

Supplementary material for the article
 ”Estimating the posterior probability that genomewide association findings
 are true or false”
 by József Bukszár, Joseph L. McClay and Edwin J.C.G. van den Oord

1 Computation of the likelihood function

In this section we summarize the method we use to compute the likelihood function

$$L(m_1, \Delta) = \frac{1}{\binom{m}{m_1}} \sum_H f_{H_1 \cdot \Delta}(t_1) \dots f_{H_m \cdot \Delta}(t_m) = \frac{1}{\binom{m}{m_1}} \left(\prod_{i=1}^m f_0(t_i) \right) \sum_{\{i_1, \dots, i_{m_1}\} \subseteq \{1, \dots, m\}} \frac{f_\Delta(t_{i_1})}{f_0(t_{i_1})} \dots \frac{f_\Delta(t_{i_{m_1}})}{f_0(t_{i_{m_1}})}, \quad (1)$$

where f_Δ (f_0) denotes the alternative (null) density function. Due to enormous number of terms in it, the sum in (1) cannot be evaluated directly. We therefore developed the method below that is based on recursive series. Define $S(n)$ as

$$S(n) = \sum_{\{i_1, \dots, i_n\} \subseteq \{1, \dots, m\}} a_{i_1} \dots a_{i_n} \quad (2)$$

for $n = 1, \dots, m$, and $S(0) = 1$, where $a_i = \frac{f_\Delta(t_i)}{f_0(t_i)}$ for $i = 1, \dots, m$. Then the likelihood function can be re-written as

$$L(m_1, \Delta) = \frac{1}{\binom{m}{m_1}} \left(\prod_{i=1}^m f_0(t_i) \right) S(m_1). \quad (3)$$

One can verify the following sieve-formula

$$S(n) = \frac{1}{n} \sum_{i=1}^n (-1)^{i+1} R(i) S(n-i), \quad (4)$$

where $R(i) = \sum_{j=1}^m a_j^i$. By this formula we can calculate $S(m_1)$ in the real likelihood function in a recursive way starting from $S(0) = 1$ and $S(1) = \sum_{j=1}^m a_j$. The large spectrum of values of a_i 's in combination with the recursive use of them will cause numerical problems when evaluating $S(n)$. Our computer implementation avoided these problems using a variety of techniques. A major technique involved is partitioning the set of a_i s. That is, the distribution of a_i 's is such that the vast majority of hypotheses have values with small range. For this set we can use the recursive formula. For the remaining hypotheses that have a very large range of a_i 's, we created bins of 10 hypotheses. Because there are only 10 hypotheses in each bin, we don't need the recursive formula for which a large range is problematic. Instead, we calculated $S(n)$ directly using (2). The $S(n)$'s of all bins were then combined to calculate the $S(n)$ for all hypotheses. The R codes are freely downloadable from the website <http://www.people.vcu.edu/~jbukzar>.

2 Comparison of the real and mixture model maximum likelihood estimators

As an alternative of the maximum likelihood function, in principal we can use the mixture model log-likelihood function

$$\ell_{\text{mix}}(m_1, \Delta) = \sum_{i=1}^m \log \{m_0 f_0(t_i) + m_1 f_\Delta(t_i)\}. \quad (5)$$

In the mixture model H_1, \dots, H_m are independent Bernoulli random variables with $\Pr(H_i = 0) = m_0/m$ and $\Pr(H_i = 1) = m_1/m$, whereas in our model $H = (H_1, \dots, H_m)$ is 0-1 vector that has *exactly* m_1 coordinates that are 1. As a result, in the mixture model the number of true alternatives is a random variable that follows binomial distribution $b(m_1/m; m)$, whereas in our model the number of true alternatives is a constant. The

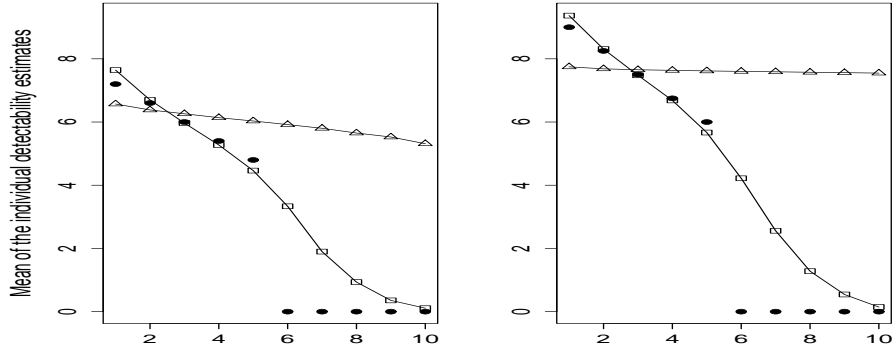


Figure 1: The mean of the real maximum likelihood (squares) and mixture model maximum likelihood (triangles) individual detectability estimates on 500 simulations are plotted. No stopping rule was used. The actual detectabilities are 7.2, 6.6, 6, 5.4, 4.8 in the left panel, and they are 9, 8.25, 7.5, 6.75, 6 in the right panel, plotted by dots. The number of hypotheses was 100,000 in each simulation.

rationale in our model is that in an experiment the number of true alternatives is truly a constant albeit unknown. The mathematical connection between the mixture model and "real" likelihood function can be given as

$$\frac{1}{m^m} L_{\text{mix}}(m_1, \Delta) = E(L(M_1, \Delta)) = \sum_{k=0}^m \binom{m}{k} (1 - m_1)^k m_1^{m-k} L(k, \Delta), \quad (6)$$

where $L_{\text{mix}}(m_1, \Delta)$ is the mixture model likelihood function, $L(m_1, \Delta)$ is the "real" likelihood function (1), and $M_1 \sim (1 - p_0; m)$ (see Appendix). Intuitively, M_1 being random impose additional uncertainty in the mixture model.

As we did in the article based on the real maximum likelihood estimator (ML), we obtain individual detectability estimates based on the mixture model maximum likelihood method (qML) by calculating first the average detectability estimates

$$\hat{\Delta}_k := \max_{\Delta} L_{\text{mix}}(k, \Delta). \quad (7)$$

and then calculating the individual detectabilities

$$\hat{\varepsilon}_k = k\hat{\Delta}_k - (k - 1)\hat{\Delta}_{k-1}, \quad (8)$$

recursively for $k = 1, 2, \dots$, where $\hat{\Delta}_0$ is defined 0. That is we use L_{mix} rather than L in (7) to calculate average detectability estimates.

In Figure 1 we study the "real" maximum likelihood (ML) and mixture model maximum likelihood (qML) individual detectability estimates without using any stopping rule on 500 simulations. In each simulation we generated five alternative test statistic values with detectabilities 7.2, 6.6, 6, 5.4, and 4.8 in the left panel, and 9, 8.25, 7.5, 6.75, 6 in the right panel, and the rest of the 100,000 test statistic value were generated according to the null distribution. The mean of the individual detectability estimates obtained by the ML (qML) method are plotted with squares (triangles), the actual detectabilities are plotted with dots on the figure. The actual detectabilities in the right panel are 25% higher than those in the left panel, which, in practice can be a result of increased sample size. Whereas the real ML individual detectability estimates become less biased as the actual detectabilities increase, the qML individual detectability estimates become almost equal with each other, thus, they become even more biased as the detectabilities increase. As one might expect, this phenomenon is not specific to this particular numerical example, but is general characteristic of the two estimators.

We obtained Figure 2 by repeating the experiment using the stopping rule this time (see method section). Since the estimators of the number of true alternative hypothesis becomes more accurate as the detectabilities increase, using the stopping rule will not change "the asymptotic behavior" of the maximum likelihood estimator. The same phenomenon as in Figure 1 can be observed in Figure2.

In Table 1 we examine the performance of the local FDR estimator when mixture model likelihood rather than the "real" likelihood was used. We used the same simulated data that was used to create Table 1 in the article. Both the STD and the estimated radius (conf) of zero-centered 90% confidence interval of MDs

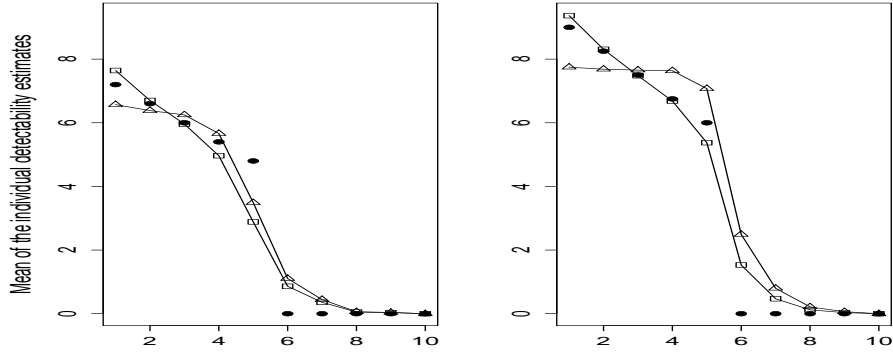


Figure 2: The mean of the real maximum likelihood (squares) and mixture model maximum likelihood (triangles) individual detectability estimates on 500 simulations are plotted. No stopping rule was used. The actual detectabilities are 7.2, 6.6, 6, 5.4, 4.8 in the left panel, and they are 9, 8.25, 7.5, 6.75, 6 in the right panel, plotted by dots. The number of hypotheses was 100,000 in each simulation.

is higher when the mixture model likelihood is used. The difference between STD and the confidence interval radius of the MDs of the two likelihood estimators is even bigger for higher range of positive detectabilities. The mean of MDs is significantly closer to zero when the real likelihood is used for higher range of positive detectabilities, and there is no noticeable difference between the two for lower range of positive detectabilities. In summary, the local FDR estimator is less biased and more accurate when the real likelihood is used than when the mixture model likelihood is used.

Table 1: Mean, standard deviation, and the estimated radius (conf) of zero-centered 90% confidence interval of the MDs (maximum differences) between the real and estimated local FDR curves for different range, number and distribution of detectabilities, when **mixture model ML method was used**. The number of markers was **100,000**

Range	#effects	Equid			Conc Endp		
		mean	STD	conf	mean	STD	conf
3.2-4.8	5	.200	.28	.639	.173	.27	.537
	10	.184	.13	.329	.143	.12	.241
	20	.199	.07	.289	.162	.07	.247
4.0-6.0	5	.068	.18	.267	.062	.18	.306
	10	.093	.11	.238	.113	.14	.293
	20	.109	.08	.216	.149	.10	.294
4.8-7.2	5	.088	.22	.406	.166	.24	.498
	10	.105	.16	.324	.278	.18	.517
	20	.121	.11	.26	.32	.13	.486

3 Conservative estimator

In this section, we give an estimator of the proportion of true null hypotheses $p_0 = m_0/m$. This p_0 estimator does not rely on the test statistic distribution under the alternative but capitalizes on the fact that in large scale genetic studies p_0 is close to 1.

We calculate a cut-off value c in such a way that the probability that a null marker has test statistic value higher than c is k/m , where k is a *fine tuning parameter*. If we denote the total number of markers whose test statistic value is higher than c as d , then this estimate of p_0 is

$$\hat{p}_0 = 1 - \frac{d - k}{m}.$$

Note that the expected number of null markers with test statistic value higher than cut-off c is km_0/m rather than k . This estimator can therefore be expected to be conservatively biased, hence we will call it Conservative estimator. However, because $p_0 = m_0/m$ is close to 1 we would expect the bias to be small.

Let us denote the effect sizes of alternative hypotheses as $\Delta_1, \dots, \Delta_{m_1}$. By taking the expectation we obtain

$$E(\hat{p}_0) = p_0 + \frac{1}{m} \sum_{i=1}^{m_1} \left(1 + k/m - \pi^{(i)}\right),$$

where $\pi^{(i)} = 1 - F_{\Delta_i}(F_0^{-1}(1 - k/m))$, and F_0 and F_{Δ} are the proper cumulative distribution functions under the null and alternative hypothesis, respectively. Note that the bias of \hat{p}_0 is positive and $E(\hat{p}_0) < 1$. This latter one holds because $k/m + F_{\Delta_i}(F_0^{-1}(1 - k/m)) < 1$ or equivalently $F_0^{-1}(1 - k/m) < F_{\Delta_i}^{-1}(1 - k/m)$.

4 Numerical Results

In all simulations, test statistic values of true null hypotheses will be drawn from central chi-squared distribution with d.f.1. Alternative test statistic values will be drawn from non-central chi-square distribution with 1 degree of freedom (d.f. 1) whose non-centrality parameter is the square of the detectability. By definition, detectability is the product of the effect size and the square root of sample size. For instance, this may be the approximating distribution for Pearson's statistic in allele-based case-control studies. Throughout this section the individual detectabilities rather than the effect sizes will be estimated. Note that in a real-life application detectabilities and effect sizes can readily be calculated from each other.

4.1 Distribution of median differences between the real and estimated local FDR curves

We redid tables 1-3 in the article using **median difference (MeD)** instead of maximum difference (MD) between the function that assigns the true ℓ FDR and the function that assigns the estimated ℓ FDR to test statistic values. In summary, the trends are the same, however, the absolute results (mean, standard deviation, and the estimated radius of zero-centered 90% confidence interval) look much better than when maximum difference (MD) was used.

The predominance of the positive mean MeDs in Table 2 indicates the upward (conservative) bias of the ℓ FDR estimator, i.e. it overestimates the ℓ FDR. In particular, the mean of the MeDs is higher for the lower range of positive detectabilities. Although in most cases the mean of the MeDs goes up slightly as the number of positive detectabilities gets larger, the mean of the MeDs is mainly dependent on the range of the positive detectabilities. The mean of the MeDs noticeably differs across the types of distributions of detectabilities when the range of the detectabilities is low (3.2 – 4.8). However, this difference becomes marginal for the higher range of detectabilities. Table 2 shows that the higher the number and the size of the positive detectabilities, the less the standard deviation of the MeDs. Moreover, the type of the distribution of the detectabilities has no substantial influence on the STD.

Table 2: Mean, standard deviation, and the estimated radius (conf) of zero-centered 90% confidence interval of the **MeDs** (median differences) between the real and estimated local FDR curves for different range, number and distribution of detectabilities. The number of markers was **100,000**

Range	#effects	Equid			Conc Endp		
		mean	STD	conf	mean	STD	conf
3.2-4.8	5	.049	.0922	.198	.034	.0747	.067
	10	.030	.0348	.048	.025	.0460	.035
	20	.032	.0109	.045	.027	.0093	.039
4.0-6.0	5	.005	.0166	.016	.004	.0085	.013
	10	.009	.0068	.018	.009	.0070	.018
	20	.012	.0055	.019	.014	.0059	.022
4.8-7.2	5	-.004	.0064	.013	-.003	.0078	.014
	10	-.001	.0048	.009	.003	.0069	.014
	20	0	.0034	.006	.006	.0050	.013

In Table 3 we increased the number of markers from 100,000 to 400,000, and kept all other conditions the same. Although the proportion of positive detectabilities in the total set of markers is much smaller now, the

estimator performed only slightly different for the same conditions in Table 2, indicated by the radius of the confidence interval. The exceptions are the lower range (3.2 – 4.8) and low number (5 or 10) of detectabilities

Table 3: Mean, standard deviation, and the estimated radius (conf) of zero-centered 90% confidence interval of the **MeDs** (median differences) between the real and estimated local FDR curves for different range, number and distribution of detectabilities. The number of markers was **400,000**.

Range	#effects	Equid			Conc Endp		
		mean	STD	conf	mean	STD	conf
3.2-4.8	5	.032	.0398	.108	.027	.0356	.100
	10	.036	.0435	.072	.024	.0257	.041
	20	.032	.0138	.049	.026	.0098	.037
4.0-6.0	5	.008	.0121	.018	.006	.0102	.016
	10	.011	.0071	.019	.01	.0069	.018
	20	.014	.0051	.02	.014	.0058	.021
4.8-7.2	5	-.001	.0054	.009	0	.0069	.011
	10	.001	.0041	.007	.005	.0059	.012
	20	.003	.0031	.007	.008	.0051	.014

Table 4: Mean, standard deviation, and the estimated radius (conf) of zero-centered 90% confidence interval of the **MeD's** (median differences) between the real and estimated local FDR curves for different correlation structures and number of detectabilities. The number of markers was **100,000**, moreover, 5 or 20 of them had real detectabilities equidistantly chosen from [4.0, 6.0] including the limits of the interval.

Block size	Block Corr.	5 dets			20 dets		
		mean	STD	conf	mean	STD	conf
0	0	.008	.0121	.018	.012	.0055	.019
5	.5	.012	.0414	.019	.012	.0071	.021
	.75	.020	.0650	.026	.013	.0092	.025
	.9	.022	.0716	.031	.013	.0114	.025
10	.5	.010	.0362	.019	.012	.0082	.022
	.75	.020	.0657	.026	.014	.0329	.027
	.9	.025	.0795	.043	.017	.0489	.029

In Table 4 we studied the performance of the ℓ FDR estimator in the context of substantial correlation or linkage disequilibrium between the markers. The mean MeD slightly changes across the different correlation structures, meaning that higher correlation results in a marginally higher bias. As one might expect, the higher the within-block correlation or the size of the block, the higher the STD. The changes in the mean and STD of the MeDs are also reflected in the greater radius of the confidence interval, although this change is not dramatic even in the extreme condition (within-block correlation 0.9, block size 10).

In summary, the accuracy of the ℓ FDR estimator is mainly dependent on the range and the number of the positive detectabilities. It only slightly depends on the total number of markers or on how the positive detectabilities are distributed, except for lower ranges of positive detectabilities, where the estimator is not precise anyway. The correlation structure has some but by no means dramatic influence on the performance of the ℓ FDR estimator.

5 Appendix

Here we derive (6).

$$E(L(M_1, \Delta)) = \sum_{k=0}^m \binom{m}{k} (m_1/m)^k (m_0/m)^{m-k} L(k, \Delta) = \frac{1}{m^m} \sum_{k=0}^m \binom{m}{k} (m_1)^k (m_0)^{m-k} L(k, \Delta) =$$

$$\begin{aligned}
& \frac{1}{m^m} \sum_{k=0}^m \binom{m}{k} (m_1)^k (m_0)^{m-k} \frac{1}{\binom{m}{k}} \left(\prod_{i=1}^m f_0(t_i) \right) \sum_{\{i_1, \dots, i_k\} \subseteq \{1, \dots, m\}} \frac{f_\Delta(t_{i_1})}{f_0(t_{i_1})} \cdots \frac{f_\Delta(t_{i_k})}{f_0(t_{i_k})} = \\
& \frac{1}{m^m} \left(\prod_{i=1}^m f_0(t_i) \right) \sum_{k=0}^m (m_1)^k (m_0)^{m-k} \sum_{\{i_1, \dots, i_k\} \subseteq \{1, \dots, m\}} \frac{f_\Delta(t_{i_1})}{f_0(t_{i_1})} \cdots \frac{f_\Delta(t_{i_k})}{f_0(t_{i_k})} = \\
& \frac{1}{m^m} \left(\prod_{i=1}^m f_0(t_i) \right) \prod_{j=1}^m \left(m_0 + m_1 \frac{f_\Delta(t_j)}{f_0(t_j)} \right) = \frac{1}{m^m} \prod_{j=1}^m (m_0 f_0(t_j) + m_1 f_\Delta(t_j)) = \frac{1}{m^m} L_{\text{mix}}(m_1, \Delta)
\end{aligned}$$