Supplementary Material: Fast and Accurate Significance Approximation for Genome-wide Association Studies

Yu Zhang[1,*] and Jun S Liu[2]

[1] Department of Statistics, The Pennsylvania State University
422A Thomas Building, University Park, PA 16803
yuzhang@stat.psu.edu

[2] Department of Statistics, Harvard University
715 Science Center, 1 Oxford St, Cambridge, MA 02138
jliu@stat.harvard.edu

* Correspondence: yuzhang@stat.psu.edu

1. WEIGHT OF IMPORTANCE SAMPLING

We generate elevated test statistics of SNP i by replacing its genotype frequency contrast vector z_i by $z_i^* = z_i^2 + r^2$, where z_i is a (d-1)-dim random vector following a standard multivariate normal distribution. It is therefore sufficient to calculate the weight w_i as

$$w_i = h(z_i^*)/h(z_i)|J| = e^{-r^2/2}|J|$$

where $h(z)$ denotes the probability density function of a (d-1)-dim standard normal distribution, and $|J|$ denotes the Jacobian determinant of converting z_i to z_i^*. Let $B = diag(1+r^2/z_i^2)$ denote a (d-1)-dim square matrix, $u = (z_{ij}r^2/z_i^2)_{j=1,\cdots,d-1}^T$ denote a (d-1)-dim column vector, and $v = -u$. It can be shown that $J = (1+r^2/z_i^2)^{-1/2}(B+uv^T)$. By Sylvester's determinant theorem, we have $|J| = (1+r^2/z_i^2)^{(d-1)/2-1}$, and hence the weight function can be expressed as

$$\begin{aligned} w_i &= e^{-r^2/2}(1+r^2/z_i^2)^{(d-1)/2-1} \\ &= f_{(d-1)}(z_i^2+r^2)/f_{(d-1)}(z_i^2) \end{aligned} \quad (1)$$

where $f_{(d-1)}(x)$ denotes the probability density function of a chi-square distribution with $(d-1)$ degrees of freedom.

2. VARIANCE OF IMPORTANCE SAMPLING

Given a clump centered at SNP i, we write the probability $P(R_i = 1)$ as

$$P(R_i = 1) = P(T_i \geq t)P(\cup_{j=i-k_i}^{i-1} T_j < T_i, \cup_{j=i+1}^{i+l_i} T_j \leq T_i | t_i \geq t) = pq$$

where p denotes the nominal p-value of SNP i at threshold t and q denotes the conditional probability of the test statistics of neighboring SNPs below T_i given T_i exceeds the threshold t. Our importance sampling method generates n independent realizations of SNP data in the clump and computes $P(R_i = 1)$ numerically by $\hat{P}(R_i = 1) = \frac{1}{n}\sum_{j=1}^{n} r_{ij} w_{ij}$. We can write the variance of $\hat{P}(R_i = 1)$ as

$$Var(\hat{P}(R_i = 1)) = \frac{1}{n} Var(R_i w_i) = \frac{1}{n}\Big(E(R_i^2 w_i^2) - (pq)^2\Big)$$

It therefore suffices to evaluate $E(R_i^2 w_i^2)$ respect to $(pq)^2$ of one-step importance sampling, and multiple sampling iterations will decrease the variance at the rate of $\sqrt{n}$. For notation simplicity, we skip the subscript i hereafter.

According to our method, we simulate a large chi-square random variable of k degrees of freedom by first simulating a random variable x of chi-square(k) and then adding a shifting parameter r to x, i.e., $x^* = r + x$. We require that $0 \leq r \leq t$, and let $x_0 = t - r$ denote the difference. We show in the following that, for any moderate or large threshold t (e.g., when $t \geq k$), we can choose a proper value of r such that the variance of one-step importance sampling is bounded within the scale of p. As a result, a few hundreds of importance sampling iterations will be sufficient to produce an accurate approximation of p-value at threshold t, with the standard error in a smaller magnitude.

Let $f_k(\cdot)$ denotes the density function of chi-square(k). The weight w of one-step importance sampling is calculated as

$$w = \frac{f_k(x^*)}{f_k(x)} = \frac{(r+x)^{k/2-1} e^{-(r+x)/2}}{x^{k/2-1} e^{-x/2}} = \frac{(r+x)^{k/2-1}}{x^{k/2-1}} e^{-r/2} \tag{2}$$

When $k > 2$, the weight function (2) decreases with respect to x. We therefore can obtain an upper bound of $E(R^2 w^2)$ as

$$E(R^2 w^2) = \int_0^\infty R^2 w^2 f_k(x) dx = \int_0^\infty R w f_k(r+x) dx$$

$$
\begin{aligned}
&\leq \max_{R=1}(w)\int_0^\infty Rf_k(r+x)dx \\
&= \frac{(r+x_0)^{k/2-1}}{x_0^{k/2-1}}e^{-r/2}\int_0^\infty Rf_k(x+r)dx \\
&= \frac{(r+x_0)^{k/2-1}}{x_0^{k/2-1}}e^{-r/2}pq \qquad (3)
\end{aligned}
$$

Furthermore, we can obtain a lower bound of p when $k > 2$ by

$$
\begin{aligned}
p = P(\chi_k^2 \geq r + x_0) &= \frac{1}{2^{k/2}\Gamma(k/2)}\int_{r+x_0}^\infty x^{k/2-1}e^{-x/2}dx \\
&\geq \frac{(r+x_0)^{k/2-1}}{2^{k/2}\Gamma(k/2)}\int_{r+x_0}^\infty e^{-x/2}dx \\
&= \frac{(r+x_0)^{k/2-1}}{2^{k/2-1}\Gamma(k/2)}e^{-(r+x_0)/2} \qquad (4)
\end{aligned}
$$

Since our interest is to evaluate the magnitude of $E(R^2w^2)$ with respect to $(pq)^2$, we can compute an upper bound of the ratio between the two quantities. Let $\gamma_{\max}(x_0)$ denote the upper bound of the ratio, we compute $\gamma_{\max}(x_0)$ by dividing (3) by the square of (4) and by q^2, i.e.,

$$\gamma_{\max}(x_0) = \frac{2^{k/2-1}\Gamma(k/2)}{x_0^{k/2-1}}e^{x_0/2}q^{-1}$$

This bound of the maximum ratio can be minimized at $x_0 = k - 2$. As a result, we may set the shifting parameter $r = t - k + 2$ for all thresholds $t \geq k - 2$, when $k > 2$, and we set $r = 0$ for $t < k - 2$. By Stirling's approximation for factorials ($n! \simeq \sqrt{2\pi n}(\frac{n}{e})^n$), we have the minimum bound of the maximum ratio at $x_0 = k - 2$ equal to

$$\gamma_{\max}(k-2) \simeq \sqrt{2\pi(k/2-1)}q^{-1} \qquad (5)$$

In summary, when the degrees of freedom $k > 2$, we can control the variance of one-step importance sampling with an upper bound of $(\sqrt{2\pi(k/2-1)} - q)p^2q$, and hence the standard deviation of one-step importance sampling in the scale of $k^{1/4}p\sqrt{q}$. Depending on the conservativeness q incurred by SNP LD, a few hundreds of importance sampling iterations may be sufficient to reduce the sampling errors to a smaller magnitude with respect to the approximated p-values. The major determining factor is q. Since q increases with respect

to the threshold t, the variance will decrease as t increases. For association tests with larger degrees of freedom, we need to further increase the number of sampling iterations at the rate of $\sqrt{k}$.

Now we consider the situation where $k < 2$. We first compute an upper bound of $E(R^2w^2)$ as

$$\begin{aligned} E(R^2w^2) &= \int_0^\infty Rw^2 f_k(x)dx = \frac{e^{-r/2}}{2^{k/2}\Gamma(k/2)} \int_0^\infty R\frac{(r+x)^{k-2}}{x^{k/2-1}} e^{-(r+x)/2}dx \\ &\leq \frac{(r+x_0)^{k-2}e^{-r}}{2^{k/2}\Gamma(k/2)} \int_0^\infty Rx^{1-k/2}e^{-x/2}dx \\ &\leq \frac{(r+x_0)^{k-2}e^{-r}}{2^{k-2}\Gamma(k/2)} \int_0^\infty Ry^{1-k/2}e^{-y}dy, \text{ where } y = \frac{x}{2} \\ &\leq \frac{(r+x_0)^{k-2}e^{-r}}{2^{k-2}\Gamma(k/2)}\Gamma(2-k/2) \end{aligned} \tag{6}$$

We then compute a lower bound of p using integration by parts

$$\begin{aligned} p &= P(\chi_k^2 \geq r+x_0) = \frac{1}{2^{k/2}\Gamma(k/2)} \int_{r+x_0}^\infty x^{k/2-1}e^{-x/2}dx \\ &= \frac{1}{2^{k/2}\Gamma(k/2)}(-2)\left[x^{k/2-1}e^{-x/2} \mid_{r+x_0}^\infty -(k/2-1)\int_{r+x_0}^\infty x^{k/2-2}e^{-x/2dx}\right] \\ &\geq \frac{2(r+x_0)^{k/2-1}e^{-(r+x_0)/2}}{2^{k/2}\Gamma(k/2)} + (k-2)p, \text{ when } r+x_0 \geq 1 \end{aligned}$$

and hence

$$p \geq \frac{(r+x_0)^{k/2-1}e^{-(r+x_0)/2}}{(3-k)2^{k/2-1}\Gamma(k/2)}, \text{ when } r+x_0 \geq 1 \tag{7}$$

By taking the ratio between (6) and the square of (7), divided by q^2, we obtain an upper bound of the maximum ratio between $E(R^2w^2)$ and $(pq)^2$, for $k < 2$, as

$$\gamma_{\max}(x_0) = e^{x_0}\frac{(3-k)^2}{2^{k/2-1}}\Gamma(k/2)\Gamma(2-k/2)q^{-2}$$

which is minimized at $x_0 = 0$ as

$$\gamma_{\max}(0) = \frac{(3-k)^2}{2^{k/2-1}}\Gamma(k/2)\Gamma(2-k/2)q^{-2}$$

In summary, when $k < 2$, we have the variance of one-step importance sampling bounded by $(\frac{(3-k)^2}{2^{k/2-1}}\Gamma(k/2)\Gamma(2-k/2) - q^2)p^2$. In association tests, $k = 1$ is the only value < 2, and

hence the number of sampling iterations needed solely depends on the conservativeness q, but not on the threshold t.

Finally, it is easily checked that, when $k = 2$, the variance of one-step importance sampling equals to $p^2q(1-q)$.

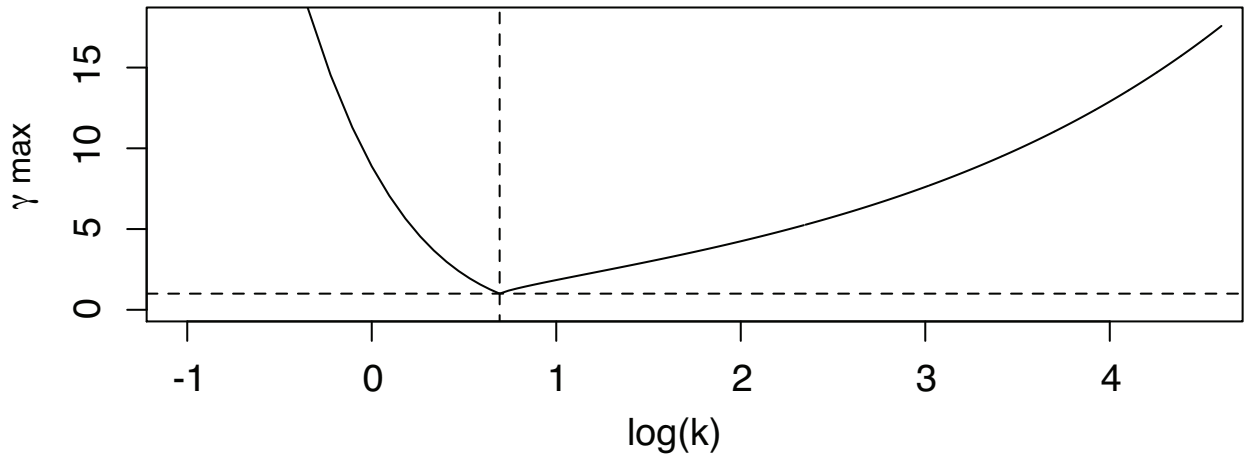

Figure 1: Maximum bounds γ_{max} of $E(R^2w^2)$ in the unit of p^2, assuming $q = 1$, for association tests with different degrees of freedom k.

3. EXAMPLES OF POISSON APPROXIMATION

We checked whether the number of significant clumps based on our definition really follows a Poisson distribution. We partitioned the WTCCC1 data (2000 cases and 3000 controls) into about 50 smaller datasets, with 10,000 consecutive SNPs in each dataset covering an average of 60Mb region in the human genome. We selected three thresholds T =12.5, 15.0, and 17.5, respectively, so that the mean numbers of significant clumps per dataset were 1, 3, 10, respectively. Figure 2 shows that the number of significant clumps follows closely to Poisson distributions. The number of significant SNPs, on the other hand, has a highly inflated variance. It is worthy to point out that, although the distributions of the two numbers (of clumps vs. of SNPs) differ significantly, the probability of both numbers equal to 0 is the same. That is, they have the same family-wise false positive rate. This validates our approach of approximating genome-wide significance using clumps.

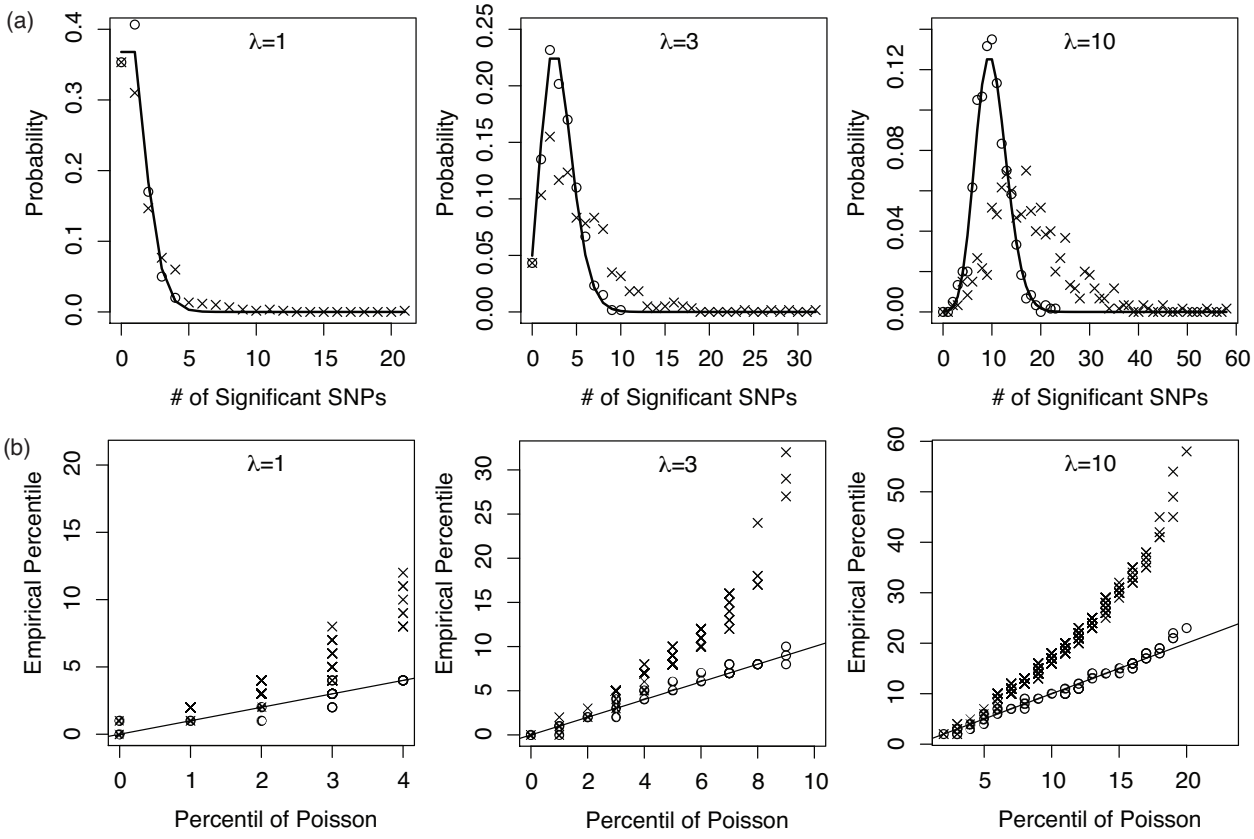

Figure 2: Distribution comparison of the number of significant association signals. The number of significant clumps (circles) and the number of significant SNPs (crosses) are calculated from the WTCCC1 data of 10,000 SNPs per dataset from 500 permuted datasets. The significance thresholds are chosen to control the expected number of false positive clumps (λ) shown on top of each plot. (a) Empirical distributions compared with Poisson(λ) distributions (solid line). (b) Corresponding quantile-quantile plots together with reference lines.

4. ENCODE REGIONS

ENCODE regions consist of 1% of the human genome selected to best represent the human genome diversities. We downloaded the ENCODE SNP data of CEU and YRI samples from the HapMap website and applied our method to approximate the genome-wide significance extrapolated from each ENCODE region. There are 10 ENCODE regions for which the SNP data are available, including both dbSNPs and novel SNPs detected by the ENCODE resequencing project. The ENCODE SNPs capture almost the complete set of common SNPs in 1% of the human genome. After filtering out non-polymorphic SNPs within CEU and YRI samples, the average SNP density of each ENCODE region is 2 SNPs per kb for CEU and 2.46 SNPs per kb for YRI, which roughly correspond to 6 million SNPs genome-wide.

Figure 3a shows the estimated deflation rates (effective number of independent SNPs divided by the total number of SNPs) at threshold 37.155 from the 10 ENCODE regions. The threshold corresponds to the Bonferroni adjusted p-value 0.05 for 6 million SNPs. Assuming 1000 cases and 1000 controls in a dataset, we observed considerably different deflation rates from the 10 ENCODE regions. In particular, *ENm010* is much less conservative than *ENm013* and *ENm014*, although they are all located on chromosome 7, indicating that the significance of associations is highly variable among local regions. Similar to the HapMap results, we again observed a consistent pattern of deflation rates across ENCODE regions between CEU and YRI samples, with correlation coefficient 0.925 (p-value 0.0001). We also observed in Figure 3b a significant correlation between deflation rates and recombination rates across ENCODE regions (0.90 in the CEU sample with p-value 0.0008 and 0.87 in the YRI sample with p-value 0.002), excluding *ENm010*.

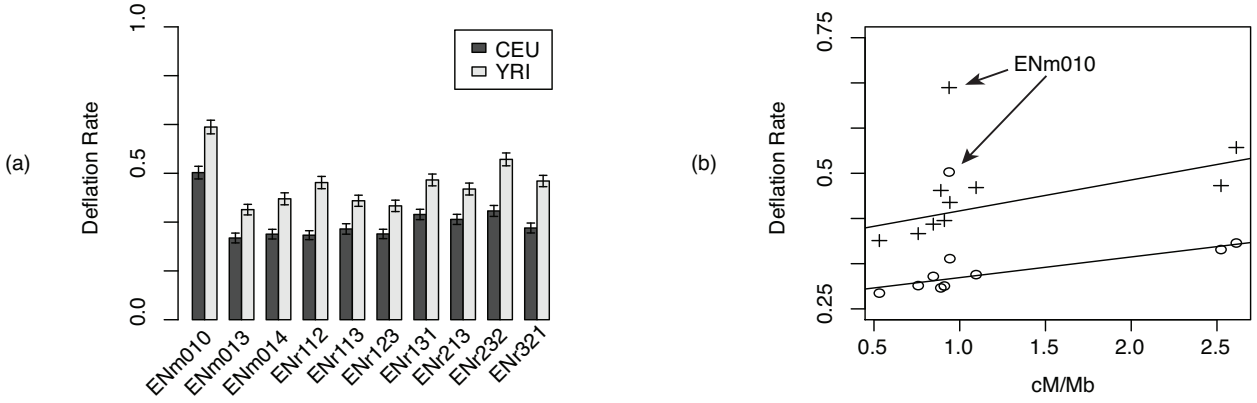

Figure 3: Deflation rates approximated from ENCODE regions, assuming 6 million SNPs genome-wide in 1000 cases and 1000 controls. (a) Bar-plot of the deflation rates estimated from each ENCODE region to the whole-genome in CEU and YRI populations, respectively, with standard errors. (b) Scatter plot of the deflation rates against the recombination rate of ENCODE regions in CEU (circle) and YRI (cross) populations, respectively.