbol{t}$ have dimension d in (6)–(8).

3.2 | Approximated tail probabilities with continuity correction

The survival function (right-tail probability) $S(u) = P(U_\gamma \geq u \mid \boldsymbol{U}_\beta = \boldsymbol{0})$ can be approximated as given by Barndorff-Nielsen[8],

$$\hat{S}(u) = 1 - \Phi\Big(w - \frac{1}{w}\ln\frac{v}{w}\Big), \tag{9}$$

where Φ denotes the standard normal cumulative distribution function. To approximate the conditional survival function of a lattice random variable we have chosen the double saddlepoint survival approximation with the so-called second continuity correction. Using $f(\boldsymbol{t}_1, \boldsymbol{t}_2)$ as shorthand for $f\left(\begin{pmatrix} \boldsymbol{t}_1^{\mathrm{T}} & \boldsymbol{t}_2^{\mathrm{T}} \end{pmatrix}^{\mathrm{T}}\right)$, where f is a function and $\boldsymbol{t}_1, \boldsymbol{t}_2$ vectors, we have

$$\begin{aligned} w &= \operatorname{sgn}(\hat{t}_\gamma)\sqrt{2\Big(-K(\hat{\boldsymbol{t}}_\beta, \hat{t}_\gamma) + \hat{t}_\gamma\Big(u - \frac{1}{2}\Big)\Big)} \quad \text{and} \\ v &= 2\Big(\sinh\frac{\hat{t}_\gamma}{2}\Big)\sqrt{\frac{\det H(\hat{\boldsymbol{t}}_\beta, \hat{t}_\gamma)}{\det H_\beta(\boldsymbol{0})}}, \end{aligned} \tag{10}$$

where $\begin{pmatrix} \hat{\boldsymbol{t}}_\beta^{\mathrm{T}} & \hat{t}_\gamma \end{pmatrix}^{\mathrm{T}}$ is the *saddlepoint* satisfying $\nabla K(\hat{\boldsymbol{t}}_\beta, \hat{t}_\gamma) = \begin{pmatrix} \boldsymbol{0}^{\mathrm{T}} & u - 1/2 \end{pmatrix}^{\mathrm{T}}$(Skovgaard[9], see Butler[7], p.114). In general, also the d-dimensional vector $\tilde{\boldsymbol{t}}_\beta$ satisfying $\nabla K_\beta(\tilde{\boldsymbol{t}}_\beta) = \boldsymbol{0}$ is involved in the expressions for w and v, but $\tilde{\boldsymbol{t}}_\beta = \boldsymbol{0}$ in our case (see Appendix C). Left-tail probabilities can be approximated, taking into account that U_γ is a lattice variable with step 1, by $P(U_\gamma \leq u \mid \boldsymbol{U}_\beta = \boldsymbol{0}) = 1 - S(u+1)$.

3.3 | Two-sided p-values

By assuming the score test statistic to have a normal distribution, and for some observation u, a two-sided p-value is reasonable and given by $P(|U_\gamma| \geq |u| \mid \boldsymbol{U}_\beta = \boldsymbol{0})$ (under the null). However, as the score test statistic has a lattice distribution, the point $-u$ might not be on the grid. If so, the closest grid point to $-u$ farthest away from zero is obtained by $u_{inv} = u - \operatorname{sgn}(u)\cdot\lceil 2\cdot|u|\rceil$. We define a two-sided p-value, assuming u positive, to be $P(U_\gamma \geq u \mid \boldsymbol{U}_\beta = \boldsymbol{0}) + P(U_\gamma \leq u_{inv} \mid \boldsymbol{U}_\beta = \boldsymbol{0})$, and vice versa when u is negative.

An example is given in Figure 1a where the exact lattice distribution of the score test statistic under the null hypothesis is given for the intercept model with a genotype vector simulated with MAF = 0.05 and a case proportion of 0.05 ($n = 1000$). Included is the support of the lattice distribution $[u^{min}, u^{max}] = [-5.5, 46.5]$. An observed $u = 4.5$ will then give $u_{inv} = -4.5$, a situation where $u_{inv} = -u$. The intercept model conditional p-value (of Observation 2) is then equal to the sum of the bars

coloured in orange. The deviation from the normal distribution increases for decreasing case proportion, as can be seen when comparing Figure 1a to 1b, where the case proportion is reduced to 0.01 while keeping the same genotype vector in Figure 1b. In fact, the skewness increases for decreasing case proportion such that the probability mass of the distribution is concentrated on the left, with a longer right tail. Consequently, the score test statistic is asymmetric as well as bounded, which means the point u_{inv} might be outside the support of the lattice distribution. In that case, a one-sided p-value will be computed as seen in Figure 1b with bars coloured orange only to the right of the observed $u = 1.9$ ($u_{inv} = -2.1 < u^{min} = -1.1$). The same observation of increased skewness can be seen with a fixed case proportion, but decreasing MAF.

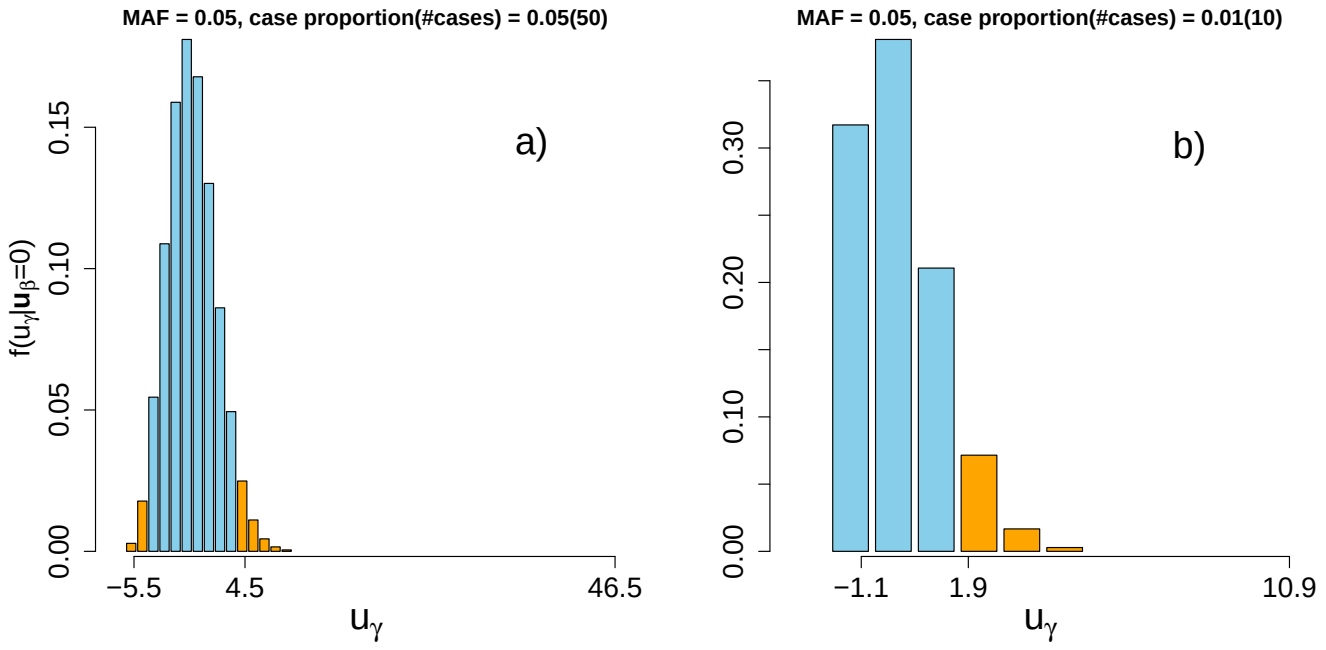

FIGURE 1 The exact lattice distribution of the score test statistic for the intercept model for different case proportions (genotype vector fixed, 1000 individuals). Included is the support $[u^{min}, u^{max}]$ of the lattice distribution in each case together with an example of an observed statistic in between, as well as the corresponding computed p-value coloured in orange. The deviation from the normal distribution increases for decreasing case proportion. When the distribution is sufficiently skewed, a one-sided p-value is computed.

4 | SINGLE SADDLEPOINT APPROXIMATION USING THE EFFICIENT SCORE

Our proposed method is related to the SPA-test by Dey et al.[3], which is also based on a saddlepoint approximation to the distribution of a score test statistic. In this section, we provide a novel interpretation of the SPA-test as a two-step approximation to conditional inference, and propose a modification.

We implicitly introduced the score test statistic $\boldsymbol{g}^T(\boldsymbol{Y} - \hat{\boldsymbol{\mu}})$, where $\hat{\boldsymbol{\mu}}$ is the maximum likelihood estimate of $\boldsymbol{\mu}$ under the null hypothesis, solved by $\boldsymbol{U}_\beta = \boldsymbol{0}$. Rather than approximating the distribution of this test statistic directly, the common procedure for score test statistics in the presence of nuisance parameters is to use conditional inference by conditioning U_γ on $\boldsymbol{U}_\beta = \boldsymbol{0}$.

Other methods for approximate conditional inference in the presence of nuisance parameters include orthogonal parameterization[10] and projective methods.[11] The first-order projective score, perhaps better known as the *efficient score*, is for our model (Equation (1)) defined by

$$\tilde{U}_\gamma = U_\gamma - \boldsymbol{F}_{\gamma\beta} F_{\beta\beta}^{-1} \boldsymbol{U}_\beta.$$

As noted by Bickel et al.[12], the efficient score may be interpreted in general as the score corresponding to a reparameterization $(\boldsymbol{\beta}, \gamma) \rightarrow (\boldsymbol{\alpha}, \gamma)$, by letting $\boldsymbol{\beta}(\boldsymbol{\alpha}, \gamma) = \boldsymbol{\alpha} - F_{\beta\beta}^{-1} \boldsymbol{F}_{\gamma\beta}^{T} \gamma$. With this reparameterization of the logistic regression model, $\text{logit}(\mu_i) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}(\boldsymbol{\alpha}, \gamma) + \gamma g_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\alpha} + \gamma \tilde{g}_i$, where $\tilde{g}_i = g_i - \boldsymbol{x}_i^{\mathrm{T}} F_{\beta\beta}^{-1} \boldsymbol{F}_{\gamma\beta}^{\mathrm{T}}$. Let $\tilde{F}$ denote the expected Fisher information of $\tilde{\boldsymbol{U}} = \left(\tilde{\boldsymbol{U}}_{\boldsymbol{\alpha}}^{\mathrm{T}} \; \tilde{U}_{\gamma}\right)^{\mathrm{T}}$, the reparameterized score vector. With this reparameterization, the parameter γ and the nuisance parameters $\boldsymbol{\alpha}$ are locally information orthogonal at $\gamma = 0$, which means that $\tilde{\boldsymbol{F}}_{\boldsymbol{\alpha}\gamma}$ and $\tilde{\boldsymbol{F}}_{\gamma\boldsymbol{\alpha}}$ in the expected Fisher information $\tilde{F}$ are zero-vectors (see e.g. Lindsey[13]). In this case, asymptotically $\tilde{\boldsymbol{U}}$ has a normal distribution, however additionally $\text{Cov}(\tilde{\boldsymbol{U}}_{\boldsymbol{\alpha}}(\hat{\boldsymbol{\mu}}), \tilde{U}_{\gamma}(\hat{\boldsymbol{\mu}})) \rightarrow \boldsymbol{0}$ when $\gamma = 0$ and $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}$. With $\tilde{\boldsymbol{U}}$ asymptotically multivariate normal, so will $\tilde{\boldsymbol{U}}_{\boldsymbol{\alpha}}$ and $\tilde{U}_{\gamma}$ (univariate) be. As covariance equal to zero for two normal distributed random variables implies independence, this means that the distribution of $\tilde{U}_{\gamma}$ conditional on $\tilde{\boldsymbol{U}}_{\boldsymbol{\alpha}} = \boldsymbol{0}$ is asymptotically the same as the unconditional distribution of $\tilde{U}_{\gamma}$ when the null hypothesis is true with $\hat{\boldsymbol{\mu}}$ treated as a plug-in constant for $\boldsymbol{\mu}$.

In our case with expected Fisher information given in (4),

$$\begin{aligned}
\tilde{U}_{\gamma} &= \boldsymbol{g}^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{\mu}) - \boldsymbol{g}^{\mathrm{T}} W X (X^{\mathrm{T}} W X)^{-1} X^{\mathrm{T}} (\boldsymbol{Y} - \boldsymbol{\mu}) \\
&= (\boldsymbol{g}^{\mathrm{T}} - \boldsymbol{g}^{\mathrm{T}} W X (X^{\mathrm{T}} W X)^{-1} X^{\mathrm{T}})(\boldsymbol{Y} - \boldsymbol{\mu}) \\
&= (\boldsymbol{g} - X (X^{\mathrm{T}} W X)^{-1} X^{\mathrm{T}} W \boldsymbol{g})^{\mathrm{T}} (\boldsymbol{Y} - \boldsymbol{\mu}) \\
&= \tilde{\boldsymbol{g}}^{\mathrm{T}} (\boldsymbol{Y} - \boldsymbol{\mu}),
\end{aligned}$$

with $\tilde{\boldsymbol{g}} = \boldsymbol{g} - X (X^{T} W X)^{-1} X^{T} W \boldsymbol{g}$ the vector of all components $\tilde{g}_i$, and defined as in Dey et al.[3] Observe that when $\boldsymbol{U}_{\beta} = X^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{\mu}) = \boldsymbol{0}$, the observed efficient score, $\tilde{u}$, is equal to u, the original observed score. Moreover, $E(\tilde{U}_{\gamma} | \tilde{\boldsymbol{U}}_{\boldsymbol{\alpha}} = \boldsymbol{0}) = E(U_{\gamma} | \boldsymbol{U}_{\beta} = \boldsymbol{0}) = 0$, and $\text{Var}(\tilde{U}_{\gamma} | \tilde{\boldsymbol{U}}_{\boldsymbol{\alpha}} = \boldsymbol{0}) = \text{Var}(U_{\gamma} | \boldsymbol{U}_{\beta} = \boldsymbol{0}) = \tilde{\boldsymbol{g}}^{\mathrm{T}} W \tilde{\boldsymbol{g}}$ with $\tilde{\boldsymbol{U}}_{\boldsymbol{\alpha}} = \boldsymbol{U}_{\beta}$ under the null hypothesis. At last, observe that asymptotically as $\hat{\boldsymbol{\mu}} \xrightarrow{p} \boldsymbol{\mu}$ under the null hypothesis,

$$\begin{aligned}
&\text{Cov}(\tilde{U}_{\gamma}(\hat{\boldsymbol{\mu}}), \tilde{\boldsymbol{U}}_{\boldsymbol{\alpha}}(\hat{\boldsymbol{\mu}})) = F_{\gamma\boldsymbol{\alpha}}(\hat{\boldsymbol{\mu}}) = E\left(\tilde{U}_{\gamma}(\hat{\boldsymbol{\mu}}) \tilde{\boldsymbol{U}}_{\boldsymbol{\alpha}}(\hat{\boldsymbol{\mu}})^{\mathrm{T}}\right) \\
&= E\left(\left(\boldsymbol{g} - X \left(X^{\mathrm{T}} \hat{W} X\right)^{-1} X^{\mathrm{T}} \hat{W} \boldsymbol{g}\right)^{\mathrm{T}} (\boldsymbol{Y} - \hat{\boldsymbol{\mu}}) (\boldsymbol{Y} - \hat{\boldsymbol{\mu}})^{\mathrm{T}} X\right) \\
&= \boldsymbol{g}^{\mathrm{T}} E\left((\boldsymbol{Y} - \hat{\boldsymbol{\mu}}) (\boldsymbol{Y} - \hat{\boldsymbol{\mu}})^{\mathrm{T}}\right) X - E\left(\boldsymbol{g}^{\mathrm{T}} \hat{W} X \left(X^{\mathrm{T}} \hat{W} X\right)^{-1} X^{\mathrm{T}} (\boldsymbol{Y} - \hat{\boldsymbol{\mu}}) (\boldsymbol{Y} - \hat{\boldsymbol{\mu}})^{\mathrm{T}}\right) X \\
&\rightarrow \boldsymbol{g}^{\mathrm{T}} W X - \boldsymbol{g}^{\mathrm{T}} W X \left(X^{\mathrm{T}} W X\right)^{-1} X^{\mathrm{T}} W X = \boldsymbol{0}^{T},
\end{aligned}$$

where $\hat{W}$ is the diagonal matrix with $\hat{\mu}_i(1 - \hat{\mu}_i)$ as the ii entry. Hence, we have shown indeed that $\tilde{U}_{\gamma}$ and $\tilde{\boldsymbol{U}}_{\boldsymbol{\alpha}}$ are asymptotically independent under the null hypothesis.

Under the null hypothesis, using $\tilde{\boldsymbol{U}}$ leads asymptotically to the same *unconditional* inference of $\tilde{U}_{\gamma}(\hat{\boldsymbol{\mu}})$ as the *conditional* inference of U_{γ} given $\boldsymbol{U}_{\beta} = \boldsymbol{0}$. In other words, $f(\tilde{U}_{\gamma}) \xrightarrow{d} N(0, \tilde{F}_{\gamma\gamma})$, with $\tilde{F}_{\gamma\gamma}$ given in (5). However, this will still be inaccurate for an imbalanced response and a skewed covariate of interest. Under this framework, we interpret the test proposed by Dey et al.[3] as a two-step approach, where the first step is to apply the efficient score, and in the second step the corresponding unconditional statistic is approximated by a single saddlepoint method via the univariate cumulant generating function of $\tilde{U}_{\gamma}$, given by

$$K(t) = \sum_{i=1}^{n} \ln(1 - \hat{\mu}_i + \hat{\mu}_i e^{\tilde{g}_i t}) - t \, \tilde{\boldsymbol{g}}^{T} \hat{\boldsymbol{\mu}}.$$

Since such a two-step approach does not require a double saddlepoint approximation, this method is computationally more efficient. In Dey et al.[3], the efficient score test statistic is assumed to have a continuous distribution. However, when $g_i \in \{0, 1, 2\}$, the efficient score test statistic in fact has a lattice distribution. Therefore, we propose to use a continuity correction. Similarly to the continuity-corrected double saddlepoint method outlined in the previous section, left-tail probabilities are estimated as in Equation (9), now with

$$w = \text{sgn}(\hat{t}) \sqrt{2(\hat{t}(u_{\gamma} - 1/2) - K(\hat{t}))}, \text{ and } v = 2\sinh(\hat{t}/2) \sqrt{K''(\hat{t})},$$

where $\hat{t}$ is the saddlepoint obtained by solving $K'(\hat{t}) = u_{\gamma} - 1/2$. Furthermore, we apply the same algorithm for obtaining two-sided p-values as in Section 3.3.

5 | COMPARISON OF METHODS

For a specified significance level α, a *valid* test satisfies $P(\text{type I error}) \leq \alpha$. In our setting, we find it relevant to distinguish between conditional and overall (unconditional) validity. To clarify what is meant by this, consider a simple logistic regression model with no nuisance covariates (intercept only model). The covariate vector $\boldsymbol{g}$ is fixed while the response vector $\boldsymbol{Y}$ is random. Under the null, $Y_i \sim \text{binom}(1, \mu)$ for all $i = 1, \ldots, n$, where $\mu = \exp(\beta_0)/(1 + \exp(\beta_0))$. For a particular realization $\boldsymbol{y}$, the observed score test statistic $u_\gamma = \boldsymbol{g}^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) = \boldsymbol{g}^T(\boldsymbol{y} - \bar{y}\boldsymbol{1})$ may be compared to the conditional null distribution of U_γ, i.e. the distribution of $\boldsymbol{g}^T(\boldsymbol{Y} - \bar{y}\boldsymbol{1})$ given that $\boldsymbol{Y}$ is restricted by $\sum_i Y_i = n\bar{y} = v$ (Observation 2). Thus, for all datasets in which the realization $\boldsymbol{y}$ satisfies $\sum_i y_i = v$, a test is *conditionally* valid only when $P(\text{type I error}| \sum_i Y_i = v) \leq \alpha$. On the other hand, the *overall* probability of type I error is given by

$$\sum_v \left[P\left(\text{type I error} \mid \sum_{i=1}^n Y_i = v \right) P\left(\sum_{i=1}^n Y_i = v \right) \right]. \tag{11}$$

A test that is conditionally valid for all v, will also be valid overall. The exact test derived in Observation 2 satisfies this property. An approximation to the exact test may be conditionally valid for some v, but invalid overall, or valid overall but conditionally invalid for some v. In the case with nuisance covariates, Equation (11) may be generalized to:

$$\sum_{X, \boldsymbol{y} \,:\, \boldsymbol{U}_\beta = \boldsymbol{0}} \left[P\left(\text{type I error} \mid \boldsymbol{U}_\beta = \boldsymbol{0} \right) P\left(\boldsymbol{U}_\beta = \boldsymbol{0} \right) \right].$$

To evaluate the performance of our proposed methods, we consider both conditional and overall validity for models where the exact test is available. Approximation methods are evaluated based on their ability to control the overall type I error rate as well as the proportion of tests that are conditionally invalid.

5.1 | Intercept model

In this section, we consider the intercept model with no nuisance parameters. We compare two discrete and two continuous conditional inference approximation methods with the exact test. The discrete methods are the double saddlepoint method with continuity correction as described in Section 3, henceforth termed DSPA-CC, and the continuity-corrected single saddlepoint method based on the efficient score as described in Section 4, henceforth termed ESPA-CC. The continuous methods are the normal approximation and the single saddlepoint method based on the efficient score (henceforth termed ESPA). To the best of our knowledge, the ESPA method mimics the SPA-test of Dey et al.[3] as implemented in the SPAtest-package in R[14]. We present a simple example in order to highlight some of the key differences between the methods.

Let $n = 1000$ and let $\boldsymbol{g}$ be the covariate vector with $n_0 = 980$ and $n_1 = 20$ and $n_2 = 0$. Without specifying what μ is, we first calculate the probabilities $P\left(\text{type I error} \mid \sum_{i=1}^n Y_i = v\right)$, for all $v = 1, 2, \ldots, n-1$. For a particular realization v, and discrete sample space within the support $[u_L, u_U]$ of the conditional null distribution of U_γ, where u_L and u_U need not be integers, we obtain the rejection region $\{u_L, \ldots, c_L\} \cup \{c_U, \ldots, u_U\}$ of the exact test. This can be achieved by a grid search from the left to obtain c_L as well as a separate grid search from the right to obtain c_U since the probability distribution is not symmetric. Then, $P\left(\text{type I error} \mid \sum_{i=1}^n Y_i = v\right) = P(U_\gamma \leq c_L \cup U_\gamma \geq c_U \mid \sum_{i=1}^n Y_i = v)$. For the approximation methods DSPA-CC, SPA-CC and SPA, we similarly use a grid search to identify lower (c_L^*) and upper (c_U^*) critical values that lead to rejection at the specified significance level. For the normal approximation, we obtain a critical value c^* from the normal distribution with mean 0 and variance $\frac{v}{n}(1 - \frac{v}{n})\left[n_0(0 - \frac{n_1}{n})^2 + n_1(1 - \frac{n_1}{n})^2\right]$, and then obtain the proper lower and upper critical values by the nearest grid points c_L^* and c_U^* to $-c^*$ and c^* such that $c_L^* \leq -c^*$ and $c_U^* \geq c^*$. Then, for rejection regions $\{u_L, \ldots, c_L^*\} \cup \{c_U^*, \ldots, u_U\}$, we calculate the exact conditional probability of erroneously rejecting the null hypothesis using the different approximation methods. For a specified value of μ, we obtain probabilities $P\left(\sum_{i=1}^n Y_i = v\right)$ for each observed v. The overall probability of type I error can be computed according to Equation (11). In addition, the probability of a conditionally invalid test for each method and for each μ can be computed by observing which values v where $P(\text{type I error} \mid \sum_i Y_i = v) > \alpha$, and add together the probabilities $P\left(\sum_{i=1}^n Y_i = v\right)$ for each such v. See Figure 2.

From this example, we make four observations;

1. The exact test is always conservative (see Figure 2). When a significance level α is specified, the discrete nature of the test results in an achieved significance level less than α. This observation is of course well-known for discrete test statistics.

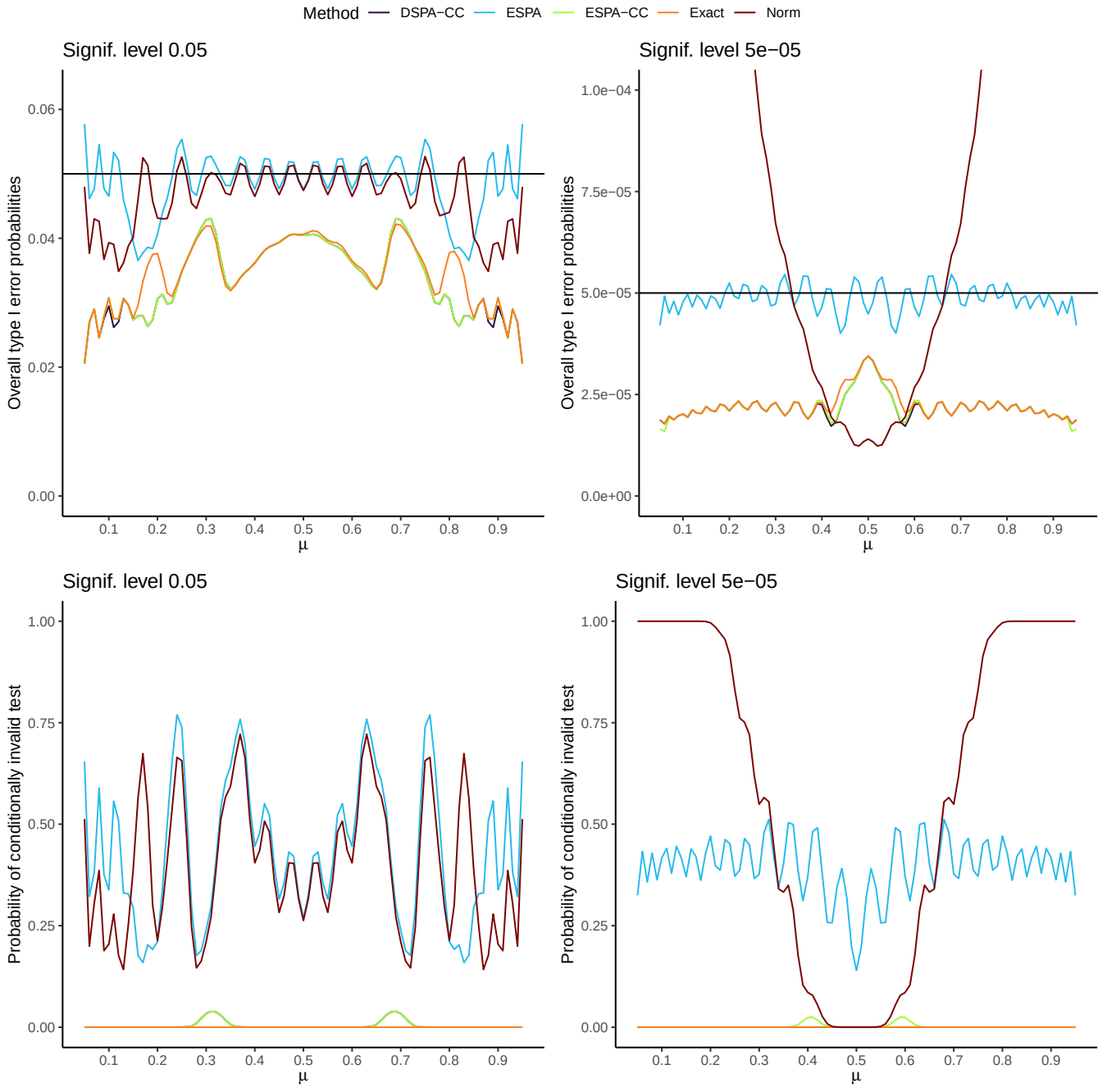

FIGURE 2 Exact overall type I error probabilities as well as probability of conditionally invalid tests for the different approximations methods for the distribution of the score test statistic, for different values of μ in the intercept model. We compare with the exact test using the known distribution of the score test statistic.

2. Both of the discrete approximations (DSPA-CC and SPA-CC) closely resemble the exact test in terms of overall type I error rates (Figure 2). At significance level $\alpha = 0.05$, both methods gave conditionally invalid tests in four situations; $v = 301$, $v = 325$, $v = 675$, and $v = 699$. For instance for $\mu = 0.31$ and $\mu = 0.69$, this results in probabilities ≈ 0.04 of sampling a dataset where these methods are conditionally invalid. At significance level $\alpha = 5 \cdot 10^{-5}$, DSPA-CC is conditionally valid for any v, while SPA-CC is conditionally invalid for $v = 406$ and $v = 594$. For instance for $\mu = 0.41$ and $\mu = 0.59$, this results in a slight probability (≈ 0.02) of sampling a dataset where the SPA-CC method is conditionally invalid.

3. Even at significance level $\alpha = 0.05$, the normal approximation is invalid for different μ-values (Figure 2). For significance level $5 \cdot 10^{-5}$, the normal approximation is valid when the response is balanced ($\mu \approx 0.5$). However, for skewed responses (small or large μ), the normal approximation becomes severely unreliable. At significance level $\alpha = 0.05$, the normal approximation was conditionally invalid in around 40% of possible realizations of $\sum_i Y_i$. At significance level $\alpha = 5 \cdot 10^{-5}$, this number had increased to around 64%. The majority of situations where the normal approximation was conditionally invalid was for small or large number of cases v, which is in-line with the observations made of overall type I error rates for skewed responses (Figure 2).

4. The ESPA method is less conservative than the exact test, and at times anti-conservative. At significance level $\alpha = 0.05$, the ESPA method was conditionally invalid around 43% of possible realizations of $\sum_i Y_i$, and at significance level $\alpha = 5 \cdot 10^{-5}$, the ESPA method was conditionally invalid around 39% of situations. As opposed to the normal approximation method where invalid tests clustered towards skewed response distributions, the ESPA method fluctuates relatively evenly between conditionally valid and conditionally invalid as the number of cases v increases for both significance levels 0.05 and $5 \cdot 10^{-5}$. Therefore, the test is approximately equally good at any μ (Figure 2). Furthermore, the absolute differences in type I error rate control improves as the significance level decreases. This observation has a simple explanation. For some data sets, the ESPA method yields the same critical region as the exact test, while at times the critical region is shifted by as little as one unit ($c_U^* = c_U - 1$ or $c_L^* = c_L + 1$). At a significance level of $\alpha = 0.05$, this shift can result in a substantial inflation in type I error rates, while at small significance levels, point probabilities are of such small magnitudes that the shift is less notable. As critical regions oscillate between correct and slightly shifted, conditional type I error rates oscillate above and below α, and averaging out to produce an overall type I error rate $\approx \alpha$.

5.2 | Simulations of genetic association studies with an imbalanced response

The purpose of the following simulation study is to compare methods in a setting resembling a genome-wide association study with an imbalanced response, for which exact tests are not available. The simulation set-up is motivated by Dey et al.[3] by conditioning on the number of cases, and we estimate the type I error rate *conditional* on the number of cases. The sample size considered is $n = 20000$, with case proportion 2% and 0.2%. We consider the logistic regression model

$$\text{logit}(\mu_i) = \beta_0 + x_{i,1} + x_{i,2} + \gamma g_i,$$

with $X_1 \sim$ Bernoulli(0.5), $X_2 \sim N(0, 1)$ and $G \sim$ binom(2, MAF) with the MAF taking the values 0.05, 0.005, 0.0005 and 0.00025. Since we are evaluating validity of tests, we set $\gamma = 0$. Finally, we set $\beta_0 = -5.6$ such that the disease prevalence is 1% in the population.

The covariates $x_{i,1}$ and $x_{i,2}$ are sampled conditionally on their respective phenotype value y_i, while the genotype value is sampled independently of this under the null hypothesis. See Supplementary File for details. This ensures that the number of cases is equal for all simulations. For each set of case proportion and MAF, we simulate 10^9 data sets and record the number of times the null hypothesis is rejected at the $\alpha = 5 \cdot 10^{-8}$ significance level when using (1) the double saddlepoint approximation with continuity correction (DSPA-CC), (2) the continuity-corrected univariate saddlepoint approximation based on the efficient score (ESPA-CC), and (3) the continuous univariate saddlepoint approximation of the efficient score (ESPA). The resulting empirical type I error rates are presented in Figure 3, along with 95% Clopper-Pearson confidence intervals.

The simulation results closely follow the observations made in the previous section. The DSPA-CC and ESPA-CC are conservative (overall probability of type I error $< \alpha$), while the type I error rate of the ESPA method is $\approx \alpha$. The results are comparable with the pattern for conditionally invalid tests in Figure 2, specifically for the small case proportion, in that we sense a large fluctuation in the probability of invalid tests for ESPA, while both ESPA-CC and DSPA-CC have a small probability of invalid test, which is decreasing for decreasing MAF. We also observe that the type I error rate, conditional on the number of cases, for EPSA is increasing for decreasing MAF. The simulation study with case proportion 0.002 serves to illustrate deviations between the DSPA-CC and ESPA-CC method, and we observe that the ESPA-CC is somewhat more conservative in this setting.

6 | APPLICATION TO UK BIOBANK DATA

We consider a recent GWAS in the UK Biobank with motivation from Rogne et al.[1] The phenotype of interest is skin and soft tissue infections (SSTIs), and individuals are defined as cases if they have been hospitalized with main ICD-10 codes

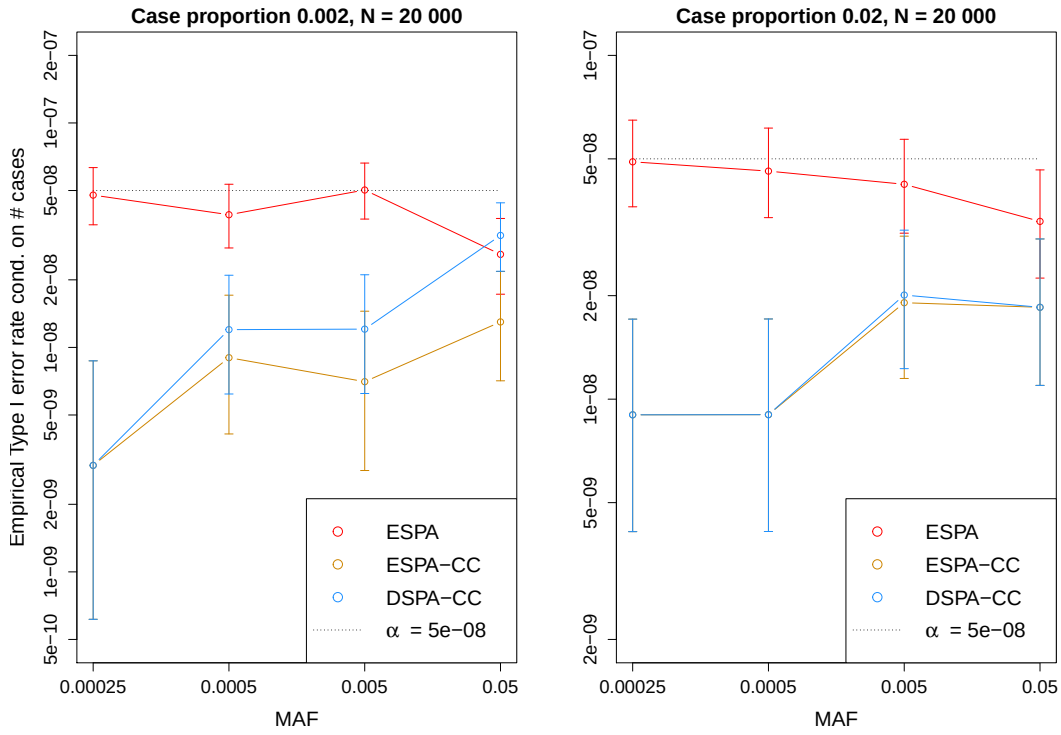

FIGURE 3 Approximated expected type I error rates - conditional on the number of cases - for ESPA, ESPA-CC and DSPA-CC from simulations with case proportions 0.02 and 0.002, and for small MAFs when nuisance covariates are included.

A46 (erysipelas), L03 (cellulitis and acute lymphangitis), or M72.6 (necrotizing fasciitis) in the period between the end of the recruitment period (2010-10-01) and April 2017 (2017-03-31). Individuals who had reported ICD-10 codes, or corresponding ICD-9 codes (035 and 729.4), before 2010-10-01 are removed as well as individuals with date of death reported after 2010-10-01 in the death register (see Data-Field 40000 in the UK Biobank data). As nuisance covariates we include age when attended assessment centre, genetic sex, and four principal components. To avoid complexities due to cryptic relatedness we only include unrelated individuals reported as White British (achieved through Data-Field 22006 and 22020 in UK Biobank). The principal components are calculated using EIGENSOFT (version 6.1.4) SmartPCA.[15,16] Only directly genotyped SNPs are considered, and phenotype-independent quality control of the genetic data is completed using PLINK1.9, with details given in the Supplementary File. This results in a total of 293 964 individuals and 529 024 SNPs with 2051 individuals defined as cases and 291 913 controls, resulting in a case proportion of 0.7 %. All SNPs are first investigated by computing p-values using the normal approximation to the score test statistic. As this test is proven to be too optimistic, SNPs with p-values less than $\alpha = 5 \cdot 10^{-5}$ are investigated more thoroughly by computing p-values using the DSPA-CC and ESPA-CC methods as implemented by us, as well as the SPA-test of Dey et al.[3], denoted ESPA. In Dey et al.[3], a computationally more efficient approximation to their SPA-test is also proposed by essentially assuming that the nuisance covariates are balanced. In a double saddlepoint setting, this assumption may be generalized to argue that the score vector $\boldsymbol{U}_\beta$ approximately has a multivariate normal distribution under the null hypothesis. By taking a similar approach, we may partition the joint CDF of $\boldsymbol{U}_\beta$ and U_γ into a sum over all individuals with genotype value $g_i > 0$ and those with $g_i = 0$. For the latter sub-sample, the CGF simplifies to a CGF of the score vector $\boldsymbol{U}_\beta^*$ including individuals with $g_i = 0$. Assuming that also $\boldsymbol{U}_\beta^*$ is normal, this part of the joint CGF may be replaced by a normal CGF, and by pre-computing the variance of $\boldsymbol{U}_\beta^*$, an approximated double saddlepoint method may be computed based only on the sub-sample individuals with genotypes $g_i > 0$. Details may be found in the Supplementary File. For comparative purposes, we also compute p-values based on the fastSPA method of Dey et al.[3] and our similar fastDSPA-CC approach.

Test results for the SNPs with the smallest normal-approximated p-values are given in Table 1. In this setting, we no longer know whether the null hypothesis is true or not for each variant. However, we expect only a tiny proportion of all variants where the null hypothesis is false. Even though no SNPs reached the significance level $\alpha = 5 \cdot 10^{-8}$, we see a pattern similar to the results for the intercept model and our simulation results. The normal approximation is the most optimistic, followed by ESPA

and fastSPA tests. The DSPA-CC test is more conservative, while the most conservative test is ESPA-CC. The fastDSPA-CC is slightly less conservative than DSPA-CC. The greatest difference between test results is observed for the SNP with a small minor allele frequency (rs113113104, MAF = 0.03). The difference between the *p*-values reduces for increasing MAFs. For the SNP rs566530 with MAF = 0.48, the SPA test gives a smaller *p*-value than the normal approximation, while the other methods give consistently larger *p*-values.

TABLE 1 The common variants with the smallest computed *p*-values using normal approximation to the score test statistic for the GWAS of skin and soft tissue infections. Alternative *p*-value computations are included for comparison.

SNP	CHR	MAF	Norm	ESPA	fastSPA	SPA-CC	DSPA-CC	fastDSPA-CC
rs113113104	6	0.03	2.39e-07	5.97e-07	6.04e-07	7.27e-07	7.10e-07	6.52e-07
rs6551253	3	0.28	8.38e-06	8.47e-06	8.78e-06	9.18e-06	9.00e-06	8.92e-06
rs78404737	2	0.10	8.50e-06	9.63e-06	9.78e-06	1.08e-05	1.06e-05	1.00e-05
rs78696065	7	0.02	8.80e-06	1.54e-05	1.55e-05	1.89e-05	1.87e-05	1.75e-05
rs479947	6	0.11	1.19e-05	1.29e-05	1.33e-05	1.44e-05	1.42e-05	1.35e-05
rs566530	6	0.48	1.46e-05	1.40e-05	1.48e-05	1.50e-05	1.47e-05	1.48e-05
rs56355912	10	0.03	1.51e-05	2.16e-05	2.16e-05	2.57e-05	2.54e-05	2.38e-05
rs72733294	5	0.36	1.58e-05	1.60e-05	1.60e-05	1.72e-05	1.69e-05	1.69e-05
rs11074743	16	0.40	1.69e-05	1.68e-05	1.71e-05	1.80e-05	1.77e-05	1.77e-05
rs1562963	11	0.07	2.02e-05	1.99e-05	2.33e-05	2.26e-05	2.23e-05	2.13e-05

6.1 | Rare variants

The difference between the methods becomes even larger when investigating rare variants. We consider the UK Biobank exome sequence data consisting of 45 596 unrelated individuals of European origin. We limit ourselves to White British individuals using the same requirements for the definition of SSTIs as for the common variants. This results in a total number of 30 210 individuals to investigate with 210 individuals defined as cases, once again leading to a case proportion of about 0.7 %. See the Supplementary File for further information about quality control. The principal components are computed as for the common variants analysis, however separately on these 30 210 individuals. We will only consider chromosome 6 rare variants with a minimum minor allele count (MAC) equal to 3. The results are given in Table 2.

TABLE 2 The rare variants with the smallest computed *p*-values using normal approximation to the score test statistic for the GWAS of skin and soft tissue infections. Alternative *p*-value computations are included for comparison.

SNP	CHR	MAC	Norm	ESPA	fastSPA	ESPA-CC	DSPA-CC	fastDSPA-CC
6:26045407:G:A	6	4	2.07e-36	4.31e-05	4.31e-05	2.2e-04	2.2e-04	2.2e-04
6:41097421:T:C	6	4	2.21e-32	4.92e-05	4.92e-05	2.6e-04	2.6e-04	2.5e-04
6:24852645:G:T	6	4	1.37e-25	8.93e-05	8.93e-05	4.4e-04	4.3e-04	4.2e-04
6:31772925:C:A	6	5	6.36e-23	1.3e-04	1.3e-04	6.0e-04	6.0e-04	5.8e-04
6:20402579:C:T	6	3	4.19e-22	0.0020	0.0020	0.010	0.010	0.010
6:132588925:C:T	6	6	8.78e-22	1.5e-04	1.5e-04	6.9e-04	6.9e-04	6.7e-04
6:17675831:G:A	6	3	8.94e-22	0.0020	0.0020	0.010	0.010	0.010
6:110960684:T:G	6	3	2.05e-21	0.0017	0.0017	0.0049	0.0049	0.0049
6:7894854:T:C	6	16	1.88e-20	3.07e-05	3.073e-05	1.2e-04	1.2e-04	1.0e-04
6:148514044:G:T	6	3	1.94e-20	0.0022	0.0022	0.011	0.011	0.011

It is clear that the normal approximation to the score test statistic can be very inaccurate in this setting. However, we also see that the difference between ESPA and the other saddlepoint approximations with continuity correction differ in about one order of magnitude. As a result, we expect the importance of the continuity correction to be most consequential for rare variants. Another observation is that ESPA-CC and DSPA-CC are practically identical in this case. We also see that the speed-up approximation

methods are more accurate which can be explained by observing that the accuracy of the multivariate normal approximation of $\mathbf{U}^*_{\boldsymbol{\beta}}$ in fastDSPA-CC, depends on the number of individuals with $g_i = 0$, which increases for decreasing MACs. The same applies for the approximation of the corresponding normal distribution in fastSPA.

7 | DISCUSSION

We have investigated different saddlepoint approximations for GWAS with binary phenotypes in order to achieve valid *p*-values. We have shown how the saddlepoint approximation introduced in Dey et al.[3] can be interpreted as a two-stage procedure in which one first applies the efficient score to approximate the conditional score test statistic as an unconditional statistic, and then performs single-saddlepoint approximation. We further show how to apply the double saddlepoint approximation to directly approximate the conditional score test statistic.

We distinguish between conditional and overall type I error rate. Taking into account both these measures, we conclude that continuity-corrected saddlepoint approximations are most appropriate for GWAS with case imbalance modelled by logistic regression. The continuity-corrected double saddlepoint approximation, DSPA-CC, and single-saddlepoint approximation, ESPA-CC, using the efficient score are both considered to perform well, however there are situations in which ESPA-CC is somewhat more conservative than DSPA-CC, indicating DSPA-CC to be somewhat more powerful.

There are additional continuity correction variants, and the one used here is called the *second continuity correction*. A first and a third continuity correction are alternatives[7], and specifically the first continuity correction was also investigated with very similar results as when using the second continuity correction, however slightly more inaccurate when considering the intercept model, see Supplementary File. An alternative saddlepoint approximation to the CDF of a random variable is the one introduced in Lugannani and Rice[17], and the expressions for w and v in 10 are intended for the Lugannani and Rice formula, as presented in Butler[7], p.114. This approximation gives the same results as the approximation by Barndorff-Nielsen[8] in most situations. However, we observed in simulations that when the case proportion and MAF approaches zero, the approximation by Lugananni and Rice is inaccurate, see Supplementary File. See for instance Booth and Wood[18] for similar observations in a different application.

An alternative to the double saddlepoint approximation is the sequential saddlepoint approximation, for example Butler[7], p. 323, but the sequential saddlepoint approximation is computationally more demanding than the double saddlepoint approximation. Reid (1996)[19], p. 147, advocates the use of the Barndorff-Nielsen approximation with the sequential saddlepoint approximation (in a situation without lattice variables). Reid (1996)[19] and Reid (2003)[20] presents higher order asymptotics with saddlepoint approximation to tail areas from a likelihood perspective.

Consider the case where one wants to include imputed SNPs. For most imputation methods, the output for each imputed SNP is a probability that the minor allele count is equal to 0, 1 or 2, denoted p_0, p_1 and p_2. Then one must be aware of the fact that when the imputed genotype is set to be the expected minor allele count, $p_1 + 2p_2$, the score test statistic will no longer have a lattice distribution, and so continuity correction does no longer apply. However, to account for imputed SNPs in our method one can instead set the imputed minor allele count to be equal to the most likely allele count according to the imputation method.

Single-variant tests on rare variants are often low-powered, and therefore several region-based tests including several SNPs in the same genetic region have been proposed to gain power. However, many of these methods again rely on single-variant tests as building blocks, among them SKAT and ACAT.[21,22] It is therefore essential that the single-variant tests are sufficiently accurate. How the insights into the score test statistic introduced in this work would impact region-based tests, could be the topic of future research.

8 | ACKNOWLEDGEMENTS

This research was supported by the Norwegian Research Council grant 272402 (PhD Scholarships at SINTEF) as well the funding for research stays abroad for doctoral and postdoctoral fellows financed by the Norwegian Research Council. The research has been conducted using the UK Biobank Resource under Application Number 32285. We thank the Yale Center for Research Computing for guidance and use of the research computing infrastructure. We thank the The Gemini Center for Sepsis Research for establishing cooperation with Yale School of Public Health.

9 | DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from UK Biobank. Restrictions apply to the availability of these data. Data are available for bona fide researchers through an open application.

10 | CODE AVAILABILITY

Source code in R [23] is available at https://github.com/palVJ/SaddlePointApproxInBinaryGWAS.

□

APPENDIX

A PROOFS OF OBSERVATIONS 1-3

Proof of Observation 1. When $g_i \in (0, 1, 2)$, we note that $\boldsymbol{g}^{\mathrm{T}}\boldsymbol{Y}$ is an integer and $\boldsymbol{g}^{\mathrm{T}}\boldsymbol{\mu}$ a constant, so that $U_\gamma = \boldsymbol{g}^{\mathrm{T}}\boldsymbol{Y} - \boldsymbol{g}^{\mathrm{T}}\boldsymbol{\mu}$ has support on a subset of a lattice with step 1. The minimum is obtained for $\boldsymbol{Y} = \boldsymbol{0}$ and the maximum for $\boldsymbol{Y} = \boldsymbol{1}$ (a vector of ones), and the result follows. □

Proof of Observation 2. We assume throughout the proof that the null hypothesis is true, $\gamma = 0$. Denote by V_j the sum of responses Y_i among individuals with genotype $g_i = j$, $j = 0, 1, 2$, and let $V = V_0 + V_1 + V_2 = \sum_{i=1}^{n} Y_i$ be the total sum of responses. With this notation, $U_\gamma = V_1 + 2V_2 - (n_1 + 2n_2)\mu$, and $U_\beta = V - n\mu$, so that the condition $U_\beta = 0$ is equivalent to $V = n\mu$.

The V_j are independent, and V_j is binomially distributed with parameters n_j and μ, $j = 0, 1, 2$, and V is binomially distributed with parameters n and μ. Assume that $v_0 + v_1 + v_2 = n\mu$ with v_j in the support of V_j. Then

$$P(V_0 = v_0, V_1 = v_1, V_2 = v_2 \mid V = n\mu) = \frac{P(V_0 = v_0)P(V_1 = v_1)P(V_2 = v_2)}{P(V = n\mu)}$$
$$= \frac{\binom{n_0}{v_0}\mu^{v_0}(1-\mu)^{n_0-v_0}\binom{n_1}{v_1}\mu^{v_1}(1-\mu)^{n_1-v_1}\binom{n_2}{v_2}\mu^{v_2}(1-\mu)^{n_2-v_2}}{\binom{n}{n\mu}\mu^{n\mu}(1-\mu)^{n-n\mu}} = \frac{\binom{n_0}{v_0}\binom{n_1}{v_1}\binom{n_2}{v_2}}{\binom{n}{n\mu}},$$

a trivariate hypergeometric probability.

Now, $P(U_\gamma = u \mid U_\beta = 0) = P\left(V_1 + 2V_2 = u^* \,\middle|\, V = n\mu\right)$ can be found by summing the above probabilities over $(v_0, v_1, v_2) \in S$. This gives the first sum of the Observation. The more explicit second version of the sum is obtained by solving the two equations in the definition of S for v_0 and v_1 in terms of $k = v_2$. The limits of the sum is determined by the inequalities in the definition of S. □

Proof of Observation 3. We assume throughout the proof that the null hypothesis is true, $\gamma = 0$. Denote by V_j the sum of responses Y_i among individuals with $x_i = 0$ and genotype $g_i = j$, $j = 0, 1, 2$, and let $V = V_0 + V_1 + V_2$. Define similar sums W_j and W for individuals with $x_i = 1$. With this notation, $U_\gamma = V_1 + 2V_2 - (l_1 + 2l_2)\mu_0 + W_1 + 2W_2 - (m_1 + 2m_2)\mu_1$, and $\boldsymbol{U}_\beta^{\mathrm{T}} = \left(V + W - l\mu_0 - m\mu_1 \;\; W - m\mu_1\right)$, so that the condition $\boldsymbol{U}_\beta = \boldsymbol{0}$ is equivalent to $V = l\mu_0$ and $W = m\mu_1$.

All the V_j and W_j are independent, and V_j is binomially distributed with parameters l_j and μ_0, and W_j with parameters m_j and μ_1, $j = 0, 1, 2$. As in the proof of Observation 2, the conditional point probabilites of (V_0, V_1, V_2) given $V = l\mu_0$ and (W_0, W_1, W_2) given $W = m\mu_1$ are trivariate hypergeometric probabilities, and by independence of the two triples, the conditional joint point probability is the product of the two. Then $P(U_\gamma = u \mid \boldsymbol{U}_\beta = \boldsymbol{0})$ can be found by summing those probabilities over $\boldsymbol{s} \in S$. □

B SUPPORT OF THE CONDITIONAL SCORE TEST STATISTIC

Consider the score test statistic of U_γ conditional on $\boldsymbol{U}_\beta = \mathbf{0}$, given by $\boldsymbol{g}^T(\boldsymbol{Y} - \hat{\boldsymbol{\mu}})$. We have $-\hat{\boldsymbol{\mu}} \leq \boldsymbol{Y} - \hat{\boldsymbol{\mu}} \leq \mathbf{1} - \hat{\boldsymbol{\mu}}$ (elementwise inequalities), where $\mathbf{1}$ is a vector of ones. Since all $g_i \geq 0$, premultiplying the inequalities with $\boldsymbol{g}^{\mathrm{T}}$ gives bounds on the support of $\boldsymbol{g}^T(\boldsymbol{Y} - \hat{\boldsymbol{\mu}})$:

$$-\boldsymbol{g}^{\mathrm{T}}\hat{\boldsymbol{\mu}} \leq U_\gamma \leq \boldsymbol{g}^{\mathrm{T}}(\mathbf{1} - \hat{\boldsymbol{\mu}}). \tag{B1}$$

The first equality holds when $\boldsymbol{g}^{\mathrm{T}}\boldsymbol{Y} = 0$ and the second when $\boldsymbol{g}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{g}^{\mathrm{T}}\mathbf{1}$. However, this combination is not achievable if it does not satisfy $\boldsymbol{U}_\beta = \boldsymbol{X}^T(\boldsymbol{Y} - \hat{\boldsymbol{\mu}}) = \mathbf{0}$. Specifically, the minimal and maximal achievable values of the conditional score test statistic is given by the constraint optimization problems:

$$\begin{aligned} \min(U_\gamma) = \min_{\boldsymbol{y}} \quad & \boldsymbol{g}^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) \\ \text{such that} \quad & X^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) = \mathbf{0}, \end{aligned}$$

and

$$\begin{aligned} \max(U_\gamma) = \max_{\boldsymbol{y}} \quad & \boldsymbol{g}^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) \\ \text{such that} \quad & X^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) = \mathbf{0}. \end{aligned}$$

As an example, consider the intercept model with $n = 1000$ and $\boldsymbol{g}$ as in Section 5.1 with $n_0 = 980$, $n_1 = 20$ and $n_2 = 0$ as well as the observation $\sum_{i=1}^{1000} Y_i = 10$. Then $\hat{\mu}_i = 10/1000 = 0.01$ satisfies $U_{\beta_0} = \sum_{i=1}^{1000}(Y_i - \mu_i) = 0$. Then the minimum achievable value is indeed $\min(U_\gamma) = -\boldsymbol{g}^T\hat{\boldsymbol{\mu}} = -0.2$, since we may have a combination where $Y_i = 0$ for all $g_i > 0$, and still get $\sum_{i=1}^{1000} Y_i = 10$. However, $\max(U_\gamma) = 10 - \boldsymbol{g}^T\hat{\boldsymbol{\mu}} = 9.8$ since $\boldsymbol{g}^T\boldsymbol{Y}$ can be no larger than the combinations where $g_i = 1$ for all $Y_i = 1$, which can only occur ten times in order to satisfy $\sum_{i=1}^{1000} Y_i = 10$.

C Solution to $\nabla_{\boldsymbol{t}_\beta} K_\beta(\tilde{\boldsymbol{t}}_\beta) = \mathbf{0}$

Given the marginal cumulant generating function of $\boldsymbol{U}_\beta$, defined by $K_\beta(\boldsymbol{t}_\beta)$ (a function of d variables) with

$$K_\beta(\boldsymbol{t}_\beta) = \sum_{i=1}^{n} \ln(1 - \mu_i + \mu_i \exp(\mathbf{x}_i^T\boldsymbol{t}_\beta)) - \boldsymbol{t}_\beta^T X^T\boldsymbol{\mu}, \tag{C2}$$

and corresponding gradient

$$\nabla_{\boldsymbol{t}_\beta} K_\beta(\boldsymbol{t}_\beta) = \sum_{i=1}^{n} \mu_i \mathbf{x}_i \left(\frac{1}{(1 - \mu_i)\exp(-\mathbf{x}_i^T\boldsymbol{t}_\beta) + \mu_i} - 1 \right). \tag{C3}$$

First, one can easily observe that $\tilde{\boldsymbol{t}}_\beta = \mathbf{0}$ is a solution to $\nabla_{\boldsymbol{t}_\beta} K_\beta(\boldsymbol{t}_\beta) = \mathbf{0}$. Second, if one can prove that the CGF is a convex function, then $\tilde{\boldsymbol{t}}_\beta = \mathbf{0}$ is a unique solution to $\nabla_{\boldsymbol{t}_\beta} K_\beta(\boldsymbol{t}_\beta) = \mathbf{0}$.

Proof. In fact, convexity of a cumulant generating function with *any* random variable $\boldsymbol{U}$, $K(\boldsymbol{t}) = \ln E(e^{\boldsymbol{t}^T\boldsymbol{U}})$, in general follows from the Hölder inequality, $E(|X|^c|Y|^{1-c}) \leq (E|X|)^c(E|Y|)^{1-c}$ for all c in $(0, 1)$, where X and Y are random variables. A function f is convex if $f(c\boldsymbol{t}_1 + (1-c)\boldsymbol{t}_2) \leq cf(\boldsymbol{t}_1) + (1-c)f(\boldsymbol{t}_2)$ for all c in $(0, 1)$. Now,

$$\begin{aligned} K(c\boldsymbol{t}_1 + (1-c)\boldsymbol{t}_2) &= \ln E e^{(c\boldsymbol{t}_1 + (1-c)\boldsymbol{t}_2)^{\mathrm{T}}\boldsymbol{U}} = \ln E\left(e^{c\boldsymbol{t}_1^{\mathrm{T}}\boldsymbol{U}} e^{(1-c)\boldsymbol{t}_2^{\mathrm{T}}\boldsymbol{U}}\right) \\ &\leq \ln\left(\left(Ee^{\boldsymbol{t}_1^{\mathrm{T}}\boldsymbol{U}}\right)^c \left(Ee^{\boldsymbol{t}_2^{\mathrm{T}}\boldsymbol{U}}\right)^{1-c}\right) = c \ln Ee^{\boldsymbol{t}_1^{\mathrm{T}}\boldsymbol{U}} + (1-c)\ln Ee^{\boldsymbol{t}_2^{\mathrm{T}}\boldsymbol{U}} \\ &= cK(\boldsymbol{t}_1) + (1-c)K(\boldsymbol{t}_2), \end{aligned}$$

showing that K is convex. □

References

1. Rogne T, Liyanarachi KV, Rasheed H, et al. GWAS Identifies LINC01184/SLC12A2 as a Risk Locus for Skin and Soft Tissue Infections.. *J Invest Dermatol* 2021.

2. Ma C, Blackwell T, Boehnke M, Scott LJ. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genetic Epidemiology* 2013; 37(6): 539–550.

3. Dey R, Schmidt EM, Abecasis GR, Lee S. A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. *American Journal of Human Genetics* 2017; 101(1): 37–49.

4. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993; 80(1): 27–38.

5. Jannot AS, Ehret G, Perneger T. $P < 5 \times 10^{-8}$ has emerged as a standard of statistical significance for genome-wide association studies. *Journal of Clinical Epidemiology* 2015; 68(4): 460–465.

6. Smyth GK. Pearson's goodness of fit statistic as a score test statistic. *Lecture notes-monograph series* 2003: 115–126.

7. Butler RW. *Saddlepoint Approximations with Applications*. Cambridge Series in Statistical and Probabilistic Mathematics Cambridge University Press . 2007.

8. Barndorff-Nielsen OE. Approximate Interval Probabilities. *Journal of the Royal Statistical Society. Series B (Methodological)* 1990; 52(3): 485–496.

9. Skovgaard IM. Saddlepoint expansions for conditional distributions. *Journal of Applied Probability* 1987; 24(4): 875–887.

10. Cox DR, Reid N. Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society: Series B (Methodological)* 1987; 49(1): 1–18.

11. Waterman RP, Lindsay BG. A simple and accurate method for approximate conditional inference applied to exponential family models. *Journal of the Royal Statistical Society: Series B (Methodological)* 1996; 58(1): 177–188.

12. Bickel PJ, Klaassen CA, Ritov Y, Wellner JA. *Efficient and adaptive estimation for semiparametric models*. 4. Johns Hopkins University Press Baltimore . 1993.

13. Lindsey JK. *Parametric statistical inference*. Oxford University Press . 1996.

14. Dey R, Lee S. *SPAtest: Score Test and Meta-Analysis Based on Saddlepoint Approximation*. 2020. R package version 3.1.2.

15. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 2006; 38(8): 904–909.

16. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. *PLOS Genetics* 2006; 2(12): e190.

17. Lugannani R, Rice S. Saddle point approximation for the distribution of the sum of independent random variables. *Advances in Applied Probability* 1980; 12(2): 475–490. doi: 10.2307/1426607

18. Booth JG, Wood ATA. An example in which the Lugannani-Rice saddlepoint formula fails. *Statistics & Probability Letters* 1995; 23(1): 53–61.

19. Reid N. Likelihood and Higher-Order Approximations to Tail Areas: A Review and Annotated Bibliography. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* 1996; 24(2): 141–166.

20. Reid N. Asymptotics and the Theory of Inference. *The Annals of Statistics* 2003; 31(6): 1695–1731.

21. Wu M, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *American Journal of Human Genetics* 2011; 89(1): 82–93.

22. Liu Y, Chen S, Li Z, Morrison AC, Boerwinkle E, Lin X. ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *The American Journal of Human Genetics* 2019; 104(3): 410–421.

23. R Core Team . *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; Vienna, Austria: 2021.

24. Davison AC. Approximate Conditional Inference in Generalized Linear Models. *Journal of the Royal Statistical Society: Series B (Methodological)* 1988; 50: 445–461. doi: https://doi.org/10.1111/j.2517-6161.1988.tb01740.x

25. Lin DY. An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics* 2005; 21(6): 781–787.

26. Brown LD, Cai TT, DasGupta A. Interval Estimation for a Binomial Proportion. *Statistical Science* 2001; 16(2): 101–133.

Saddlepoint approximations in logistic regression for analysing genome-wide association studies Supplementary File

Pål Vegard Johnsen[1,2], Øyvind Bakke[2], Thea Bjørnland[2], Andrew Thomas DeWan[3], and Mette Langaas[2]

[1]SINTEF Digital, Oslo, Norway
[2]Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway
[3]Department of Chronic Disease Epidemiology and Center for Perinatal, Pediatric and Environmental Epidemiology, Yale School of Public Health

1 Quality assessment of UK Biobank Genetic Data

1.1 Common variants

Analyses were limited to autosomal variants covered by both genotype arrays used over the course of the study and that passed the batch-level quality control. SNPs were included if the call rate was above 99%, the Hardy-Weinberg equilibrium p-value was less than $5 \cdot 10^{-8}$, and the minor allele frequency was larger than 1%. 529 024 SNPs passed these filters.

Individuals were removed if the genetic and reported sex did not match and if the sex chromosomes were not XX or XY. Outliers in heterozygosity and missing rates were removed. The analyses were limited to those identified as Caucasian through the UK Biobank's PCA analysis (field 22006). All individuals had an individual call rate larger than 99%. 366 752 individuals passed these filters. Individuals were removed if the genetic and reported sex did not match and if the sex chromosomes were not XX or XY.

1.2 Rare variants

Analyses were limited to autosomal variants at chromosome 6. Details of quality assessment of the sequenced exomes from 49 960 UKB participants is given in Van Hout et al. (2019). For further quality assessment, out of these 49 960 participants, the analysis were limited to those identified as Caucasian through the UK Biobank's PCA analysis (field 22006). Individuals were removed if the genetic and reported sex did not match and if the sex chromosomes were not XX or XY. The SNPs had a MAF less than 0.01, but a MAC larger than two. All SNPs had a missing rate less than 0.01, and all individuals had an individual call rate larger than 99%.

2 Approximating the double saddlepoint method by a normal approximation to $\boldsymbol{U}_\beta$

By a double saddlepoint method we may, under the null, estimate

$$P(U_\gamma = u_\gamma | \boldsymbol{U}_{\boldsymbol{\beta}} = \mathbf{0}) = \frac{f(\mathbf{0}, u_\gamma)}{f_{\boldsymbol{\beta}}(\mathbf{0})}, \tag{1}$$

by using saddlepoint techniques to approximate the joint distribution f of $\boldsymbol{U}_{\boldsymbol{\beta}}$ and U_γ at $\boldsymbol{U}_{\boldsymbol{\beta}} = \mathbf{0}$, and the marginal distribution $f_{\boldsymbol{\beta}}$ of $\boldsymbol{U}_{\boldsymbol{\beta}}$ at $\boldsymbol{U}_{\boldsymbol{\beta}} = \mathbf{0}$.

2.1 The joint cumulant generating function

Let $\boldsymbol{t}$ be a vector of dimension $d+1$, which we partition into the vector $\boldsymbol{t}_\beta$ of dimension d and the scalar t_γ. Since $Y_i \sim \text{binomial}(\mu_i)$, the CGF of $\boldsymbol{U}$ may then be expressed as

$$K(\boldsymbol{t}) = K(\boldsymbol{t}_{\boldsymbol{\beta}}, t_\gamma) = \sum_{i=1}^{n} \ln\left(1 - \mu_i + \mu_i \exp(g_i t_\gamma + \boldsymbol{t}_{\boldsymbol{\beta}}^T \boldsymbol{x}_i)\right) - t_\gamma \sum_{i=1}^{n} g_i \mu_i - \boldsymbol{t}_{\boldsymbol{\beta}}^T \sum_{i=1}^{n} \boldsymbol{x}_i \mu_i. \tag{2}$$

Derivatives of $K(\boldsymbol{t}_\beta, t_\gamma)$ with respect to $\boldsymbol{t}_\beta$ and t_γ, denoted $\nabla_{\boldsymbol{t}_{\boldsymbol{\beta}}} K(\boldsymbol{t}_\beta, t_\gamma)$ and $\frac{\partial}{\partial t_\gamma} K(\boldsymbol{t}_\beta, t_\gamma)$ respectively, are

$$\nabla_{\boldsymbol{t}_{\boldsymbol{\beta}}} K(\boldsymbol{t}_{\boldsymbol{\beta}}, t_\gamma) = \sum_{i=1}^{n} \mu_i \left(\frac{\exp(g_i t_\gamma + \boldsymbol{t}_{\boldsymbol{\beta}}^T \boldsymbol{x}_i)}{\left(1 - \mu_i + \mu_i \exp(g_i t_\gamma + \boldsymbol{t}_{\boldsymbol{\beta}}^T \boldsymbol{x}_i)\right)} - 1 \right) \boldsymbol{x}_i,$$

and

$$\frac{\partial}{\partial t_\gamma} K(\boldsymbol{t}_{\boldsymbol{\beta}}, t_\gamma) = \sum_{i=1}^{n} \mu_i \left(\frac{\exp(g_i t_\gamma + \boldsymbol{t}_{\boldsymbol{\beta}}^T \boldsymbol{x}_i)}{\left(1 - \mu_i + \mu_i \exp(g_i t_\gamma + \boldsymbol{t}_{\boldsymbol{\beta}}^T \boldsymbol{x}_i)\right)} - 1 \right) g_i,$$

so that the gradient of $K(\boldsymbol{t}_{\boldsymbol{\beta}}, t_\gamma)$, denoted $\nabla K(\boldsymbol{t}_{\boldsymbol{\beta}}, t_\gamma)$, may be expressed as

$$\nabla K(\boldsymbol{t}_{\boldsymbol{\beta}}, t_\gamma) = \begin{pmatrix} \nabla_{\boldsymbol{t}_{\boldsymbol{\beta}}} K(\boldsymbol{t}_{\boldsymbol{\beta}}, t_\gamma) \\ \frac{\partial}{\partial t_\gamma} K(\boldsymbol{t}_{\boldsymbol{\beta}}, t_\gamma) \end{pmatrix}.$$

Let $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\beta}^{\mathrm{T}} & \gamma \end{pmatrix}^{\mathrm{T}}$ denote the full parameter set, and define a diagonal matrix $M^{\boldsymbol{\theta}}$ with entries

$$M_{ii}^{\boldsymbol{\theta}} = \frac{\mu_i(1-\mu_i)\exp(-g_i t_\gamma - \boldsymbol{t}_{\boldsymbol{\beta}}^T \boldsymbol{x}_i)}{\left((1-\mu_i)\exp(-g_i t_\gamma - \boldsymbol{t}_{\boldsymbol{\beta}}^T \boldsymbol{x}_i) + \mu_i\right)^2}.$$

The Hessian of K, denoted $H(\boldsymbol{t})$, can be expressed as

$$H(\boldsymbol{t}) = \begin{bmatrix} \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} K(\boldsymbol{t}_{\boldsymbol{\beta}}, t_\gamma) & \frac{\partial^2}{\partial \boldsymbol{\beta} \partial t_\gamma} K(\boldsymbol{t}_{\boldsymbol{\beta}}, t_\gamma) \\ \frac{\partial^2}{\partial t_\gamma \partial \boldsymbol{\beta}} K(\boldsymbol{t}_{\boldsymbol{\beta}}, t_\gamma) & \frac{\partial^2}{\partial t_\gamma^2} K(\boldsymbol{t}_{\boldsymbol{\beta}}, t_\gamma) \end{bmatrix} = \begin{bmatrix} X^T M \boldsymbol{X} & X^T M \boldsymbol{g} \\ \boldsymbol{g}^T M X & \boldsymbol{g}^T M \boldsymbol{g} \end{bmatrix}.$$

2.2 The marginal cumulant generating function

The cumulant generating function of $\boldsymbol{U}_{\boldsymbol{\beta}}$, denoted $K_{\boldsymbol{\beta}}(\boldsymbol{t}_{\boldsymbol{\beta}})$, is given by

$$K_{\boldsymbol{\beta}}(\boldsymbol{t}_{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \ln\left(1 - \mu_i + \mu_i \exp\left(\boldsymbol{t}_{\boldsymbol{\beta}}^T \boldsymbol{x}_i\right)\right) - \sum_{i=1}^{n} \boldsymbol{t}_{\boldsymbol{\beta}}^T \boldsymbol{x}_i \mu_i. \tag{3}$$

The gradient, denoted $\nabla K_{\boldsymbol{\beta}}(\boldsymbol{t}_{\boldsymbol{\beta}})$, is

$$\nabla K_{\boldsymbol{\beta}}(\boldsymbol{t}_{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \mu_i \left(\frac{\exp(\boldsymbol{t}_{\boldsymbol{\beta}}^T \boldsymbol{x}_i)}{\left(1 - \mu_i + \mu_i \exp(g_i t_\gamma + \boldsymbol{t}_{\boldsymbol{\beta}}^T \boldsymbol{x}_i)\right)} - 1 \right) \boldsymbol{x}_i,$$

and the Hessian, denoted $H_{\boldsymbol{\beta}}(\boldsymbol{t}_{\boldsymbol{\beta}})$, is

$$H_{\boldsymbol{\beta}}(\boldsymbol{t}_{\boldsymbol{\beta}}) = X^T M^{\boldsymbol{\beta}} X,$$

where $M^{\boldsymbol{\beta}}$ is a diagonal matrix with entries

$$M_{ii}^{\boldsymbol{\beta}} = \frac{\mu_i(1-\mu_i)\exp(-\boldsymbol{t}_{\boldsymbol{\beta}}^T \boldsymbol{x}_i)}{\left((1-\mu_i)\exp(-\boldsymbol{t}_{\boldsymbol{\beta}}^T \boldsymbol{x}_i) + \mu_i\right)^2}.$$

We note that

$$K(\boldsymbol{t}_{\boldsymbol{\beta}}, 0) = K_{\boldsymbol{\beta}}(\boldsymbol{t}_{\boldsymbol{\beta}}). \tag{4}$$

2.3 Double-saddlepoint approximation

The saddlepoint approximation of the probability distribution of the score vector $\boldsymbol{U}$ evaluated at $\boldsymbol{U}_{\boldsymbol{\beta}} = 0$ is given by

$$\hat{f}(\mathbf{0}, u_\gamma) = (2\pi)^{-(d+1)/2} |H(\hat{\boldsymbol{t}})|^{-1/2} \exp\left\{K(\hat{\boldsymbol{t}}_{\boldsymbol{\beta}}, \hat{t}_\gamma) - \hat{t}_\gamma u_\gamma)\right\},$$

where $\begin{pmatrix}\hat{\boldsymbol{t}}_{\boldsymbol{\beta}}^{\mathrm{T}} & \hat{t}_\gamma\end{pmatrix}^{\mathrm{T}}$ is the $(d+1)$-dimensional saddlepoint that solves $K'(\hat{\boldsymbol{t}}_{\boldsymbol{\beta}}, \hat{t}_\gamma) = \begin{pmatrix}\mathbf{0}^{\mathrm{T}} & u_\gamma\end{pmatrix}^{\mathrm{T}}$. The saddlepoint approximation of the marginal distribution of $\boldsymbol{U}_{\boldsymbol{\beta}}$, evaluated at $\boldsymbol{U}_{\boldsymbol{\beta}} = \mathbf{0}$ is similarly

$$\hat{f}_{\boldsymbol{\beta}}(\mathbf{0}) = (2\pi)^{-d/2} |H_{\boldsymbol{\beta}}(\tilde{\boldsymbol{t}}_{\boldsymbol{\beta}})|^{-1/2} \exp\left\{K_{\boldsymbol{\beta}}(\tilde{\boldsymbol{t}}_{\boldsymbol{\beta}})\right\},$$

where $\tilde{\boldsymbol{t}}_{\boldsymbol{\beta}}$ is the d-dimensional saddlepoint that solves $\nabla K_{\boldsymbol{\beta}}(\boldsymbol{t}_{\boldsymbol{\beta}}) = \mathbf{0}$. We showed in Appendix B that $\tilde{\boldsymbol{t}}_{\boldsymbol{\beta}} = \mathbf{0}$, hence

$$\hat{f}_{\boldsymbol{\beta}}(\mathbf{0}) = (2\pi)^{-d/2} |H_{\boldsymbol{\beta}}(\mathbf{0})|^{-1/2} \exp\left\{K_{\boldsymbol{\beta}}(\mathbf{0})\right\}.$$

2.4 Speed-up algorithm

Starting with the joint CGF $K(\boldsymbol{t}_{\boldsymbol{\beta}}, t_\gamma)$ in Equation (2) we split the sum into the sets of individuals with and without minor alleles:

$$\begin{aligned} K(\boldsymbol{t}_{\boldsymbol{\beta}}, t_\gamma) = &\sum_{i=1}^{m} \ln\left(1 - \mu_i + \mu_i \exp(g_i t_\gamma + \boldsymbol{t}_{\boldsymbol{\beta}}^T \boldsymbol{x}_i)\right) - t_\gamma \sum_{i=1}^{m} g_i \mu_i - \boldsymbol{t}_{\boldsymbol{\beta}}^T \sum_{i=1}^{m} \boldsymbol{x}_i \mu_i \\ &+ \sum_{i=m+1}^{n} \ln\left(1 - \mu_i + \mu_i \exp(\boldsymbol{t}_{\boldsymbol{\beta}}^T \boldsymbol{x}_i)\right) - \boldsymbol{t}_{\boldsymbol{\beta}}^T \sum_{i=m+1}^{n} \boldsymbol{x}_i \mu_i. \end{aligned}$$

By comparing with Equation (3), the last two terms is in fact the part of the cumulant generating function of $\boldsymbol{U}_\beta$ restricted to individuals with $g_i = 0$, denoted $K^*_{\boldsymbol{\beta}}(\boldsymbol{t}_{\boldsymbol{\beta}})$. As discussed in Dey et al. (2017), if the non-genetic covariates are not particularly skewed, then a normal approximation to $\boldsymbol{U}_{\boldsymbol{\beta}}$ may be accurate. If there are few individuals with $g_i > 0$, which is typically the case, this would imply that a normal approximation of $\boldsymbol{U}^*_{\boldsymbol{\beta}}$, the part of $\boldsymbol{U}_{\boldsymbol{\beta}}$ with $g_i = 0$, may also be accurate. Therefore, let $X_{g_i=0}$, $\boldsymbol{Y}_{g_i=0}$ and $\boldsymbol{\mu}_{g_i=0}$ denote the part of X, $\boldsymbol{Y}$ and $\boldsymbol{\mu}$ with $g_i = 0$, and so $\boldsymbol{U}^*_{\boldsymbol{\beta}} = X^T_{g_i=0}(\boldsymbol{Y}_{g_i=0} - \boldsymbol{\mu}_{g_i=0})$ with $E(\boldsymbol{U}^*_{\boldsymbol{\beta}}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{U}^*_{\boldsymbol{\beta}}) = X^T_{g_i=0} W_{g_i=0} X_{g_i=0}$, with $W_{g_i=0}$ the submatrix of the diagonal matrix W with entries $\mu_i(1-\mu_i)$ among those individuals with $g_i = 0$. By approximating $\boldsymbol{U}^*_{\boldsymbol{\beta}}$ to have a normal distribution, the approximation of the CGF of $U^*_{\boldsymbol{\beta}}$ is $K^*_{\boldsymbol{\beta}}(\boldsymbol{t}_{\boldsymbol{\beta}}) \approx \frac{1}{2} \boldsymbol{t}^T_{\boldsymbol{\beta}} \mathrm{Cov}(\boldsymbol{U}^*_{\boldsymbol{\beta}}) \boldsymbol{t}_{\boldsymbol{\beta}}$. And consequently, the original CGF may be approximated as

$$\begin{aligned} K(\boldsymbol{t}_{\boldsymbol{\beta}}, t_\gamma) \approx &\sum_{i=1}^{m} \ln\left(1 - \mu_i + \mu_i \exp(g_i t_\gamma + \boldsymbol{t}_{\boldsymbol{\beta}}^T \boldsymbol{x}_i)\right) - t_\gamma \sum_{i=1}^{m} g_i \mu_i - \boldsymbol{t}_{\boldsymbol{\beta}}^T \sum_{i=1}^{m} \boldsymbol{x}_i^T \mu_i \\ &+ \frac{1}{2} \boldsymbol{t}^T_{\boldsymbol{\beta}} \mathrm{Cov}(\boldsymbol{U}^*_{\boldsymbol{\beta}}) \boldsymbol{t}_{\boldsymbol{\beta}}. \end{aligned}$$

This approximation to the CGF does not represent any reasonable speed-up yet, since $\mathrm{Cov}(\boldsymbol{U}^*_{\boldsymbol{\beta}})$ must be computed for each genetic variant and requires $O(n-m)$ calculations. However, we may express for each variant $\mathrm{Cov}(\boldsymbol{U}^*_{\boldsymbol{\beta}}) = \mathrm{Cov}(\boldsymbol{U}_{\boldsymbol{\beta}}) - \mathrm{Cov}(\boldsymbol{U}^\dagger_{\boldsymbol{\beta}})$, with $\mathrm{Cov}(\boldsymbol{U}_{\boldsymbol{\beta}})$ the same for all variants, while $\boldsymbol{U}^\dagger_{\boldsymbol{\beta}}$ is the part of $\boldsymbol{U}_{\boldsymbol{\beta}}$ with $g_i > 0$, and so $\mathrm{Cov}(\boldsymbol{U}^\dagger_{\boldsymbol{\beta}}) = X^T_{g_i>0} W_{g_i>0} X_{g_i>0}$. As $\mathrm{Cov}(\boldsymbol{U}_{\boldsymbol{\beta}})$ can be precomputed for all variants, this requires $O(m)$ calculations. Hence, the approximation

$$\begin{aligned} K(\boldsymbol{t}_{\boldsymbol{\beta}}, t_\gamma) \approx &\sum_{i=1}^{m} \ln\left(1 - \mu_i + \mu_i \exp(g_i t_\gamma + \boldsymbol{t}_{\boldsymbol{\beta}}^T \boldsymbol{x}_i)\right) - t_\gamma \sum_{i=1}^{m} g_i \mu_i - \boldsymbol{t}_{\boldsymbol{\beta}}^T \sum_{i=1}^{m} \boldsymbol{x}_i^T \mu_i \\ &+ \frac{1}{2} \boldsymbol{t}^T_{\boldsymbol{\beta}} \left(\mathrm{Cov}(\boldsymbol{U}_{\boldsymbol{\beta}}) - \mathrm{Cov}(\boldsymbol{U}^\dagger_{\boldsymbol{\beta}})\right) \boldsymbol{t}_{\boldsymbol{\beta}} \end{aligned}$$

is computed only for those individuals with $g_i > 0$ which leads to a substantial reduction in running time from $O(n-m)$ to $O(m)$ calculations when $m << n$, which is typically the case, and particularly relevant for rare variants.

Similarly as for the normal approximation of $K(\boldsymbol{t}_{\boldsymbol{\beta}}, t_\gamma)$, the normal approximation of $K_{\boldsymbol{\beta}}(\boldsymbol{t}_{\boldsymbol{\beta}})$ is given by

$$K_{\boldsymbol{\beta}}(\boldsymbol{t}_{\boldsymbol{\beta}}) \approx \frac{1}{2}\boldsymbol{t}_{\boldsymbol{\beta}}^T \operatorname{Cov}(\boldsymbol{U}_{\boldsymbol{\beta}})\boldsymbol{t}_{\boldsymbol{\beta}}.$$

3 Simulations of genetic association studies with an imbalanced response

We will in detail explain the simulations given in Section 5.3. The simulation set-up is motivated by Dey et al. (2017) by conditioning on the which individuals are cases and controls, say $\boldsymbol{y} = (\mathbf{0}_c, \mathbf{0}_{n-c}^T)$, with $\mathbf{0}_c$ and $\mathbf{0}_{n-c}$ vectors of zeros with size c (the number of cases) and $n-c$ (the number of controls). We consider the logistic regression model

$$\operatorname{logit}(\mu_i) = \beta_0 + x_{i,1} + x_{i,2} + \gamma g_i,$$

with $X_1 \sim$ Bernoulli(0.5), $X_2 \sim N(0,1)$ and $G \sim \operatorname{binom}(2, \text{MAF})$ mutually independent. Since we are evaluating validity of tests, we set $\gamma = 0$. For each iteration, the genotype vector is sampled independently of $\boldsymbol{y}$, while the nuisance covariates for each individual are sampled conditionally on $\boldsymbol{y}$. In each iteration, and for each method, we record whether a false rejection has occurred. With a total of 10^9 iterations, we estimate the type I error rate conditioned on the constant phenotype vector $\boldsymbol{y}$ when applying SPA, ESPA-CC and DSPA-CC. We want β_0 to be such that the disease prevalence to be 1% in the population, i.e. $P(Y=1) = 0.01$. That is we want:

$$\begin{aligned}
P(Y=1) &= \int_{x_1=-\infty}^{\infty} \sum_{x_1=0}^{1} P(Y=1|x_1,x_2)P(x_1,x_2)dx_1 \\
&= P(Y=1) = \int_{x_1=-\infty}^{\infty} \sum_{x_1=0}^{1} P(Y=1|x_1,x_2)P(x_1)P(x_2)dx_1 \\
&= 0.5\sqrt{\frac{1}{2\pi}} \cdot \int_{x_2=-\infty}^{\infty} \exp(-0.5x_2^2)\left(\frac{1}{1+\exp(-\beta_0 - x_2)} + \frac{1}{1+\exp(-\beta_0 - 1 - x_2)}\right) dx_2 = 0.01
\end{aligned} \tag{5}$$

A solution is $\beta_0 = -5.6$.

Given the vector of phenotype values $\boldsymbol{y}$, the nuisance covariates need to be sampled according to their conditional probabilities:

$$\begin{aligned}
P(x_1|y) &= P(x_1,y)/P(y) = P(y|x_1)P(x_1)/P(y) \\
&= P(x_1)/P(y_1)\int_{x_2=-\infty}^{\infty} P(y|x_1,x_2)P(x_2)dx_2,
\end{aligned} \tag{6}$$

and

$$\begin{aligned}
P(x_2|y) &= P(x_2,y)/P(y) = P(y|x_2)P(x_2)/P(y) \\
&= P(x_2)/P(y_1)\sum_{x_1=0}^{1} P(y|x_1,x_2)P(x_2),
\end{aligned} \tag{7}$$

Therefore from (6) (with prevalence 0.01):

$$P(X_1 = x_1|y=1) = 50\sqrt{\frac{1}{2\pi}} \cdot \int_{x_2=-\infty}^{\infty} \frac{\exp(-0.5x^2)}{1+\exp(-\beta_0 - x_1 - x_2)}dx_2, \tag{8}$$

and similarly one can compute $P(X_1 = x_1|y=0)$ for each value of x_1. For the sampling of x_2 conditional on y we will get for instance:

$$P(x_2|y=1) = 50\sqrt{\frac{1}{2\pi}}\exp(-0.5x^2)\left(\frac{1}{1+\exp(-\beta_0 - x_2)} + \frac{1}{1+\exp(-\beta_0 - 1 - x_2)}\right). \quad (9)$$

The continuous probability distribution of $P(x_2|y=1)$, as well as for $P(x_2|y=0)$, does not belong to any known distribution class, however one can see that $P(x_2|y=1) \leq \phi(x_2)$ for all x_2 with $\phi()$ the standard normal distribution. Therefore the standard normal can be used as a proposal distribution in a *rejection sampling* procedure. However, the large amount of simulations needed requires a faster approach as the efficiency in the rejection sampling depends on how close the proposal distribution resembles the true distribution. It can be shown in this particular case that the efficiency decreases for decreasing prevalence ($P(Y=1)$). Since the probability distribution of $P(x_2|y=1)$ as well as $P(x_2|y=0)$ can be shown to be log-concave, one can apply the much more efficient *adaptive rejection sampling* procedure, where the proposal distribution is adaptively improved during the iterations.

4 PCA plots

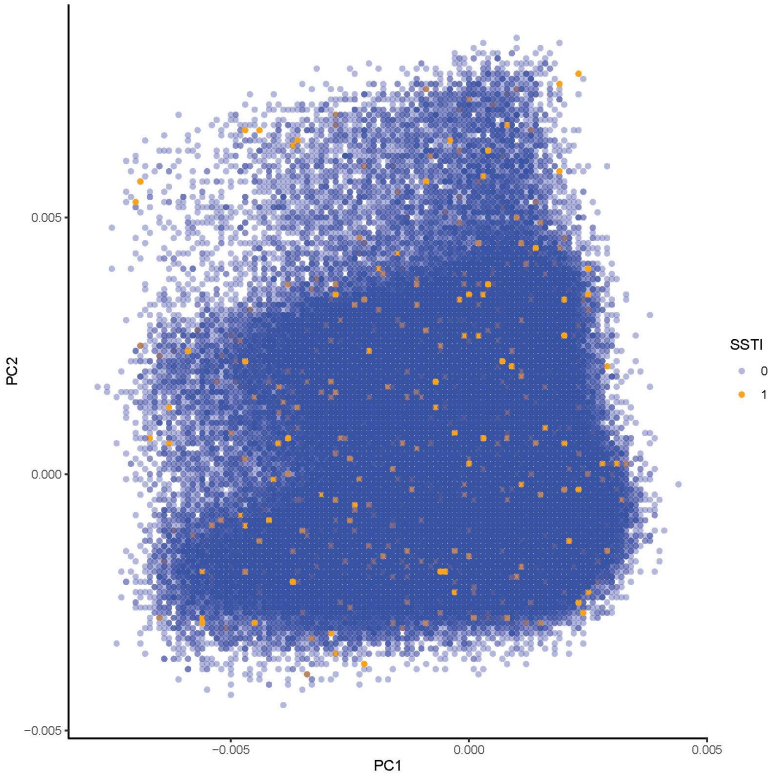

Figure 1: PCA plot of first and second principal components when analysing the common variants.

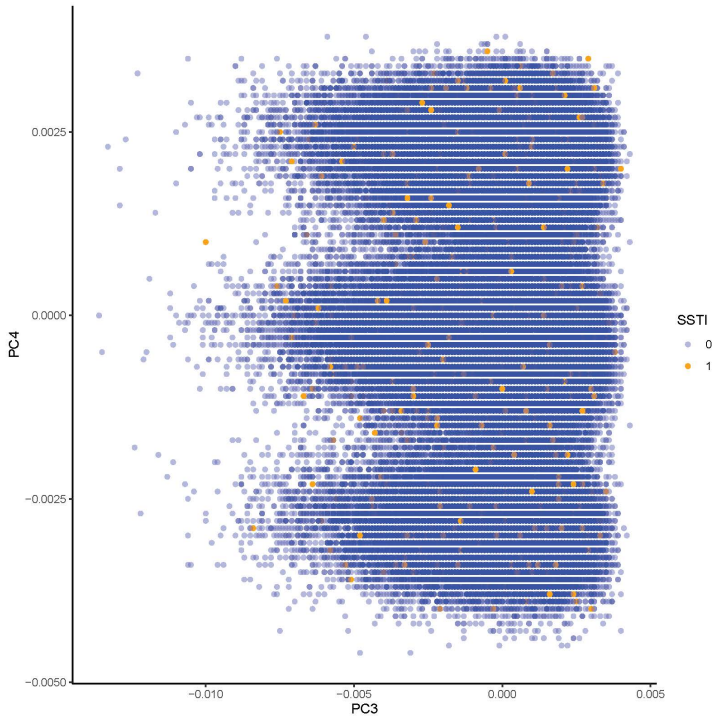

Figure 2: PCA plot of third and fourth principal components when analysing the common variants.

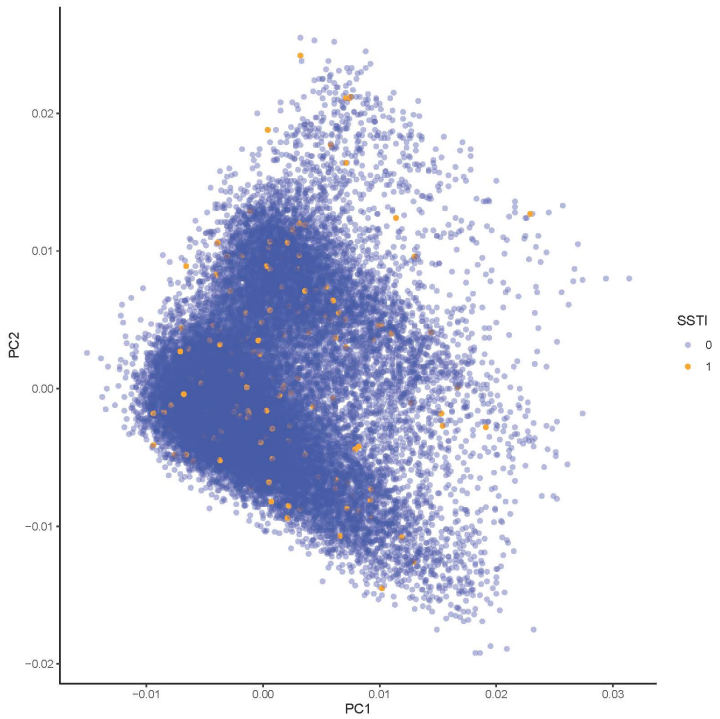

Figure 3: PCA plot of first and second principal components when analysing the rare variants.

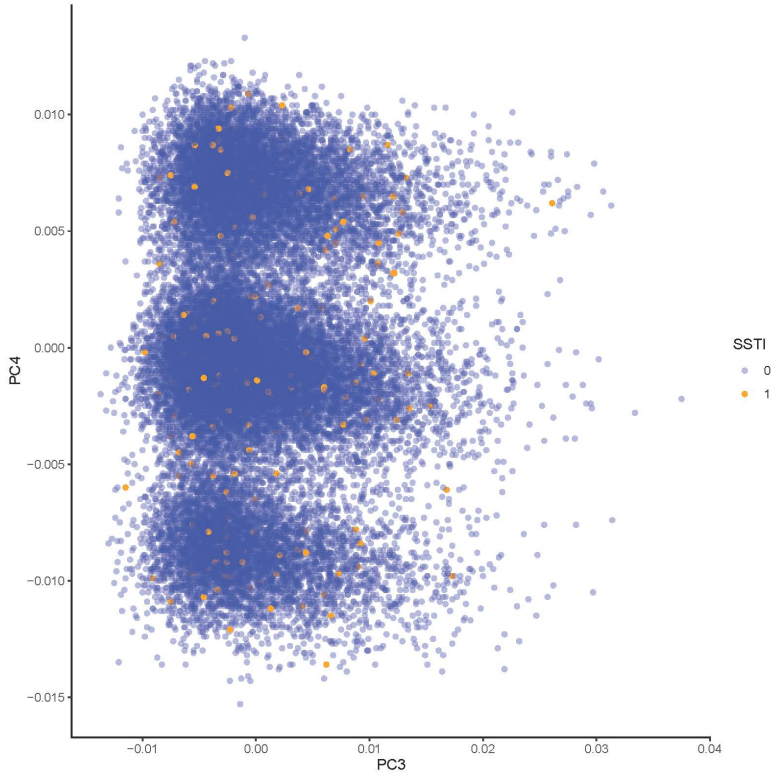

Figure 4: PCA plot of first and second principal components when analysing the rare variants.

References

Dey, R., Schmidt, E. M., Abecasis, G. R., & Lee, S. (2017). A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. *American Journal of Human Genetics*, *101*(1), 37–49.

Van Hout, C. V., Tachmazidou, I., Backman, J. D., Hoffman, J. X., Ye, B., Pandey, A. K., Gonzaga-Jauregui, C., Khalid, S., Liu, D., Banerjee, N., Li, A. H., Colm, O., Marcketta, A., Staples, J., Schurmann, C., Hawes, A., Maxwell, E., Barnard, L., Lopez, A., ... on behalf of the Regeneron Genetics Center. (2019). *Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank* (tech. rep.).