2-11-2009

EFFICIENT EVALUATION OF RANKING PROCEDURES WHEN THE NUMBER OF UNITS IS LARGE WITH APPLICATION TO SNP IDENTIFICATION

Thomas A. Louis
Johns Hopkins University Bloomberg School of Public Health, tlouis@jhsph.edu

Ingo Ruczinski
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

EFFICIENT EVALUATION OF RANKING PROCEDURES WHEN THE NUMBER OF UNITS IS LARGE WITH APPLICATION TO SNP IDENTIFICATION[1]

Thomas A. Louis [2] and Ingo Ruczinski
Department of Biostatistics
Johns Hopkins Bloombergh SPH

February 9, 2009

Summary

Simulation-based assessment is a popular and frequently necessary approach to evaluation of statistical procedures. Sometimes overlooked is the ability to take advantage of underlying mathematical relations and we focus on this aspect. We show how to take advantage of large-sample theory when conducting a simulation using the analysis of genomic data as a motivating example. The approach uses convergence results to provide an approximation to smaller-sample results, results that are available only by simulation. We consider evaluating and comparing a variety of ranking-based methods for identifying the most highly associated SNPs in a genome-wide association study, derive integral equation representations of the pre-posterior distribution of percentiles produced by three ranking methods, and provide examples comparing performance. These results are of interest in their own right and set the framework for a more extensive set of comparisons.

KEYWORDS: Efficient Simulation, Ranking Procedures, SNP identification

[1] Prepared in honor of Hans van Houwelingen.
[2] To whom email should be addressed, tlouis@jhsph.edu

1 INTRODUCTION

Hans van Houwelingen's impact on statistical theory and practice is both broad and deep. His many contributions are detailed elsewhere in this volume and we mention only a few. Hans has made signature methodological contributions to non- and semi-parametric empirical Bayes, to large sample theory in several contexts, to cross-validation and other sample reuse methods. His applied collaborations are extensive and, as with his methods work, his impact is "zeer groot."

Simulation-based assessment is a popular and frequently necessary approach to evaluation of statistical procedures. Approaches range from the most straightforward, but generally inefficient (prepare a straightforward program and just blast off) to more sophisticated approaches using importance sampling (to enable generation of the appropriate stochastic sequences or to improve efficiency), rejection sampling or Markov Chain Monte-Carlo (to enable generation of the appropriate stochastic sequences). Carlin and Louis (2009) provide details on these and others. Sometimes overlooked is the ability to take advantage of underlying mathematical relations and we focus on this aspect. We show how to take advantage of large-sample convergence of empirical distributions to develop mathematical relations that provide an excellent approximation to the results of a large-scale simulation. When components of these relations are available analytically, the approximations can replace simulations. If some components are themselves available only by simulation, then in some cases the mathematical relations can be used to improve efficiency. This use of two types of CPUs (one biologic and one in silico) most definitely represents a strategic approach that Hans would take. We trust that this article, at least to a small degree, captures the spirit of his innovation, creativity and eclectic use of statistical theory and simulation-based evaluations.

Hans has made important contributions to genomics analysis, in particular with regard to statistical methodology for dimension reduction in association studies (Goeman et al. (2006), Heidema et al. (2007), Uh et al. (2007)), and efficient statistical methods for linkage and association studies that comprehensively employ all information avaiable (Houwing-Duistermaat et al. (2007), Lebrec et al. (2008)). We motivate our approach and provide examples in the setting of Genome-Wide Association studies (GWAs). Specifically, we consider evaluating and comparing performance of ranking-based methods that can be used to identify the most highly associated SNPs in a GWAs.

1.1 Genome-wide Association Studies

A single-nucleotide polymorphism (SNP) is a locus in the DNA where some appreciable variation in the nucleotide sequence between the members of a species occurs. In humans for example, each person has one pair of each of the autosomes (chromosomes 1-22), with one copy inherited from each parent. While almost all loci have identical nucleotides in each DNA strand for each person, about every 500 base pair positions some differences between people exist. Interestingly, the vast majority of all SNPs are bi-allelic, that is, only 2 out of the 4 nucleotides occur in the population. For a locus with variation to be defined as a SNP, the frequency with which the minor allele occurs in the population is usually required to be above 1%. Since there are substantial differences in the genetic background between ethnic groups or geographically separated populations, the actual minor allele frequencies (MAFs) between sub-populations can differ significantly. The bi-allelic SNPs are typically encoded as nucleotide pairs within a person's matched chromosome pairs (e. g. (CC,CT,TT)), or specifically for SNPs typed in genomic arrays, as a generic factor with three levels (e. g. (AA,AB,BB)). In the statistical modeling of SNP data, the SNPs are typically encoded as the number of variant alleles (0, 1, or 2), where the variant allele (as opposed to the common allele) is defined as the allele that occurs less frequent in the population. In a "homogeneous" population (and under certain assumptions such as random mating), the frequencies of the nucleotide pairs that define a SNP is governed by the Hardy-Weinberg equilibrium, producing prevalences of $(p^2, 2pq, q^2)$, with p being the major and $q = 1 - p$ the minor allele frequency. A typical GWAs assesses association between candidate genotypes and a disorder or a quantitative trait (the phenotype), often via Z-scores, testing the hypothesis of no association between the genotype (AA, AB, BB) and the phenotype. Then, these association measures are ranked and the most highly associated SNPs identified. See Ruczinski et al. (2009) for additional details. To detect the genetic causes of complex human disease, genome-wide association studies have become the method of choice for many researchers around the globe. Consistent with this is the sheer abundance of GWAs papers that have been published in the last couple of years.

1.2 Issues in Analyzing GWAs Data

The predominant strategy for analyzing GWAs data is to carry out SNP specific (marginal) tests such as the Cochran-Armitage trend test, sort the results from the smallest to the largest p-value, and declare significance based on a Bonferroni correction to limit the family-wise error rate at a fixed level (typically, 5%). The benefit of this approach is its straightforward and reproducible implementation, and has led to the identification of important genes for many diseases (see Manolio et al. (2008) for an overview). However, both from scientific and statistical perspectives, this approach is sub-optimal for various reasons.

Enforcing a tight control on the family-wise error rate, i.e. the probability of declaring at least one SNP significant that is not associated with the phenotype, comes at the expense of the truly associated SNPs which do not achieve "genome wide" significance. It appears that this strategy works in strong opposition to the aims of a GWAs, typically declared to be *hypothesis generating* and *discovery* oriented. This stringent type I error control is particularly worrisome in the investigation of truly complex diseases such as asthma or schizophrenia, since many SNPs, each typically with a very modest effect size, contribute to the disease risk. Unless the sample size is huge, none of the SNPs might reach overall significance after multiple comparisons corrections. From a Bayesian perspective, a Bonferroni adjustment with very small α can be justified if one has a strong prior belief that all the nulls are true (Westfall et al. (1997)), however, the opposite is the case in GWAs. Moreover, controlling the family wise error rate via Bonferroni completely ignores type II errors, and employs a uniform threshold for significance that ignores power. Using such a constant threshold in a frequentist setting leads to an inconsistent procedure, and viewed from a decision theory perspective, leads to an inadmissible procedure (Wakefield (2007, 2008a,b)).

Employing the family wise error rate is certainly reasonable if the confirmation of a list of candidate SNPs believed to be associated with the phenotype is the objective, for example in sequencing or fine mapping of genomic regions. However, for many complex diseases the a priori information of the genetic base is very limited, and GWAs are usually carried out as truly hypothesis generating. Thus, the objective is to generate a list of SNPs for further investigation about their role in the genetic underpinning of the disease, or their correlation with truly disease causing variants via fine mapping. The generation of such a SNP list obviously relies on proper ranking methods, and should take the conditional power of each variant into account. In this manuscript, we focus on such ranking procedures.

Fully Bayesian approaches that directly address ranking rather than type I error control have been developed (Lin et al., 2006) and we are interested in comparing the performance these with the current approaches. However, a GWAs can include more than 1 million candidate SNPs (and the number is growing), so computational demands in a simulation-based approach can be substantial. However, fortunately, when the number of candidate SNPs is large, asymptotic consistency of empirical distributions can be used to speed up the assessment.

1.3 Organization

In section 2 we outline candidate ranking methods including Z-score bases and Bayesian optimal approaches (see Lin et al. (2006)). Sections 2.3 and 3 derive distributions of percentile ranks for a spiked SNP when K is large. Section 4 reports on comparisons of candidate ranking methods for mathematically evaluable models. Section 5 presents some cautions and section 6 sums up.

2 RANKING METHODS

A variety of ranking methods are in use (see Lin et al. (2006)). We focus on ranks/percentiles based on hypothesis test Z-scores and two Bayes optimal approaches. We rank $(K+1)$ SNPs, where $k = 1, \ldots, K$ are "null" SNPs (with null or moderate association with the phenotype), and $k = 0$ is the spiked SNP, a SNP with a stochastically larger association than for the null SNPs.

2.1 The Z-score Approach

For each SNP, compute a Z-score which is used to test the hypothesis of no association, producing $(Z_0, Z_1, \ldots, Z_K)$. Rank these and divide the ranks by $(K + 2)$ (one more that the number of SNPs), producing $\ddot{P}$ and $\ddot{P}_k$. In a GWAs, Z-scores are often produced by a trend test. In the appendix, we outline the score test approach in Ruczinski et al. (2009) that also accommodates uncertain genotype calls.

2.2 Bayes Optimal Approaches

Lin et al. (2006) show that Bayesian ranking/percentiling procedures can outperform those computed by ranking Z-scores or MLEs. We define two versions, one that is "general purpose" and one targeted at identifying a specific number of SNPs. The Bayesian approach requires a full model with a prior distribution for a parameter of interest and a sampling distribution, conditional on the parameter. In applications, especially with a large K, empirical Bayes approaches are very attractive (see Schwender and Ickstadt (2008)).

We assume that θ is the true magnitude of the genotype/phenotype association and construct a model so that it is also the conditional mean function for observed data. For clarity, we assume that other parameters in the sampling model are known or well-estimated and that the maximum likelihood estimate $\hat{\theta}$ of θ is sufficient for computing the posterior distribution. For ease of notation, we denote $\hat{\theta}_k$ by Y_k and assume a Gaussian sampling distribution (these assumptions are not necessary in general).

For the null SNPs we have,

$$\begin{aligned}
[\theta_k, \sigma_k^2] \quad & iid \quad H \quad (H \text{ also delivers the } \sigma_k^2) \\
[Y_k \mid \theta_k, \sigma_k^2] \quad & \sim \quad f_k(Y_k \mid \theta_k, \sigma_k^2) = N(\theta_k, \sigma_k^2), k = 1, \ldots, K \\
f_k(Y_k \mid H) \quad & = \quad \int f_k(Y_k \mid \theta_k, \sigma_k^2) dH(\theta_k, \sigma_k^2) \qquad (1) \\
H_k(t \mid Y_k) \quad & = \quad pr(\theta_k \leq t \mid Y_k) = \int_{-\infty}^{t} \int_{0}^{\infty} f_k(Y_k \mid \theta_k, \sigma_k^2) dH(\theta_k, \sigma_k^2) / f_k(Y_k \mid H)
\end{aligned}$$

We will use in the sequel that,

$$E(H_k(t \mid Y_k)) = H(t), \text{ the prior distribution} \qquad (2)$$

In a likelihood-based analysis, inputs $Y_k (= \hat{\theta}_k)$ and σ^2 are the MLE and its associated variance (inverse information). However, in many contexts a score test (see the appendix) is used. As we show in section 2.2.4, a parameter estimate and its variance can be extracted from a score test.

2.2.1 Representation of ranks

For Bayes optimal ranks use the following representation for the K null SNPs in the context of ranking a total of $(K + 1)$ SNPs. The percentile for the spiked SNP is computed along with those for the null SNPs. For $k = 0, 1, \ldots, K$,

$$R_k(\boldsymbol{\theta}) = \sum_{\nu=0}^{K} I_{\{\theta_k \geq \theta_\nu\}}; P_k = R_k/(K + 2).$$

The smallest θ has rank 1.

2.2.2 $\hat{P}_k$: General purpose percentiles

First, compute the posterior expected ranks,

$$\bar{R}_k(\mathbf{Y}) = E[R_k(\boldsymbol{\theta}) \mid Y] = \sum_{\nu} pr[\theta_k \geq \theta_\nu \mid Y].$$

Then, rank the $\bar{R}_k$ to get $\hat{R}_k$ and $\hat{P}_k$.

2.2.3 $P^*(Y_0)$: Threshold-specific percentiles

Select $0 < \gamma < 1$, and let

$$\begin{aligned} H_K(t \mid \boldsymbol{\theta}) &= \frac{1}{K}\sum I_{\{\theta_k \leq t\}}, \text{ the edf of the } \theta_k \\ \bar{H}_K(t \mid \mathbf{Y}) &= E_H[H_K(t \mid \boldsymbol{\theta}) \mid \mathbf{Y}] = \frac{1}{K}\sum P_H[\theta_k \leq t \mid \mathbf{Y}] \end{aligned}$$

Then,

$$R_k^*(\gamma) = \text{rank } pr(\theta_k > \bar{H}_K^{-1}(\gamma \mid \mathbf{Y}) \mid \mathbf{Y}).$$

To optimally identify the top N SNPs, set $\gamma = 1 - N/K$.

2.2.4 Extracting $\hat{\theta}_k$ and σ_k^2 from a trend test

We need to map the trend test Z-score numerator and denominator into an estimate of the implicit θ (the log odds ratio for risk) and an estimate of its conditional variance. Using the model in section 2.1 and equation (20) in the appendix, we have for θ near 0 (omitting the subscript k),

$$\begin{aligned} E(Z \mid \theta, \mathbf{w}) = m(\theta, \mathbf{w}) &\doteq \theta n^{\frac{1}{2}} c(\mathbf{w})\{\hat{\pi}(1-\hat{\pi})\}^{\frac{1}{2}} \\ V(Z \mid \theta, \mathbf{w}) &\doteq 1 \end{aligned}$$

So,

$$\begin{aligned} \hat{\theta} &= Y = \frac{Z}{n^{\frac{1}{2}}\{\hat{\pi}(1-\hat{\pi})\}^{\frac{1}{2}} c(\mathbf{w})} \qquad (3) \\ \hat{\sigma}^2 &= V(Y \mid \theta) = \hat{V}(\hat{\theta} \mid \theta) = \frac{1}{n\{\hat{\pi}(1-\hat{\pi})\}c^2(\mathbf{w})}. \end{aligned}$$

We also need the relation,

$$E(Z \mid \theta, \sigma) = m(\theta, \sigma) = \frac{\theta}{\sigma} \; .$$

Note that σ^2 is inversely proportional to $nc(\mathbf{t})$ which depends on the MAF (see the appendix). Therefore, for a given sample size (n), the power of a test or the performance of a ranking procedure depends on the MAF and on the fraction of cases in the sample (π).

2.3 Taking Advantage of Large K

The ranking method specific, cumulative distribution of the percentile for the spiked SNP provides a complete description of performance. A fully simulation-based approach to estimating this cdf would proceed by generating data from the assumed scenario for all $(K+1)$ SNPs, computing the percentiles and saving the percentile for the spiked SNP. After a sufficient number of simulation replications, the empirical distribution function (edf) of these percentiles can be used as the estimate and edfs from competing methods can be compared. Specifically, using $P_K(Y_0)$ to denote the percentile for the spiked SNP as a function of all data, $(Y_0, Y_1, \dots, Y_K)$, we need to compute $pr(P_K(Y_0) > s), s \in [0,1]$ and compare these tail functions among ranking methods (Z-score based, Bayes) for a set of scenarios. We use the tail function because we are looking for the stochastically large distributions.

Our goal is to compute $pr(P_K(Y_0) > s)$, but consider computing $pr(P_\infty(Y_0) > s)$. If K is sufficiently large, evaluation of the latter provides a good approximation to the former. This is easily seen in the case wherein all SNPs, including the spiked SNP, come from the same distribution. In this case, $pr(P_\infty(Y_0) > s) = 1 - s$ and for finite K, the distribution of the percentile for the spiked SNP is discrete uniform on the points $(\frac{1}{K+2}, \dots, \frac{K+1}{K+2})$ and so

$$1 - \frac{[s(K+2)]+1}{K+1} \le pr(P_K(Y_0 > s) \le 1 - \frac{[s(K+2)]}{K+1}. \qquad (4)$$

The foregoing holds for a fixed s. For $s = s_K$ such that $(1-s_K)K \approx N$, then both limits in equation (4) go to 0 and in this situation the $K = \infty$ representation provides a lower bound to performance in identifying a the small number of SNPs in a large K context.

In the following sections we approximate performance for large K by developing mathematical results for $pr(P_K(Y_0) > 2)$.

2.4 Distributions for the Null SNPs and the Spiked SNP

For the null SNPs we assume that the working and actual priors both equal to H. Therefore, using standard Bayesian computations, the pre-posterior expectation of the posterior distribution is the prior H and it can be used as the posterior ensemble posterior distribution for the null SNPs. It is the $K = \infty$ approximation to the finite K distribution.

To see that this produces a good approximation for all but the most extreme percentiles, consider simulating K "null" SNPs (these are not necessarily unassociated with phenotype, but have generally low association) and a small number (we focus on 1) highly associated, "spiked" SNPs. If K is very large, for each simulation replication the empirical distribution function (edf) of the K simulated values used for ranking will be very close to the theoretical (the generating) distribution because this edf is the average of iid realizations with expectation the distribution of interest. So, for a fixed percentile, the average converges almost surely and uniformly to the distribution of interest and for K large, the edf is, for all practical purposes, equal to the induced distribution. For example, in the Z-score approach, if all K SNPs are truly null, then the edf of the Z-scores is essentially $N(0, 1)$ for each replication; more generally the edf is essentially equal to the distribution induced by the sampling model.

The ensemble posterior for the null SNPs is

$$\bar{H}(t \mid \mathbf{Y}) = \frac{1}{K}\sum_{k} H_k(t \mid Y_k). \tag{5}$$

The summands in (5) are iid bounded random variables with mean H (see equation (2)). So, the average converges almost surely and uniformly to H. For very large K, $\bar{H}(t \mid \mathbf{Y}) \approx H(t)$. Note that if for θ, H is a point mass at 0, then so will be $\bar{H}$.

The foregoing assumes that H is known, but if K is large, an empirical Bayes or Bayes empirical Bayes approach produces the same result so long as the true H is in the allowable space of estimated priors. For empirical Bayes, it is sufficient that the parametric family includes H or that a (smoothed) non-parametric approach is used. For Bayes empirical Bayes, it is sufficient that H is in the support of the hyper-prior. In this case, for all but the extreme percentiles, the ensemble posterior will be virtually identical to H. For example, if H is Gaussian and the assumed prior is also Gaussian with (μ, τ^2) estimated from the data, the result holds as it does for the (smoothed) non-parametric maximum likelihood approach.

By taking advantage of this near-constancy of the edf, the stochastic positioning of the spiked SNP among the "null" SNPs can be related either analytically or with minimal computation to this distribution, thereby efficiently computing the edf for its percentile. Evaluation proceeds by one-time generation of K realizations for the null SNPs and then repeated generation of realizations for the spiked SNP(s). Ranks and percentiles are computed by comparing these values to the relevant edf induced by the null SNPs. In some cases, for example when the null SNPs have mean 0 and Z-scores are used, the induced distribution is $N(0, 1)$ and no simulation of the null SNPs is needed. Similarly, the spiked SNP may have a mathematically tractable distribution and a full mathematical treatment is possible. In any case, to approximate performance for a large K, at least for percentiles that are not too extreme a one-time simulation (generally generating realizations for substantially more than K SNPs) is needed for the null SNPs to produce the comparison distribution and repeated simulation for the spiked. However, see section 5 for a warning about situations wherein there is little to no saving in a simulation. Taking advantage of this near-constancy of the relevant edf for the K SNPs produces considerable efficiency gains and insights for the Z-score approach, the Bayesian approach and other candidate approaches to ranking.

2.5 The Spiked SNP

For the spiked SNP $(k = 0)$, we denote a generic prior by G rather than H. Further, we use A for the actual prior and W for the working (assumed) prior. As for the null SNPs, we assume a Gaussian sampling model and denote the sampling variance by σ_a^2 (it can be a random variable),

$$[Y_0 \mid \theta_0, \sigma_a^2] \quad \sim \quad f(y_0 \mid \theta_0, \sigma_a^2) = N(\theta_0, \sigma_a^2) \tag{6}$$

with actual prior A and marginal distribution $f_A(y_0)$. That is,

$$\begin{aligned} [\theta, \sigma_a^2] &\quad iid \quad A \\ f_A(y_0) &= \int f(y_0 \mid \theta_0, \sigma_a^2) dA(\theta_0, \sigma_a^2). \end{aligned}$$

The working prior W can be different from the actual prior A, producing,

$$\begin{aligned} [\theta_0, \sigma_a^2] &\quad iid \quad W \\ f_W(y_0) &= \int f(y_0 \mid \theta_0, \sigma_a^2) dW(\theta_0, \sigma_a^2). \end{aligned}$$

The Actual and the Working priors (A and W) may be different and one or both may equal H. We can set $dW = d\theta_0$ for a "frequentist" analysis. Setting A to a point mass at a single θ-value, allows assessment of the frequentist performance of the Bayesian analysis that assumes $\theta \sim W$.

Conditioning on Y_0 generates the working posterior distribution,

$$\theta_0 \mid Y_0 \ \sim \ G_W(t \mid Y_0)$$

with pre-posterior expectation, taken with respect to the actual prior

$$\bar{G}_W(t \mid A) \ = \ E(G_W(t \mid Y_0)). \tag{7}$$

If $W = A$, G_W in equation (7) equals A and if $A = H$ it equals H. For a frequentist (θ-specific) analysis of a Bayesian working model, let A be degenerate at some θ_0.

2.5.1 "Frequentist" Bayes computations

We present in the context of the spiked SNP, but similar relations hold for the null SNPs. A "frequentist" (likelihood-based) analysis results from use of a flat, improper working prior, $dW(\theta) = d\theta_0$. The posterior is the normalized likelihood (the likelihood is integrable for the Gaussian sampling distribution). That is $[\theta_0 \mid Y_0, G_W, \sigma_w^2] \sim N(Y_0, \sigma_w^2)$. The marginal posterior for θ_0 is produced by mixing over the distribution of σ_w^2 and the pre-posterior distribution by integrating with respect to the marginal distribution of Y_0.

3 PRE-POSTERIOR DISTRIBUTIONS

We are interested in comparing performance of $\ddot{P}_k$ (section 2.1), $\hat{P}_k$ (section 2.2.2) and $P_k^*(\gamma)$ (section 2.2.3). For each, we want the pre-posterior tail distribution and to this end also need the pre-posterior distribution of Z-scores and of Bayesian posterior distributions computed from a working model.

3.1 Pre-posterior Distribution of $\hat{P}(Y)$

Here we want the posterior percentile of the spiked SNP relative to all others with the percentile defined as the expected number of competing SNPs that the spiked SNP beats. With θ_k denoting a typical null SNP, and using $P(Y_0)$ for $P_\infty(Y_0)$, we have,

$$\hat{P}(Y_0) \quad = \quad pr(\theta_0 > \theta_k \mid Y_0). \tag{8}$$

With $G_W(\theta_0 \mid Y_0)$ the working posterior for the spiked SNP, consider a randomly drawn θ_k from H. Assume that H has a density, then:

$$pr(\theta_0 > \theta_k \mid Y_0) \quad = \quad 1 - \int G_W(\xi \mid Y_0) h(\xi) d\xi = \int H(\xi) g_W(\xi \mid Y_0) d\xi. \tag{9}$$

For each Y_0, this is a value $\in [0, 1]$ and we want its distribution. In general, it won't have a closed form, but we can simulate Y_0 from $f_A(y_0)$ and get an empirical estimate without simulating the K null SNPs.

Theorem 1:

$$pr(\hat{P}(Y_0) > s) \quad = \quad 1 - \bar{G}_W(H^{-1}(s) \mid A). \tag{10}$$

Note that when $W = A = H$ (the spiked SNP is also null), equation 10 equals $H(H^{-1}(s)) = s$, the $U(0, 1)$ distribution.

Proof: Use H to compute the probability integral transform for the random variable that follows $G_W(\cdot \mid Y_0)$. Using this, the conditional distribution (given Y_0) of the percentile is $G_W(H^{-1}(s) \mid Y_0)$. Take the pre-posterior expectation to obtain (10). $\boxed{\text{QED}}$

Equation (10) can be evaluated numerically or in some situations (e.g., both $\bar{G}$ and H are Gaussian) in "closed" form. In any case, H and $\bar{G}_W(\cdot \mid A)$ need to be computed only once, a considerable computational savings. Since H may be time-consuming to invert, for graphs it will be convenient to let $s = H(t)$ and plot $(H(t), G(t))$ with t ranging over a suitably broad interval and evaluated on a sufficiently fine mesh.

The expected percentile is,

$$E(\hat{P}(Y_0) \mid A) \;=\; 1 - \int \bar{G}_W(\xi \mid A) h(\xi) d\xi \tag{11}$$

and if $W = A = H$ (spiked is also null)

$$= \; 1 - \int H(\xi) h(\xi) d\xi = \frac{1}{2}.$$

3.1.1 Gaussian W, A and H

For the Gaussian sampling distribution with a variance that does not depend on k ($\sigma_k^2 \equiv \sigma_h^2$), we can obtain closed form when all of (W, A, H) are also Gaussian, but not necessarily the same Gaussian. In this case, all marginal and conditional distributions are Gaussian. If $\theta \sim N(\mu, \tau^2)$ and $[Y \mid \theta] \sim N(\theta, \sigma^2)$, then $[\theta \mid Y] \sim N(\mu + (1-B)(Y-\mu), (1-B)\sigma^2)$ with $B = \sigma^2/(\sigma^2 + \tau^2)$ and the marginal distribution of Y is $N(\mu, \sigma^2 + \tau^2) = N(\mu, \sigma^2/B) = N(\mu, \tau^2/(1-B)$.

We use subscripts w, a, h to indicate each distribution and find,

$$\begin{aligned} \bar{G}_W(\cdot \mid A) &= N(\mu_*, \tau_*^2) \\ \mu_* &= B_w \mu_w + (1 - B_w)\mu_a \\ \tau_*^2 &= (1 - B_w)^2 \left[\frac{\sigma_w^2}{1 - B_w} + \frac{\sigma_a^2}{B_a}\right] \end{aligned} \tag{12}$$

Generally, $\sigma_w^2 = \sigma_a^2$, producing

$$\tau_*^2 = \sigma_a^2 (1 - B_w)^2 \left[\frac{1}{1 - B_w} + \frac{1}{B_a}\right]$$

Now, $H^{-1}(s) = \tau_h \Phi^{-1}(s) + \mu_h$ and so from (10),

$$pr(\hat{P}(Y) \leq s) \quad = \quad \bar{G}_W(H^{-1}(s) \mid A) = \Phi\left(\frac{\tau_h \Phi^{-1}(s) + (\mu_h - \mu_*)}{\tau_*}\right) \tag{13}$$

For a frequentist evaluation of the (W, H) combination, set $\tau_a = 0$ $(B_a = 1)$ (a point mass at μ_a).

3.2 Pre-posterior Distribution of $P^*(Y_0)$

Let $t_\gamma = H^{-1}(\gamma)$. To compute $P^*(Y_0)$, we need to find $pr(\theta_0 > t_\gamma \mid Y_0)$ (for the spiked SNP) and do the same for all of the null SNPs. Then, rank the probability for the spiked SNP relative to all the null SNPs. Thus, we need to compare $G_W(t_\gamma \mid Y_0) = G_W(H^{-1}(\gamma) \mid Y_0)$ to the collection of the K posteriors for the null SNPs. Equivalently, we need to count as follows for finite K (omitting dependence on W),

$$P^*(Y_0) \quad = \quad \frac{\#\{H(t_\gamma \mid Y_k) > G_W(t_\gamma \mid Y_0)\}}{K} \tag{14}$$

where the $H(t_\gamma \mid Y_k)$ are iid according to the distribution induced by $F_H(y)$. To see that (14) is the correct accounting, note that we want to rank the right-hand tail probabilities and when $H > G_W$, the right-hand tail is smaller.

We need $pr(H(t \mid Y_k) > u) = S(u \mid t)$. Then, $S(H(t \mid Y_k) \mid t) \sim U(0,1)$ and so,

$$\begin{aligned} pr_A\{H(t \mid Y_k) > G_W(t \mid Y_0)\} \quad &= \quad pr_A\{S(H(t \mid Y_k) \mid t) \leq S(G_W(t \mid Y_0) \mid t)\} \\ &= \quad pr_A\{U(0,1) \leq S(G_W(t \mid Y_0) \mid t)\} \\ &= \quad S(G_W(t \mid Y_0) \mid t). \end{aligned}$$

So,

$$P^*(Y_0) \quad \overset{distn}{=} \quad S(G_W(t_\gamma \mid Y_0) \mid t_\gamma), \text{computed under } A. \tag{15}$$

Note that if $W = A = H$, then $S(G_H(t_\gamma \mid Y_0) \mid t_\gamma) \sim U(0,1)$ and so $P^*(Y_0) \sim U(0,1)$.

3.2.1 Gaussian W, A and H

We use the notation in section 3.1.1 and first find S. Under H (dropping the subscript h until the end) and letting $Z_u = \phi^{-1}(u)$,

$$\begin{aligned} [\theta_0 \mid Y_0] &\sim N(B\mu + (1-B)Y_0, (1-B)\sigma^2) \\ H(t \mid Y_0) &= \Phi\left(\frac{t - (B\mu + (1-B)Y_0)}{\sigma\sqrt{1-B}}\right) \end{aligned}$$

For this to be $> u$, we need

$$Y_0 < \frac{t - B\mu}{1-B} - \frac{\sigma Z_u}{\sqrt{1-B}}.$$

But,

$$Y_0 \sim N(\mu, \sigma^2/B).$$

So,

$$S(u \mid t) = \Phi\left(\left\{\frac{t-\mu_h}{1-B_h} - \frac{\sigma_h Z_u}{\sqrt{1-B_h}}\right\}\frac{\sqrt{B_h}}{\sigma_h}\right). \tag{16}$$

Use (12) and (16) to evaluate (15).

In this all-Gaussian situation, if $W = H$ and $\sigma_w^2 = \sigma_h^2$, then the (apparently) optimal $P^*(Y_0)$ percentiles (those based on the working model) don't depend on γ and are equal to the $\hat{P}(Y)$.

3.2.2 Evaluation

To compute (15) we need to find S and then sample Y_0 from $F_A(y)$. Getting $S(u \mid t)$ may require simulation for a target t_γ and over a grid of u-values, but it is a one-time computation for each scenario. The same holds for sampling Y_0 from F_A. For efficient evaluation of a variety of scenarios, importance sampling can be used.

3.3 The Z-score Approach

We first consider the situation wherein all "null" SNPs have $\theta \equiv 0$ and the more general case.

3.3.1 All competing SNPs are null

All Z-scores are N(0,1) and so for large K their edf is N(0,1) as is $\bar{H}$. In equations (12) and (16), $\mu_h = 0, \tau_h^2 = 1$. For the spiked SNP, set $\tau_a = 0$ (a point mass at μ_a) and compute the Z-score based on $Y_0 \mid \mu_a \sim N(\mu_a, \sigma_a^2)$. Here $\mu_a = \theta$ and σ_a^2 is the variance in (3). This Z-score has mean $\xi = \mu_a/\sigma_a$ and local variance 1 (for small μ_a).

3.3.2 "Null" SNPs come from a distribution

Assume that the θ for the null SNPs come from H^* (recall that the prior delivers both the θ_k and the σ_k^2). Conditional on (θ, σ), $Z \sim N(m = \frac{\theta}{\sigma}, 1)$ and H^* induces a distribution on m, call it $\tilde{H}$. So,

$$H(u) = \int \Phi(u - m) d\tilde{H}(m) \tag{17}$$

In general, H needs to be evaluated numerically, but a few special cases are available.

Case I: If $\sigma_k^2 \equiv \sigma^2$ and θ_k iid $N(\mu, \tau^2)$ then

$$H(u) = N\left(\frac{\mu}{\sigma}, \frac{\sigma^2 + \tau^2}{\sigma^2}\right) \tag{18}$$

Case II: If σ is inverse gamma, H will be a t-distribution.

4 PERFORMANCE COMPARISONS

We provide performance evaluation of the Z-score approach based on $\ddot{P}$ and $\ddot{P}_k$ to detecting a spiked SNP, comparisons of fully Bayesian ranking via $(\hat{P}, \hat{P}_k)$ to the Z-score approach, and a comparison of "frequentist-Bayes" based on a flat prior and $(\hat{P}, \hat{P}_k)$ to $(\ddot{P}, \ddot{P}_k)$. These provide a proof of concept and a basis for additional studies that include $P_k^*(\gamma)$ and investigate performance of candidate approaches in with data-based priors.

4.1 Performance of Z-score-based Percentiles

We compute the performance for Z-score based percentiles when $\theta_k \equiv 0$ for various values of θ. The sampling variance is constant and the same value for all $(K + 1)$ SNPs. Therefore, for the null SNPs, $\bar{H} = N(0, 1)$ and the pre-posterior sampling distribution for the spiked SNP is $N(\theta/\sigma, 1)$ or equivalently $N(\theta, \sigma^2)$. Table 1 contains scenarios for which the tail functions (from $s = 0.80$ to 1.00) are plotted in Figure 1. The (θ, σ) pairs are produced by extracting estimates for θ, the log(Odds Ratio) for the genotype-phenotype association, from trend test Z-scores computed using genomic data in the HapMap project (International HapMap Consortium (2007), http://www.hapmap.org/).

These curves are computed from the foregoing mathematical relations. They are virtually identical to those produced by extensive simulation except for the extreme right tail, and so afford rapid evaluation of a wide variety of scenarios.

4.2 Comparisons for Gaussian θ with a Common Sampling Variance

For these comparisons, generically $\theta \sim N(\mu, \tau^2)$, $[Y \mid \theta] \sim N(\theta, \sigma^2)$ with the appropriate subscripts on (μ, τ^2, σ^2) to denote null SNPs or the spiked SNP. For the null SNPs, $\mu = 0$ and we study performance as a function of μ for the spiked SNP.

The Z-score approach: The generic Z-score is $Z = Y/\sigma^2$ and so the marginal distribution is $N(\mu/\sigma^2, 1/B)$ (recall that $\tau^2 = \sigma^2(1-B)/B$). This distribution is equivalent to using $N(\mu, \sigma^2 \frac{1}{B})$ in comparisons. For the null SNPs, $\mu = 0$.

Fully Bayes: Using the fully Bayes approach, the marginal posterior is the prior and so is $N(\mu, \tau^2) = N\left(\mu, \sigma^2 \frac{1-B}{B}\right)$.

Comparisons: The pre-posterior distributions for the Z-score and fully Bayesian approaches are Gaussian with the same mean. Therefore, comparison of $\ddot{P}_k$ with $\hat{P}_k$ depends on the variance. The variance for the Bayesian approach is $(1 - B)$ times that for the Z-score approach, and so the Bayesian approach will out-perform the Z approach in that for $\mu > 0$ the distribution of $\hat{P}$ will be stochastically larger than that of $\ddot{P}$. This is no surprise, since $\hat{P}$ takes advantage of the, correct, Bayesian structure.

4.2.1 Frequentist analysis

We investigate performance of the foregoing working analyses for fixed values of θ. We assume that $\theta_1 = \theta_2 = \theta_K = 0$ and study performance for a fixed θ for the spiked SNP.

Z-score: $Z = Y/\sigma^2$ and so the marginal distribution is $N(\theta/\sigma^2, 1)$ This distribution is equivalent to using $N(\theta, \sigma^2)$ in comparisons. For the null SNPs, $\theta = 0$.

Bayes: Using the fully Bayes approach as the working model, the pre-posterior for a fixed θ is,

$$N[\theta + B(\mu - \theta), \sigma^2(1-B)(2-B)].$$

For the null SNPs, $\theta = \mu = 0$ and the mean is 0.

Comparison: The variance for the Bayesian approach is $(1-B)(2-B)$ times that for the Z-score approach. However, the mean for the Bayesian approach is biased away from θ by $B(\mu-\theta)$. Therefore, the "frequentist performance" (performance for specific θ-values) will be better than the Z-score approach for θ near μ and less good for θ far from μ. Of course, "on average" over the prior for θ, Bayes performance dominates.

4.3 Z-score and Frequentist/Bayes with Constant Sampling Variance

Here, the working model for the Bayesian analysis is flat-prior, producing $[\theta \mid Y] = N(Y, \sigma^2)$. Conditioning on θ and mixing on Y produces $N(\theta, 2\sigma^2)$ and is inferior to the Z-score approach which produces $N(\theta, \sigma^2)$. Mixing on both Y and θ produces $N(\mu, 2\sigma^2 + \tau^2) = N[\mu, \frac{\sigma^2}{B}(1+B)]$, with a variance $(1+B)$ times that for the Z-score approach. Thus, frequentist/Bayes is never better than the Z-score approach, providing another example where use of an incorrect working model fails to achieve the Bayes advantage.

4.4 Z-score and Frequentist/Bayes with Non-constant Sampling Variance

The following comparisons are for $\theta = 0.223 (= \log(1.25))$, $\theta_k \equiv 0$, both when the sampling variance is constant and when it follows a discrete uniform trinomial distribution. The trinomial mass points are: $\sigma^2 = (0.224, 0.449, 0.897)$, $1/\sigma = (2.11, 1.49, 1.06)$ and are typical values from a subset of the HapMap data. The constant variance comparisons use $\sigma^2 = 0.449$.

Figures 2 and 3 display results, the second expanding the vertical scale by taking differences between the curves and $(1-s)$, the tail function for the U(0,1) distribution for $\theta = 0$. Inspection shows that there is no clear winner; the Z-score approach sometimes outperforms frequentist/Bayes and sometimes the reverse holds. Again, these curves and comparisons were produced without simulation and closely match curves produced by simulation.

5 TRADE-OFFS WHEN SIMULATIONS ARE NEEDED

The $K = \infty$ mathematical representations provide an excellent approximation to performance for large K so long as s is not too close to 1. However, simulation will still be needed if, for example H or other distributions are not mathematically evaluable and the number of draws from the distribution being estimated (R) needs to be specified. If one wants to have a good approximation for an extreme percentile, R may be sufficiently large that there is only a modest savings in computations and a standard simulation. As is the case in most statistical contexts, "there is no free lunch." However, as shown below, there may be a reduced-price lunch.

This need for a large R will is most apparent if one wants to estimate the far right-hand tail of the distribution of the percentile for the spiked SNP. Accuracy for this requires an accurate estimate of the right-hand tail of the relevant distribution for the null SNPs. To see this, consider estimating an extreme percentile of a U(0,1) distribution. This case is sufficient for the ranking context because the probability integral transform for the null SNPs can be used to produce the distribution. In a simulation of size R, the variance of the empirical s^{th} percentile is approximately $s(1-s)/R$ and the squared-coefficient of variation for the right-tail is $s/\{(1-s)R\} \approx 1/\{(1-s)R\}$ for s near 1. To control relative variation we require that this expression be no larger than V, producing, $R = 1/\{(1-s)V\}$. Now, consider the goal of identifying the top N SNPs so $(1-s) = N/(K+1)$ and $R = (K+1)/(NV)$. With $V = 10^{-4}$, a relative standard deviation of 0.01, and the goal of identifying the top $N = 100$ SNPs, $R = 100(K+1)$ and the computational equivalent of 100 simulations, each of size $(K+1)$ are needed. This is still a considerable savings over a full simulation, but a smaller V or smaller N will generate the need for an even larger initial simulation.

6 DISCUSSION

Simulation is a very effective tool for assessing the properties of statistical procedures. However, in many situations use of large-sample statistical properties can facilitate evaluations. In this context, we have shown how to use large-sample convergence of empirical distribution functions to speed up assessments of ranking procedures in some large sample size situations. This approach facilitates rapid evaluation and comparison of approaches for a broad array of scenarios and we provide a few examples as a proof of concept. More extensive comparisons and inclusion of the P^* approach will build on the current results. In structuring the assessments, we have developed integral representations of the stochastic distribution for the percentile rank of a single "spiked" SNP in competition with a large number of "null" SNPs. In some settings the integral representations will require some numerical evaluation, but these are one-time operations. Generalizations include developing representations for the maximum percentile of more than one spiked SNP.

The large K context is ideal for the use of flexible, empirical priors including the non-parametric maximum likelihood (another area in which Hans has worked) or a smooth version of it. Use of such priors should be very efficient and robust and our integral representations will facilitate rapid assessments.

Though an attractive approach, the approximations are not a panacea. For extreme percentiles they provide a lower bound for the tail of the distribution of the spiked SNP and this bound can be very close to zero. Furthermore, as shown in section 5, if simulations are needed to evaluate components of the mathematical expressions, they may not need to be very large.

Hans for your many contributions to statistics and to science, to your educational and mentoring success, and to your continued contributions we close with "dank u zeer."

Acknowledgements

Support provided by grants R01 DK061662, from the U. S. NIH, National Institute of Diabetes, Digestive and Kidney Diseases. and R01 HL090577 from the U. S. NIH, National Heart, Lung, and Blood Institute.

No conflict of interest: The authors have declared no conflict of interest.

7 APPENDIX: TREND TEST Z-SCORES

We outline the score test approach from Ruczinski et al. (2009) that accommodates uncertain ("fuzzy") genotype calls. For a specific SNP (index suppressed) from individual $i \in \{1, \dots, n\}$, we use the following notation:

X_i The binary disease status indicator for the i^{th} person.

g_i The true genotype, $g \in \{0, 1, 2\}$, assuming bi-allelic SNPs.

G_i The random variable taking on values g_i.

d_g The score assigned to genotype g. The additive genetic model uses $d_g = g, g \in \{0, 1, 2\}$.

π_g The risk relation: $\pi_g = pr(X_i = 1 \mid g) = H(\mu + \theta d_{g_i})$.

w_i The working probabilities: $w_i = pr(g_i = 1)$, $\mathbf{w}$ is the vector of these. If G is known, $\mathbf{w} = \mathbf{g}$.

To test the hypothesis of no genotype/phenotype association, H_0: $\pi_0 = \pi_1 = \pi_2 = \pi$ or equivalently, $H_0 : \theta = 0$, use the statistic:

$$Z(\mathbf{w}) \quad = \quad \frac{\sum_i (w_{i1} + 2w_{i2})(X_i - \hat{\pi})}{\sqrt{n \times \mathsf{Var}(\mathbf{w_1} + 2\mathbf{w_2}) \times \hat{\pi}(1-\hat{\pi})}} \tag{19}$$

with $\hat{\pi} = \bar{X} = \frac{1}{n}\sum_i X_i$ and $\mathsf{Var}(\mathbf{w_1} + 2\mathbf{w_2}) = \mathsf{Var}(\mathbf{w}_1) + 4 \times \mathsf{Var}(\mathbf{w}_2) + 4 \times \mathsf{Cov}(\mathbf{w}_1, \mathbf{w}_2)$.

Note that under H_0, $Z(\mathbf{w}) \sim N(0,1)$, the null distribution is correct irrespective of the working probabilities $\mathbf{w}$. Using a local alternative, we obtain for $\theta = \theta_0 \neq 0$:

$$Z \sim N(m(\theta_0, \mathbf{w}, \mathbf{g}), 1) \tag{20}$$

with

$$\begin{aligned} m(\theta_0, \mathbf{w}, \mathbf{g}) &= \theta_0 \times \sqrt{n} \times \sqrt{\hat{\pi}(1-\hat{\pi})} \times \mathsf{Corr}\left(\{\mathbf{w}_1 + 2\mathbf{w}_2\}, \{\mathbf{g}_1 + 2\mathbf{g}_2\}\right) \times \sqrt{\mathsf{Var}(\mathbf{g}_1 + 2\mathbf{g_2})} \\ &= \theta_0 \times \sqrt{n} \times \sqrt{\hat{\pi}(1-\hat{\pi})} \times c(\mathbf{w}) \end{aligned}$$

Squaring the Z-statistic produces a one degree of freedom, chi-square statistic with non-centrality $\lambda = m^2(\theta_0, \mathbf{w}, \mathbf{t})/2$.

References

Carlin, B. P. and Louis, T. A. (2009). *Bayesian Methods for Data Analysis, 3rd edition*. Chapman and Hall/CRC Press, Boca Raton, FL, 3^{nd} edition.

Goeman, J. J., van de Geer, S. A., and van Houwelingen, H. C. (2006). Testing against a high dimensional alternative. *Journal of the Royal Statistical Society: Series B* **68,** 477–493.

Heidema, A. G., Feskens, E. J. M., Doevendans, P. A. F. M., Ruven, H. J. T., van Houwelingen, H. C., Mariman, E. C. M., and Boer, J. M. A. (2007). Analysis of multiple snps in genetic association studies: comparison of three multi-locus methods to prioritize and select snps. *Genet Epidemiol* **31,** 910–921.

Houwing-Duistermaat, J. J., Uh, H. W., and van Houwelingen, H. C. (2007). A new score statistic to test for association given linkage in affected sibling pair-control designs. *BMC Proc* **1 Suppl 1,** S39.

International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million snps. *Nature* **449,** 851–861.

Lebrec, J. J. P., Putter, H., Houwing-Duistermaat, J. J., and van Houwelingen, H. C. (2008). Influence of genotyping error in linkage mapping for complex traits–an analytic study. *BMC Genet* **9,** 57.

Lin, R., Louis, T. A., Paddock, S. M., and Ridgeway, G. (2006). Loss function based ranking in two-stage, hierarchical models. *Bayesian Analysis* **1,** 915–946.

Manolio, T. A., Brooks, L. D., and Collins, F. S. (2008). A hapmap harvest of insights into the genetics of common disease. *J Clin Invest* **118,** 1590–1605.

Ruczinski, I., Li, Q., Carvalho, B., Fallin, M., Irizarry, R. A., and Louis, T. A. (2009). Association tests that accommodate genotyping errors. *in review* .

Schwender, H. and Ickstadt, K. (2008). Empirical bayes analysis of single nucleotide polymorphisms. *BMC Bioinformatics* **9,** 144.

Uh, H.-W., Mertens, B. J., van der Wijk, H. J., Putter, H., van Houwelingen, H. C., and Houwing-Duistermaat, J. J. (2007). Model selection based on logistic regression in a highly correlated candidate gene region. *BMC Proc* **1 Suppl 1,** S114.

Wakefield, J. (2007). A bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am J Hum Genet* **81,** 208–227.

Wakefield, J. (2008a). Bayes factors for genome-wide association studies: comparison with p-values. *Genet Epidemiol* **33,** 79–86.

Wakefield, J. (2008b). Reporting and interpretation in genome-wide association studies. *Int J Epidemiol* **37,** 641–653.

Westfall, P. H., Johnson, W. O., and Utts, J. M. (1997). A bayesian perspective on the bonferroni adjustment. *Biometrika* **84,** 419–427.

Scenario	OR	θ	$1/\sigma$	mean	color
Z100	1.00	0	5.44	0	black
Z110	1.10	0.095	4.28	0.41	red
Z125	1.25	0.223	2.11	0.47	green
Z150	1.50	0.405	3.87	1.57	blue
Z175	1.75	0.560	2.32	1.30	cyan
Z200	2.00	0.693	4.15	2.88	purple

Table 1: Scenarios for the Z-score plots

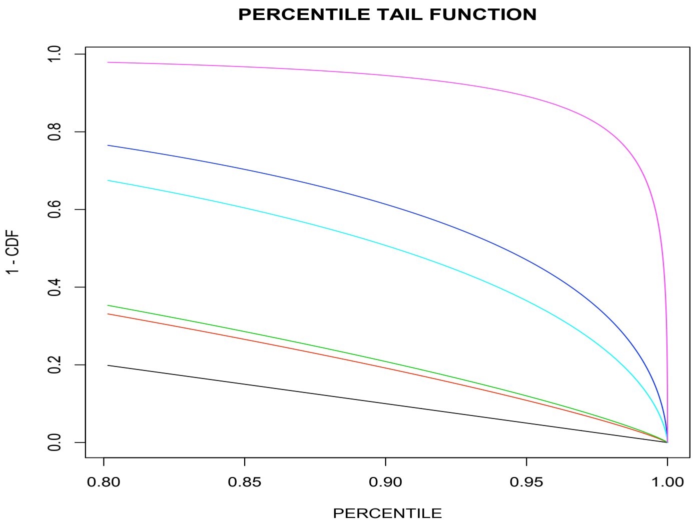

Figure 1: Tail distribution functions for $\ddot{P}$ for the scenarios in the Table 1. Color order is (black, red, green, blue, cyan, purple). The $E(\ddot{P})$, pr(percentile for the spiked SNP exceeds that for a null SNP), are: $(0.50, 0.61, 0.63, 0.82, 0.98)$.

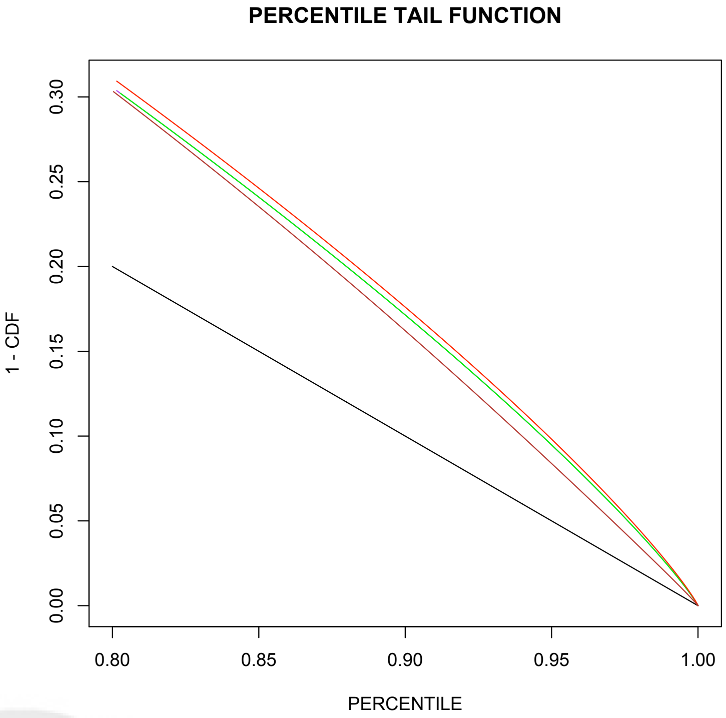

Figure 2: Tail distribution functions for $\ddot{P}$ (Z-prefix) and $\hat{P}$ (FB prefix) for OR $= 1.25$. The scenarios and color coding are black:U(0,1) (for comparison), purple:Znm, green:FBnm, red:Zmx, brown:FBmx. The pr(percentile for the spiked SNP exceeds that for a null SNP), are: $(0.50, 0.59, 0.59, 0.60, 0.59)$.

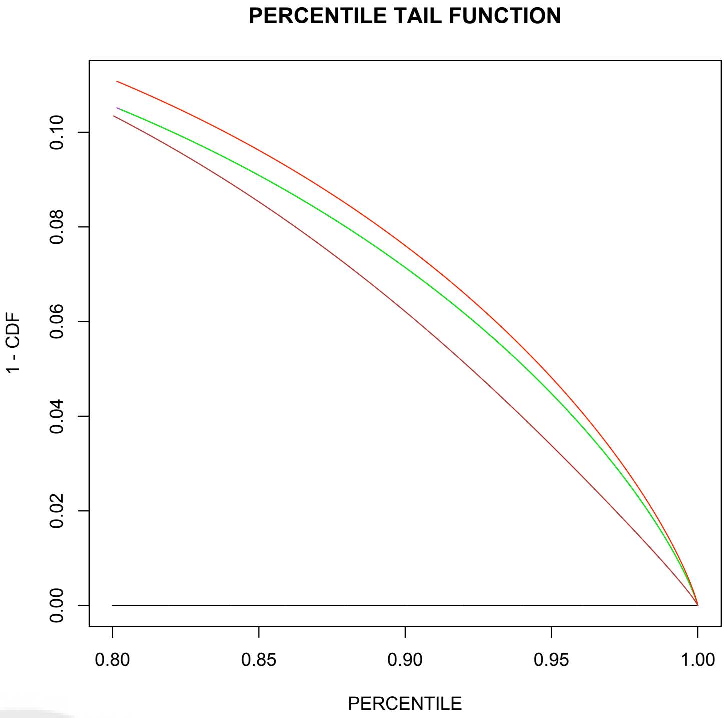

Figure 3: Difference between tail distribution functions for $\hat{P}$ and the (p, 1-p) line, for OR $=$ 1.25. The scenarios and color coding are black:U(0,1), purple:Znm, green:Bnm, red:Zmx, brown:Bmx. The $E(\hat{P})$, pr(percentile for the spiked SNP exceeds that for a null SNP), are:$(0.50, 0.59, 0.59, 0.60, 0.59)$.