A symmetric length-aware enrichment test

David Manescu and Uri Keich

School of Mathematics and Statistics, University of Sydney, Australia

Abstract. Young et al. [14] showed that due to gene length bias the popular Fisher Exact Test should not be used to study the association between a group of differentially expressed (DE) genes and a specific Gene Ontology (GO) category. Instead they suggest a test where one conditions on the genes in the GO category and draws the pseudo DE expressed genes according to a length-dependent distribution. The same model was presented in a different context by Kazemian et al. who went on to offer a dynamic programming (DP) algorithm to exactly estimate the significance of the proposed test [8]. Here we point out that while valid, the test proposed by these authors is no longer symmetric as Fisher's Exact Test is: one gets different answers if one conditions on the observed GO category than on the DE set. As an alternative we offer a symmetric generalization of Fisher's Exact Test and provide efficient algorithms to evaluate its significance.

1 Introduction

Fisher's exact test allows us to test whether two binary ("on"/"off") features assigned to a set of objects are correlated by considering the number of objects for which both features are on (loosely speaking, the objects that have both features). Equivalently, if we identify a feature with the set of objects for which the feature is "on", or which share this feature, then what we are gauging is the size of the intersection of the two feature sets.

Fisher [6] gave an example of 30 convicted criminals with same sex twin of which 13 were monozygotic twins and 17 were dizygotic twins. Ten of the 13 monozygotic twins were themselves convicted but only two of the 17 dizygotic twins were also convicted. The question we are interested in is whether the monozygotic/dizygotic feature is correlated with the conviction/no-conviction feature. Fisher's exact test allows us to quantify the level of surprise in the size of the intersection assuming these features are independently assigned to the set of 30 same sex twins.

Calculating the p-value of the test invokes the hypergeometric distribution. Assuming we have an urn with 13 black balls (monozygotic) and 17 red balls (dizygotic) we take a random sample of 12 balls (convicted) and check how many of those are black. More specifically, in Fisher's exact test we find the probability that we will see a result at least as significant as the one we saw. Depending on whether we are interested in a one-sided or a two-sided p-value "at least as significant" has a different meaning. For the one-sided p-value it means having at least as many black balls in the sample as observed in the experiment (10 in our case). For the two-sided p-value it means that the hypergeometric probability of seeing such a number of black balls in our sample is at least as small as the probability that the sample contains the number of black balls observed in the experiment [1][1].

Thinking about it in terms of the urn it is obvious why this test is also referred to as an enrichment test: is our sample of balls (convicted twins) enriched for black balls (monozygotic twins)? Such correlation or enrichment tests are routinely performed in biological research although in some cases the application of Fisher's exact test needs to be evaluated carefully. An example of this is the often posed question of whether a group of genes is enriched for a certain Gene Ontology (GO) category [4].

Young et al. observed that when testing if a set of differentially expressed (DE) genes is enriched for a certain GO category one has to take into account a potential length bias [14]. In other words, one cannot

[1] As pointed out in [1], other formulations of the two-sided p-value include doubling the one-sided p-value and considering "at least as significant" to mean that the deviation between the number of black balls in the sample and its expected value is at least as large as the observed same deviation in the original experiment.

blindly use Fisher's exact test because the underlying null assumption that both the DE genes and the GO category labels are chosen uniformly at random might be violated: longer genes are more likely to be found as DE than shorter ones and similarly, certain GO categories are associated with longer genes, while others are associated with shorter genes.

The solution that Young et al. proposed was to estimate what they refer to as a PWF (probability weighting function) which "quantifies how the probability of a gene selected as DE changes as a function of its transcript length" [14]. Then, assigning each gene a probability proportional to its PWF value they repeatedly sample m genes without replacement, where m is the number of DE expressed genes in the experiment. By noting the statistics of the number of sampled genes in the GO category under consideration they can assign a Monte Carlo (MC) derived p-value to the observed enrichment of this GO category in the actual list of DE genes.

Young et al. noted a downside to their proposed method: "The end result is that, even on a modern cluster, the time taken to calculate p-values to the accuracy necessary to adequately differentiate GO categories from one another is often prohibitive". They therefore provided an additional method which relies on the Wallenius approximation [13] requiring some additional assumptions which might or might not be acceptable: "the approximation assumes that all genes within a category have the same probability of being chosen, but this probability is different from the probability of choosing genes outside this category" [14].

Kazemian et al. faced a similar problem of length bias when studying whether a set of genes which neighbor sequence-predicted CRMs (cis-regulatory modules) is enriched for genes in an "expression gene set" [8]. The expression gene set consisted of the genes with expected expression patterns learned from a training CRM. The same training CRM was also used to estimate a variable length HMM which was in turn used to predict the said loci of novel CRMs.

To address this length bias problem Kazemian et al. introduced the LLHT (locus length-aware Hypergeometric test). The null hypothesis of this test is that each gene is selected independently with probability p_i (they assumed $p_i \propto l_i$ where l_i is essentially the length of the adjacent intergenic region) and that m genes are selected without replacement at random. Here m is the number of genes that neighbor predicted CRMs. Under this null hypothesis they compute the probability that the randomly selected set of m genes will contain at least k genes from the expression gene set. Kazemian et al. go further to provide a dynamic programming (DP) algorithm that allows them to efficiently compute that probability.

It is straightforward to combine Young et al.'s initial step of estimating the probability p_i of selecting gene i with the DP algorithm of Kazemian et al. to achieve what seems to be the ultimate solution to this length-aware enrichment problem. However, as we argue next, this solution, as well as the two previous solutions from which it is synthesized, introduce an undesired side effect into the enrichment test.

Specifically, a key feature of Fisher's exact test is its symmetry: we get the same answer whether we assume the urn contains 13 black balls (monozygotic) and 17 red balls (dizygotic) and we take a random sample of 12 balls (convicted), or we assume the urn contains 18 red balls (not convicted) and 12 black balls (convicted) and we take a sample of 13 balls (monozygotic). This symmetry is apparent when we consider the Fisher exact test as testing the independence of rows and columns in a contingency table with fixed marginals.

Unfortunately, the above length-aware generalizations of the Fisher exact test introduce an inherent asymmetry into the problem formulation: you generally get a different answer if you sample, as Young et al. do, the DE genes and note the overlap with the GO category than if you sample the genes in the GO category and note their intersection with the list of DE genes. The same goes for the two sets of genes considered by Kazemian et al.: you will generally get different answers using their LLHT if you sample the genes that are adjacent to CRMs than if you sample the genes in the expression gene set.

The problem is that in many applications both interpretations of which of the two feature (gene) sets is sampled seems equally plausible. For example, we argue that sampling the genes in the GO category is at least as plausible as sampling the DE genes. So given two different outcomes of the test each corresponding to a different interpretation, how should one decide which of the two p-values is the "right" one?

We begin with setting up a mathematical model for studying the length-aware enrichment tests and go on to demonstrate the asymmetry problem in this context. We then offer an alternative generalization

of Fisher's exact test that retains the latter's symmetry. We also provide several methods of computing the significance of the proposed test including an exact DP algorithm which might be too slow for some applications. For those we develop a fast and fairly accurate saddlepoint approximation.

2 The Model

As in the setup of Fisher's exact test we wish to examine whether the correlation we observe between two sets of features/labels is substantially larger than expected by chance. Each of N "balls" (objects/genes) is assigned a pair of on/off labels: an X label (gene is/not DE) and a Y label (gene is/not in GO category). The correlation between the labels is measured by the number of balls for which both labels are turned on.

Under the null assumption the label indicators X_i and Y_i are $2N$ independent Bernoulli random variables (RVs) with corresponding success probabilities p_i^X and p_i^Y respectively. We are interested in the distribution of the size of the intersection set which is given by the RV

$$Z = \sum_{i=1}^{N} X_i Y_i.$$

In the symmetric setup we are interested in the conditional distribution of Z given that m of the balls have feature X and k have feature Y, i.e., in

$$P\left(Z = l \mid \sum_{i=1}^{N} X_i = m, \sum_{i=1}^{N} Y_i = k\right) = P\left(Z = l \mid A_{mk}\right), \tag{1}$$

where A_{mk} is the event

$$A_{mk} = \left\{\sum_{i=1}^{N} X_i = m, \sum_{i=1}^{N} Y_i = k\right\}.$$

Let z be the observed size of the intersection. The one-sided p-value of z under our symmetric enrichment test is the conditional probability[2]

$$P\left(Z \geq z \mid A_{mk}\right) = \sum_{l \geq z} P\left(Z = l \mid A_{mk}\right). \tag{2}$$

The two sided p-value of z under our symmetric enrichment test is the conditional probability that we will observe an intersection $Z = l$ at least as surprising as the observed z. Here "at least as surprising" has the same meaning as in the two-sided Fisher's exact test, i.e., the intersection of size l is less likely than the intersection of size z:

$$P\left(Z = l \mid A_{mk}\right) \leq P\left(Z = z \mid A_{mk}\right). \tag{3}$$

More explicitly, the two-sided p-value is given by

$$\sum_{l \text{ for which (3) holds}} P\left(Z = l \mid A_{mk}\right). \tag{4}$$

It is not hard to show that the symmetric enrichment test is indeed a generalization of Fisher's exact test: if the probabilities of all X labels are the same $p_i^X \equiv p^X$ and if similarly $p_i^Y \equiv p^Y$ then this new test is identical to Fisher's exact test. Indeed, under such uniform conditions the conditional distribution (1) is the hypergeometric distribution with N balls, m of them black and a sample of size k is taken (or, equivalently, k are black and a sample of m balls is taken).

[2] This one-sided p-value is designed for an alternative that expects positive correlation between the labels. There is an analogous one-sided p-value when the alternative specifies negative correlation between the labels.

The asymmetric model of Young et al. [14] and of Kazemian et al. [8] is obtained by replacing (1) from the symmetric analysis with either the conditional probability

$$P\left(Z=l \mid \sum_{i=1}^{N} X_i = m,\, Y_1, Y_2, \ldots, Y_N\right) = P\left(\sum_{i:Y_i=1} X_i = l \mid \sum_{i=1}^{N} X_i = m\right), \tag{5}$$

or with

$$P\left(Z=l \mid X_1, X_2, \ldots, X_N,\, \sum_{i=1}^{N} Y_i = k\right) = P\left(\sum_{i:X_i=1} Y_i = l \mid \sum_{i=1}^{N} Y_i = k\right). \tag{6}$$

In both cases we can formulate a 1-sided or a 2-sided test by appropriately modifying the corresponding symmetric tests (2) and (4).

The Label Probabilities

So far we only assumed that the ith ball/object is independently assigned each of two types of on/off labels: x_i and y_i. This suffices to define our one-sided (2) or two-sided (4) symmetric enrichment test, as well as their asymmetric analogues. However, as in [14] and in [8] there is another aspect to our model, namely, we assume each object (gene) has a property (say, length) which uniquely determines its label probabilities p_i^X, and p_i^Y. In other words, we assume the existence of functions $\varphi_X, \varphi_Y : \mathbb{N} \mapsto [0,1]$ such that

$$P\left(X_i=1\right) = \varphi_X\left(l_i\right) \quad \text{and} \quad P\left(Y_i=1\right) = \varphi_Y\left(l_i\right),$$

where l_i is the length, or more generally, the value of the property that determines the label-probability of the ith object.

Going back to the example of enrichment of a GO category in a list of DE expressed genes [14], we assume the probability of the gene being DE is uniquely determined by its length. Similarly, in [8] the genes adjacent to predicted CRMs are chosen with probability proportional to the intergene's length[3] so in this latter case $\varphi_X\left(l_i\right) \propto l_i$, where l_i is the adjacent intergene length.

Young et al. estimate φ_X, which they refer to as a "probability weighting function", using a monotone increasing spline. Here we chose to forgo the monotone assumption and use the LOESS locally weighted regression procedure [3] for estimating both φ_X and φ_Y: as we address the symmetric problem we need to estimate both label probabilities functions.

Specifically, we use the `R` function `loess` [10] twice, first for regressing the observed x_i labels on the length (label probability defining property) and the second time for regressing the y_i labels. The function is called with all its default parameters which, among other things, implies it is using a quadratic local approximation.

Note that by construction $\sum_i p_i^X \approx m$ and $\sum_i p_i^Y \approx k$ – a property which will be relevant below.

3 The Asymmetry Problem

The problem with the asymmetric test is that often each one of the two interpretations (5) and (6) is equally plausible and at the same time they would typically generate different p-values. The reason is that in (5) we condition on the observed Y_i hence we factor the length bias only into the selection of the corresponding X_i and vice versa for (6).

The situation is particularly unfortunate when one of those interpretation generates a p-value below the significance cutoff while the other does not. Which one should we choose in such a case?

In what follows we first demonstrate analysis of real data in which this issue came up. We then use simulations to demonstrate the potential extent of this problem in practice.

[3] No rigorous justification for this assumption is given which should be considered as intuitively derived approximation.

Are Syntenically Conserved Intergenes Enriched for *S. cerevisiae* Replication Origins? Consider the following real biological data which motivated our interest in this problem. Processing the October 2003 version of the *S. cerevisiae* genome we have $N = 6355$ intergenic regions. Using the gene homology map of [11] (supplementary table 1) we determined the 5113 *S. cerevisiae* intergenes that are syntenically conserved (the exact protocol that was used to determine this intergene synteny is immaterial here).

At the same time we have a list of 353 *S. cerevisiae* replication origins that reside in *S. cerevisiae* intergenic regions (again the exact protocol is immaterial, it suffices to say most came from OriDB [9]). Of those 353 origins, 236 coincide with the syntenically conserved intergenic regions while 117 origins lie in non-conserved intergenes.

Is there evidence that conserved intergenic regions are enriched for origins? While over 80% of the intergenes (5113 / 6355) are conserved only 67% (236 / 353) of the intergenic origins lie in conserved intergenes. This suggests that, if anything, there is an inexplicable depletion of origins in conserved intergenes. Moreover, a two sided Fisher exact test finds this depletion statistically significant with a miniscule p-value less than $5 \cdot 10^{-10}$.

Of course, although Fisher's exact test is symmetric it ignores the length bias: an intergene is more likely to contain an origin if it is longer and at the same time it is less likely to be conserved. Redoing the statistical analysis factoring the length bias (with the length specific feature probabilities learned as explained above) using our symmetric enrichment test we get a 2-sided p-value of 0.051 using our saddlepoint approximation[4]. As this is a borderline p-value we also estimated it using 1,000,000 MC samples and obtained a 2-sided p-value of 0.052. Hence, the depletion according to our test is borderline insignificant which is very different from the highly significant result reported by Fisher's exact test.

Importantly, this question also demonstrates the problem with the asymmetry of the other length aware tests. Namely, if we condition on the observed syntenic intergenes and factor the length only into the selection of the origins we get a significant p-value of 0.01. At the same time if we condition on the observed origins and factor the length only into the selection of the conserved intergenes we get an insignificant p-value of 0.18. This is one of those cases where each of the interpretations is equally plausible so we are left with an undesirable choice in our hands. Using the symmetric enrichment test naturally removes this ambiguity.

How Often Can the Asymmetry Become a Real Issue? Is the last example an extraordinary one? How often do we expect to run into a problem where the asymmetry could be an issue in the sense that it could give two qualitatively different answers? To get some intuition into this we designed the following simulation test.

Using the label probabilities p_i^X and p_i^Y we independently drew labels x_i and y_i. We then computed the enrichment statistic $z = \sum_i x_i y_i$ and performed three length-aware enrichment tests: our symmetric one and the two asymmetric ones, first conditioning on the observed set of y_i and on $\sum x_i$ as in (5) and then conditioning on the observed set of x_i and on $\sum y_i$ as in (6). Repeating this experiment many times we were able to gauge how often and how badly do these three tests disagree on samples in the critical 0.05 region.

In all we generated four test sets using two sets of label probabilities as specified in Table 1. We also varied our label sampling procedure: in test sets 1 & 3 we sampled the labels unconditionally whereas in the other two sets we sampled conditioned on the sums specified in Table 1.

As can be seen in Tables 1 and 3 there is substantial disagreement between the two asymmetric interpretations among themselves as well as between each of them and the symmetric test. For example, in all four test sets about 50% of the times that the asymmetric test (5) rejected the null at the 0.05 level, the homologous asymmetric test, (6), did *not* reject the null at the same level.

[4] Note that this problem is "too large" for estimating the significance of the test using our exact method so an approximation method was required (more on these significance evaluation methods below).

test set	symmetric ≤ 0.05 cond on $x > 0.05$	symmetric ≤ 0.05 cond on $y > 0.05$	cond on $x \leq 0.05$ cond on $y > 0.05$	cond on $y \leq 0.05$ cond on $x > 0.05$
1	23%	38%	47%	49%
2	29%	37%	44%	49%
3	34%	36%	59%	59%
4	31%	35%	57%	55%

Table 1. Disagreement between the asymmetric and symmetric enrichment tests. To test how often we can expect the three enrichment tests (two asymmetric and one symmetric) to disagree in the critical region we looked at the percentage of samples that would be called significant according to test 1 that will be insignificant according to test 2. The significance level used here is the canonical 0.05 level and we used four different test sets. Columns 3-4 show that the two asymmetric tests ("cond on x" refers to the asymmetric test using (6) and "cond on y" to using (5)) greatly disagree among themselves on whether or not a sample in the critical region is significant. In test set 1, 1,000 samples were generated by sampling the labels independently according to the label probabilities estimated from the replication origins data mentioned above. In test set 2 we sampled 1,000 pairs of labels labels using the same label probabilities as in test set 1 except that we conditioned on $\sum_i x_i = 1242$ and $\sum_i y_i = 353$. Test set 3 was made of 10,000 pairs of label samples generated independently according to the label probabilities $p_i^X = p_i^Y = i/300$ for $i = 1, 2, \ldots, 300$. For test set 4 we generated 10,000 samples using the same label probabilities as in test set 3 only we conditioned on specific observed sums: $\sum_i x_i = \sum_i y_i = 150$. For test sets 1 & 2 the exact method was too slow for calculating the symmetric p-value so we used the saddlepoint method instead. Exact methods were used for all the asymmetric tests as well as for the symmetric tests of test sets 3 & 4.

4 Computing the p-value of the Symmetric Test

In this section we present several ways to compute or approximate the one/two-sided p-value of our symmetric enrichment test. We begin by showing how to compute the p-value exactly using first principles. This type of computation is often referred to as an exact test.

Exact Test Using DP As X_i and Y_i are independent,

$$P(A_{mk}) = P\left(\sum_{i=1}^{N} X_i = m\right) P\left(\sum_{i=1}^{N} Y_i = k\right).$$

Using a straightforward DP implementation of the convolutions, $P\left(\sum_{i=1}^{N} X_i = m\right)$ can be computed exactly in a runtime complexity of $O(mN)$ implying that $P(A_{mk})$ can be computed in $O((m+k)N)$.

Therefore, to evaluate the significance of our test using (4) we need to compute $P(Z = l, A_{mk})$ for every $l = 0, 1, \ldots, N$. These probabilities can be computed exactly using DP based on the following recursive formula: for any $n > 0$ and $l, j, r \in \{0, 1, \ldots, N\}$

$$\begin{aligned}
P\left(\sum_{i=1}^{n} X_i Y_i = l, \sum_{i=1}^{n} X_i = j, \sum_{i=1}^{n} Y_i = r\right) &= p_n^X p_n^Y P\left(\sum_{i=1}^{n-1} X_i Y_i = l-1, \sum_{i=1}^{n-1} X_i = j-1, \sum_{i=1}^{n-1} Y_i = r-1\right) \\
&+ p_n^X \left(1 - p_n^Y\right) P\left(\sum_{i=1}^{n-1} X_i Y_i = l, \sum_{i=1}^{n-1} X_i = j-1, \sum_{i=1}^{n-1} Y_i = r\right) \\
&+ \left(1 - p_n^X\right) p_n^Y P\left(\sum_{i=1}^{n-1} X_i Y_i = l, \sum_{i=1}^{n-1} X_i = j, \sum_{i=1}^{n-1} Y_i = r-1\right) \\
&+ \left(1 - p_n^X\right)\left(1 - p_n^Y\right) P\left(\sum_{i=1}^{n-1} X_i Y_i = l, \sum_{i=1}^{n-1} X_i = j, \sum_{i=1}^{n-1} Y_i = r\right).
\end{aligned}$$

The base, or the boundary condition of the recursion is:

$$P\left(\sum_{i=1}^{n} X_i Y_i = l, \sum_{i=1}^{n} X_i = j, \sum_{i=1}^{n} Y_i = r\right) = \begin{cases} 0 & 0 > \min\{l, j, r\} \text{ or } l > \min\{j, r\} \text{ or } \max\{r, j\} > n \\ 1 & n = l = j = r = 0 \end{cases}.$$

Therefore computing $P(Z = l, A_{mk})$ for all l can be done in a runtime complexity of $O(\min\{m, k\} mkN)$.

Since in a typical genomic setting N, m, k can be of the order of several thousands this exact calculation can prove too costly. For example, had we tried to analyze the ARS enrichment problem presented above using this exact method we estimate it would have taken us 87 days on a single processor machine.

Normal Approximation In situations where exact calculation of the p-value is prohibitively slow we need to look for approximations. For example, can try to approximate the conditional distribution of Z given A_{mk} using a normal $N(\mu, \sigma^2)$ distribution.

Computing the mean μ and the variance σ^2 of this conditional distribution can be done in an exact manner as described in the appendix where we show that the runtime complexity of computing these moments is $O(mN^2)$.

At $O(mN^2)$ an exact computation of the conditional moments of Z proved costly in many realistic settings. In addition, the calculation of the conditional moments is prone to significant accumulation of roundoff errors. We therefore looked for an alternative, approximate calculation of the conditional moments which we describe next.

Normal Approximation with Approximate Moments The bottleneck in computing the conditional variance both in terms of numerical stability and speed was in estimating $P\left(\sum_{l \notin \{i,j\}} X_l = m - 2\right)$ in (11). One obvious way to bypass that difficulty is to replace the exact calculation of these probabilities with their normal derived approximation. In other words our normal approximation will now use approximate moments, themselves derived from a normal approximation.

Note however that in computing the moments here we consider the *un-conditional* distribution of $\sum_{l \notin \{i,j\}} X_l$ so its mean and variance are readily computed: the mean is $\sum_{l \notin \{i,j\}} p_l^X$ and the variance is $\sum_{l \notin \{i,j\}} p_l^X \left(1 - p_l^X\right)$. We can compute these for all i, j at a total cost of $O(N^2)$.

Keep in mind that in our application $\sum_i p_i^X \approx m$ by construction so the probabilities we are estimating here using the normal approximation are near the mode of the distribution where the normal approximation is at its best.

MC Simulations For p-values which are not very small[5] we can always resort to MC sampling for approximating the p-value. Conceptually, we can draw samples of the two sets of labels x_i and y_i according to the corresponding label probabilities p_i^X and p_i^Y and reject all the sampled sets for which either $\sum_i x_i \neq m$ or $\sum_i y_i \neq k$. We can then construct the empirical distribution of $z = \sum_i x_i y_i$ from the samples that were not rejected and use it as an estimate of the conditional distribution of Z given A_{mk}. Using this empirical distribution as a surrogate for the actual (1) we can then estimate the one-sided (2) and two-sided p-values (4).

Sampling by rejection as above is conceptually straightforward and it is easy to implement however it can be very inefficient. A much faster MC simulations can be achieved if we can efficiently sample directly from the conditional distribution.

As X_i are independent of Y_i we can generate a sample from the conditional distribution by sampling the X_i conditioned on $\sum_i X_i = m$ and sampling the Y_i conditioned on $\sum_i Y_i = k$. The latter can be done efficiently using the following iterative scheme.

[5] For example, for the fairly large ARS enrichment problem with $N = 6355$ we can generate 10^6 MC samples in 9,000 seconds using a single core machine.

Sample X_1 conditioned $\sum_i X_i = m$ using (9) with $i = 1$. Then iteratively sample X_i given the sampled values $x_1, \ldots, x_{i-1}$ using

$$P\left(X_i = 1 \,|\, \sum_j X_j = m\,, X_1 = x_1, \ldots, X_{i-1} = x_{i-1}\right) = \frac{p_i^X P\left(\sum_{j>i} X_j = m - \sum_{j<i} x_j - 1\right)}{P\left(\sum_{j\geq i} X_j = m - \sum_{j<i} x_j\right)}.$$

To compute the RHS we need to find $P\left(\sum_{j>i} X_j = l\right)$ for all $l = 0, \ldots, m$ and all $i = 1, \ldots, N-1$. These can be pre-calculated using iterative convolutions at the same runtime complexity that it takes to compute the distribution of $\sum_{i=1}^N X_i$ which is $O(mN)$. Thus, we can generate a sample of n sets of labels conditioned on A_{mk} in an overall time complexity of $O((m+k+n)\,N)$.

Saddlepoint Approximation While normal approximations typically work well for moderate p-values their accuracy is often less than desirable when it comes to smaller p-values. This is the region where saddlepoint approximations generally do much better. The main downside of saddlepoint methods is that their implementation is more involved than the normal approximation. Here we chose to use the double saddlepoint approximation for conditional distribution of [12] which is conveniently summarized in [2].

The approximation uses the joint cumulant generating function (CGF) of $X = \sum_i X_i$, $Y = \sum_i Y_i$ and $Z = \sum_i Z_i$ defined as

$$K_{(X,Y,Z)}(r,s,t) = \log M_{(X,Y,Z)}(r,s,t)\,.$$

The term $M_{(X,Y,Z)}$ is the joint moment generating function (MGF) of (X,Y,Z) which can be computed using the independence of (X_i, Y_i, Z_i) from (X_j, Y_j, Z_j) for $i \neq j$:

$$M_{(X,Y,Z)}(r,s,t) = \mathrm{E}\left(e^{rX+sY+tZ}\right) = \prod_{i=1}^{N} \mathrm{E}\left(e^{rX_i+sY_i+tZ_i}\right). \tag{7}$$

Further details on the saddlepoint approximation can be found in the appendix where we show its runtime complexity is $O(N)$.

5 Comparison of the Symmetric p-value Approximation Schemes

method	complexity	uniform marginals $z = 117$	enrichment of ARSs
Exact	$O(\min\{m,k\}\,mkN)$	1672 secs	87 days (estimate)
Normal with exact moments	$O\left(mN^2\right)$	33 secs	26 hours (estimate)
Normal with estimated moments	$O\left(N^2\right)$	0.3 secs	147 secs
MC with n samples	$O((m+k+n)\,N)$	256 secs ($n = 10^6$)	2.5 hours ($n = 10^6$)
Saddlepoint	$O(N)^*$	1.1 secs	30 secs

Table 2. Theoretical and practical runtime of the two-sided p-value calculation. The table provides the theoretical runtime complexity of each of the approximations we discussed, as well as its actual runtime on a couple of realistic examples. The "uniform marginals" column specifies the actual runtime of computing the corresponding 2-sided approximation of the symmetric enrichment test ("method") for the observed value of $z = 117$ and where $p_i^X = p_i^Y = \frac{1}{2}$ for $i = 1, 2, \ldots, 300$ (which reduces to the hypergeometric case or to Fisher Exact Test). The enrichment of ARS column lists the actual (or estimated) runtime it tool to analyze our real world replication origins data mentioned above. Note that for the saddlepoint the complexity is $O(N)$ times the number of iterations the root solving algorithm take though typically that number is small.

Table 2 summarizes the theoretical complexity and the actual runtime of the approximation schemes we presented. The saddlepoint method has the smallest complexity and this is well demonstrated in the 30 seconds it took it to analyze our actual ARS enrichment problem where $N = 6355$, $m = 5113$ and $k = 353$. Compare that with the 2.5 hours it took to get 1e6 MC samples to estimate this significance, and with the estimated 87 days it would have taken for the exact method (all on a single core desktop). The normal approximation with estimated moments was also quite fast for this problem taking only 147 secs. Moreover, it was even slightly faster (0.3 vs. 1.1 seconds) than the saddlepoint method on a much smaller problem where $m = 150$, $k = 150$ and $N = 300$ and where we set $p_i^X \equiv p_i^Y = 1/2$ and an "observed" value of $z = 117$. However, while the normal approximation with estimated moments was reasonably fast its accuracy was not comparable to that of the saddlepoint as described next.

In order to study the accuracy of the approximation methods we presented we simulated 3 examples. In each case we used $N = 300$, $m = 150$ and $k = 150$ but we varied the set of label probabilities in each example as detailed in Figure 1. This figure presents plots of (the absolute value of the base 10 logarithm of) the ratio between the approximated p-values and the exact p-values for each possibly observed value.

The MC method is accurate for a limited range which extends beyond the range in which the normal approximation is accurate (note that the normal approximation with exact moments exhibited similar accuracy to the one obtained from estimated moments but was significantly slower). Importantly, the saddlepoint method stays reasonably close to the exact computation throughout the entire range of possible values (Figure 1). Moreover, except for the very extreme values, where the probability is miniscule anyhow it provides an excellent approximation to the exact p-value. For example, for the uniform label probabilities problem mentioned above, panel (A), and for $z = 0$ the exact p-value is 1.07e-89 and the saddlepoint method gives 9.25e-90, for $z = 5$ the exact method yields 3.74e-72 and the saddlepoint gives 3.69e-72 and for $z = 50$ both methods give 5.69e-09.

6 Discussion

It was previously demonstrated that the classical Fisher exact test needs to be adjusted when testing for correlation between a group of DE genes and a specific GO category [14]. Such adjustment is required more generally whenever the observations whose mutual enrichment is tested (the gene is DE and the gene belongs to the GO category) depend on some inert properties of the observed objects (the gene's length).

Young et al. proposed estimating the conditional probability that a gene is DE given its length which they refer to as a PWF. They then use this PWF in a MC sampling procedure to estimate the null distribution of the overlap between the sampled DE genes and the GO category considered.

Faced with a similar enrichment problem Kazemian et al. introduced a DP algorithm to exactly calculate the conditional probability of the observed overlap between predcited CRMs and their predicted expression pattens given the latter [8]. Rather than try to estimate the conditional probability that an intergene will contain a predicted CRM given its length, Kazemian et al. assumed this probability is proportional to the intergene length. It is however straightforward to combine Young et al.'s estimation procedure with Kazemian et al.'s exact calculation.

Both Young et al. and Kazemian et al. used an asymmetric model; they studied the overlap/enrichment conditioned on one of the set of observations: the considered GO category in the first case and the gene expression set in the second. In both of these cases the authors' choice of observations to condition on is not necessarily more plausible than conditioning on the alternative set of observations.

The problem is that unlike with Fisher's exact test where it does not matter which of the two set of observations you condition on, in these length aware generalizations you generally get different answers depending on which set you condition on. This leaves the users with an undesirable choice at their hands which, as we show, can often yield contradicting results in terms of the significance of the observed enrichment.

To address this problem we introduce the symmetric enrichment test which is appropriate where conditioning on one set of the observed labels is equally plausible as conditioning on the other set. We propose and compare several methods for computing the one-sided and two-sided p-values of this test concluding that only two of them should be recommended. Whenever feasible the exact DP calculation has the obvious

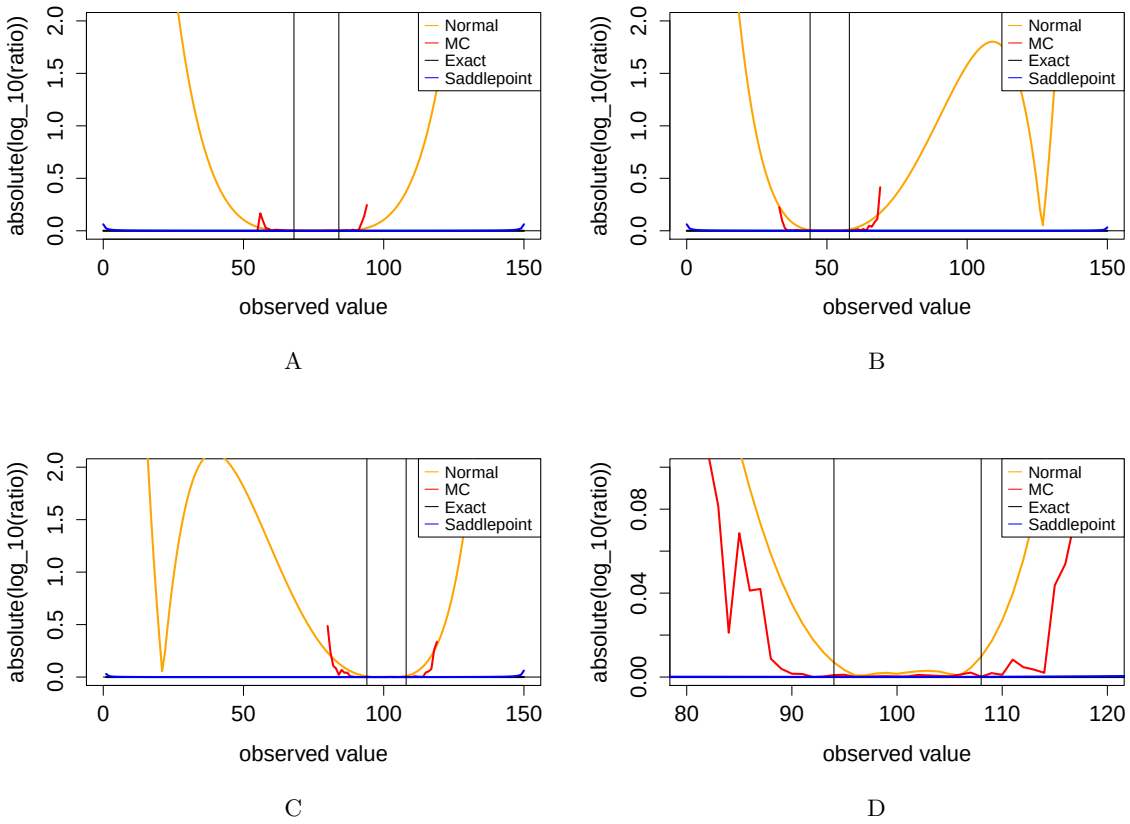

Fig. 1. Accuracy of approximations of the symmetric enrichment test. The p-value of the appropriate one-sided test was evaluated for each possible value of the statistic using an exact calculation, the normal approximation with estimated moments (using exact moments gave almost identical results), MC estimation (using 10^6 samples), and the saddlepoint approximation. Each figure gives (the absolute value of the log base 10 of) the ratio of each of the approximation methods to the exactly computed p-value of each of the theoretically observable values of the statistic. The panels correspond to: (A) The standard hypergeometric case with $p_i^X = p_i^Y = \frac{1}{2}$ for $i = 1, 2, \ldots, 300$, (B) label probabilities $p_i^X = i/300$ and $p_i^Y = (301 - i)/300$ for $i = 1, 2, \ldots, 300$, (C) label probabilities $p_i^X = p_i^Y = i/300$ for $i = 1, 2, \ldots, 300$, (D) same example as (C) but the center part is zoomed in. The two vertical lines in each plot correspond to the significance thresholds of the exact p-value calculation. The saddlepoint offers a good approximation throughout the entire range of values and an excellent approximation as long as we are not at one of the extreme values (for which the p-values are typically miniscule anyhow, e.g., for (A) the exact p-value at $z = 0$ is 1.07e-89).

advantage of giving an accurate answer. However, as the exact calculation may often be too slow for analysis of genomic enrichment tests, we recommend our saddlepoint approximation as an overall fairly accurate and relatively fast substitute.

Note that we do *not* suggest that the symmetric test should always be preferred to the asymmetric tests. In some cases the asymmetry might be inherent to the problem. For example, if we are testing the enrichment of *many* GO categories against a *single* set of DE genes then conditioning on the observed list of DE genes and sampling the genes in the GO category might be more natural than sampling both[6]. However, in general the user should be aware that conditioning on one set (DE genes) will yield different results than conditioning on the other set (GO category) and if both conditionings are equally plausible then the proposed symmetric test should be considered. That, for example, would be the case, if we only test the correlation of a single GO category with a given set of DE genes.

We should keep in mind that the symmetric test requires estimating two sets of label probabilities rather than a single set required by the asymmetric tests. Therefore, if one of those estimations seems dubious it might be preferable to use the asymmetric test that conditions on the corresponding observed labels.

Finally, another consideration of which test one should use is its power: how likely is the test to reject the null when the labels are in fact correlated. In future work we plan to compare the power of these different enrichment tests as well as other possible generalizations of Fisher's Exact Test. After all, if the label probabilities are assumed known other tests that do not condition on the observed number of labels might even be more powerful.

Scripts for conducting the proposed length aware symmetric enrichment test will be available to download from `http://www.maths.usyd.edu.au/u/uri/`

7 Acknowledgment

The authors would like to thanks anonymous reviewers for their constructive comments.

References

1. Alan Agresti. A survey of exact inference for contingency tables. *Statistical Science*, 7:131–153, 1992.
2. RW Butler. *Saddlepoint Approximations with Applications*. Cambridge University Press, 2007.
3. William S. Cleveland and Susan J. Devlin. Locally-weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83:596–610, 1988.
4. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.
5. Wayne R. Cowell, editor. *Sources and Development of Mathematical Software*. Prentice-Hall Series in Computational Mathematics, Cleve Moler, Advisor. Prentice-Hall, Upper Saddle River, NJ 07458, USA, 1984.
6. RA Fisher. *Statistical methods for research workers*. Oliver & Boyd, London, 14th ed. edition, 1970.
7. Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–.
8. Majid Kazemian, Qiyun Zhu, Marc S. Halfon, and Saurabh Sinha. Improved accuracy of supervised crm discovery with interpolated markov models and cross-species comparison. *Nucleic Acids Research*, 39(22):9463–9472, Dec 2011.
9. CA Nieduszynski, S Hiraga, P Ak, CJ Benham, and AD Donaldson. Oridb: a dna replication origin database. *Nucleic Acids Res*, 35(Database issue):D40–D46, Jan 2007.
10. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.
11. Devin R. Scannell, Oliver A. Zill, Antonis Rokas, Celia Payen, Maitreya J. Dunham, Michael B. Eisen, Jasper Rine, Mark Johnston, and Chris Todd Hittinger. The awesome power of yeast evolutionary genetics: New genome sequences and strain resources for the saccharomyces sensu stricto genus. *G3 (Bethesda)*, 1(1):11–25, Jun 2011.
12. IM Skovgaard. Saddlepoint expansions for conditional distributions. *J. Appl. Prob.*, 24:875–87, 1987.

[6] Note that presumably for computational efficiency reasons Young et al. sample the DE genes rather than the GO category [14].

13. KT Wallenius. *Biased sampling: the non-central hypegeometric probability distribution.* PhD thesis, Stanford University, 1963.
14. MD Young, MJ Wakefield, GK Smyth, and A Oshlack. Gene ontology analysis for rna-seq: accounting for selection bias. *Genome Biology , 11:R14*, 11, 2010.

8 Appendix

test set	symmetric ≤ 0.04 asymmetric > 0.06 x-cond (6)	y-cond (5)	cond on $x \leq 0.04$ (6) sym > 0.06	y-cond > 0.06	cond on $y \leq 0.04$ (5) sym > 0.06	x-cond > 0.06
1	12%	21%	3%	21%	23%	37%
2	13%	31%	17%	37%	20%	27%
3	34%	36%	25%	53%	26%	51%
4	31%	35%	27%	52%	25%	50%

Table 3. Disagreement between the asymmetric and symmetric enrichment tests Same as Table 1 except the columns now indicate the level of discrepancy between the various enrichment tests by showing the percentage of those samples which were (null-) rejected according to test 1 at the more significant 0.04 that were found insignificant even at the 0.06 level by test 2.

8.1 Normal Approximation - Exact Computation of the Conditional Moments

Let $Z_i = X_i Y_i$ then

$$\mu = \mathrm{E}\left(Z \mid A_{mk}\right) = \sum_{i=1}^{N} \mathrm{E}\left(Z_i \mid A_{mk}\right) = \sum_{i=1}^{N} P\left(Z_i = 1 \mid A_{mk}\right). \tag{8}$$

Therefore we need to find $P\left(Z_i = 1 \mid A_{mk}\right)$ which due to the independence of the RVs can be computed from

$$P\left(Z_i = 1 \mid A_{mk}\right) = P\left(X_i = 1 \mid \sum_j X_j = m\right) P\left(Y_i = 1 \mid \sum_j Y_j = k\right).$$

The terms on the right hand side (RHS) can be computed using

$$P\left(X_i = 1 \mid \sum_j X_j = m\right) = \frac{p_i^X P\left(\sum_{j \neq i} X_j = m - 1\right)}{P\left(\sum_{i=1}^{N} X_i = m\right)}, \tag{9}$$

and similarly for the corresponding term in Y.

Computing $P\left(\sum_{i=1}^{N} X_i = l\right)$ for all $l \leq m$ can be done using straightforward DP with a time complexity of $O(mN)$ and a space complexity of $O(m)$. Then using the recursive formula

$$P\left(\sum_{i=1}^{N} X_i = l\right) = p_i^X P\left(\sum_{j \neq i} X_j = l - 1\right) + \left(1 - p_i^X\right) P\left(\sum_{j \neq i} X_j = l\right), \tag{10}$$

and

$$P\left(\sum_{j \neq i} X_j = 0\right) = \prod_{j=1}^{N} \left(1 - p_j^X\right) / \left(1 - p_i^X\right),$$

we can compute $P\left(\sum_{j \neq i} X_j = m - 1\right)$ in an additional $O(m)$ steps for each i, or in a total complexity of $O(mN)$ for all i. Note that the latter is the same as the complexity of computing $P\left(\sum_{i=1}^{N} X_i = m\right)$ to begin with, so this is also the overall runtime complexity of computing (8).

Computing the conditional variance is somewhat more involved. As Z_i are Bernoulli RVs their conditional variance is given by

$$\sigma^2 = \mathrm{Var}\left(Z_i \mid A_{mk}\right) = P\left(Z_i = 1 \mid A_{mk}\right)\left[1 - P\left(Z_i = 1 \mid A_{mk}\right)\right],$$

where $P\left(Z_i = 1 \mid A_{mk}\right)$ is computed above. As for the conditional pairwise covariances we have

$$\mathrm{Cov}\left(Z_i, Z_j \mid A_{mk}\right) = P\left(Z_i = 1, Z_j = 1 \mid A_{mk}\right) - P\left(Z_i = 1 \mid A_{mk}\right) P\left(Z_j = 1 \mid A_{mk}\right).$$

Thanks again to the independence we have

$$P\left(Z_i = 1, Z_j = 1 \mid A_{mk}\right) = P\left(X_i = 1, X_j = 1 \mid \sum_l X_l = m\right) P\left(Y_i = 1, Y_j = 1 \mid \sum_l Y_l = k\right).$$

The RHS above can be found using the following analogous formula to (9)

$$P\left(X_i = 1, X_j = 1 \mid \sum_l X_l = m\right) = \frac{p_i^X p_j^X P\left(\sum_{l \notin \{i,j\}} X_l = m - 2\right)}{P\left(\sum_{i=1}^N X_i = m\right)}, \tag{11}$$

with an obvious analogue for Y.

The new term on the RHS of (11) can be found from the distribution of $\sum_{j \neq i} X_j$ (required for computing the conditional mean) using the analogue of (10) at a runtime complexity of $O(m)$ for each pairs of indices i, j, or $O\left(mN^2\right)$ in total. This term dominates the complexity of all other steps so it is the overall cost of computing the conditional variance as

$$\mathrm{Var}\left(Z \mid A_{mk}\right) = \sum_i \mathrm{Var}\left(Z_i \mid A_{mk}\right) + \sum_{i,j} \mathrm{Cov}\left(Z_i, Z_j \mid A_{mk}\right).$$

8.2 The Saddlepoint Approximation (Details)

Using (7) the CGF can be computed in a runtime of $O(N)$ using

$$\begin{aligned} K_{(X,Y,Z)}(r,s,t) &= \sum_{i=1}^N \log \mathrm{E}\left(e^{rX_i + sY_i + tZ_i}\right) \\ &= \sum_{i=1}^N \log\left[e^{r+s+t} p_i^X p_i^Y + e^r p_i^X \left(1 - p_i^Y\right) + e^s \left(1 - p_i^X\right) p_i^Y + \left(1 - p_i^X\right)\left(1 - p_i^Y\right)\right]. \end{aligned}$$

The approximation involves K', the gradient vector, as well as K'', the 3×3 Hessian matrix of $K = K_{(X,Y,Z)}$ which can be computed by differentiating the above sum term by term. For example,

$$\frac{\partial K}{\partial r} = \sum_{i=1}^N \frac{e^{r+s+t} p_i^X p_i^Y + e^r p_i^X \left(1 - p_i^Y\right)}{e^{r+s+t} p_i^X p_i^Y + e^r p_i^X \left(1 - p_i^Y\right) + e^s \left(1 - p_i^X\right) p_i^Y + \left(1 - p_i^X\right)\left(1 - p_i^Y\right)},$$

and

$$\frac{\partial^2 K}{\partial r \partial t} = \sum_{i=1}^N \frac{e^{r+s+t} p_i^X p_i^Y \left[e^s \left(1 - p_i^X\right) p_i^Y + \left(1 - p_i^X\right)\left(1 - p_i^Y\right)\right]}{\left[e^{r+s+t} p_i^X p_i^Y + e^r p_i^X \left(1 - p_i^Y\right) + e^s \left(1 - p_i^X\right) p_i^Y + \left(1 - p_i^X\right)\left(1 - p_i^Y\right)\right]^2}.$$

Each of these derivatives can be computed again in $O(N)$ which is therefore the runtime complexity of computing K' and K'' for any particular value of $(r, s, t) \in \mathbb{R}^3$.

The approximation also requires finding the roots $\hat{r}_0, \hat{s}_0 \in \mathbb{R}$ of the following two 1-dimensional equations:

$$\begin{aligned}\frac{\partial K_X}{\partial r} &= \sum_{i=1}^{N} \frac{M'_{X_i}}{M_{X_i}} = \sum_{i=1}^{N} \frac{e^r p_i^X}{e^r p_i^X + \left(1 - p_i^X\right)} = m \\ \frac{\partial K_Y}{\partial s} &= \sum_{i=1}^{N} \frac{M'_{Y_i}}{M_{Y_i}} = \sum_{i=1}^{N} \frac{e^s p_i^Y}{e^s p_i^Y + \left(1 - p_i^Y\right)} = k,\end{aligned} \tag{12}$$

where K_X and K_Y are the CGFs of $X = \sum_i X_i$ and of $Y = \sum_i Y_i$. Note that both CGFs as well as their first two derivatives can again be computed[7] for each given value of r (or s) in $O(N)$.

We solve these two equations numerically using the `python` function `scipy.optimize.fsolve` [7] which in turn is based on `MINPACK`'s `hybrj` algorithm [5]. As the Hessian (or simply the second derivative in this case) is passed to the function it typically requires only a few evaluations before converging on a value $\hat{r}_0$ (or $\hat{s}_0$) that is within the default tolerance parameter of $\approx 1.5e-8$ to the exact root.

Finally we need to solve the following set of 3 equations in 3 unknowns $\left(\tilde{r}, \tilde{s}, \tilde{t}\right) \in \mathbb{R}^3$:

$$\left(\frac{\partial K}{\partial r}, \frac{\partial K}{\partial s}, \frac{\partial K}{\partial t}\right)\Bigg|_{\left(\tilde{r},\tilde{s},\tilde{t}\right)} = K'\left(\tilde{r}, \tilde{s}, \tilde{t}\right) = (m, k, z - 0.5), \tag{13}$$

where z is the observed value of Z. We again solve (13) numerically using `scipy.optimize.fsolve` which given the Hessian requires only a few evaluations of K' and the K'', each taking $O(N)$, before converging to a solution which within the tolerance parameter of 10^{-15}.

Our approximation of the one-sided p-value (with the alternative hypothesis being that the intersection set is larger than expected by chance) is based on (4.17) from [2] which is repeated here for convenience:

$$P(Z \geq z \mid A_{mk}) \approx 1 - \Phi(\tilde{w}_2) - \phi(\tilde{w}_2)\left(\frac{1}{\tilde{w}_2} - \frac{1}{\tilde{u}_2}\right), \tag{14}$$

where Φ and ϕ are the distribution and density function of the $N(0,1)$ distribution and

$$\begin{aligned}\tilde{w}_2 &= \operatorname{sgn}\left(\tilde{t}\right)\sqrt{2\left[\left\{\left(K_X\left(\hat{r}_0\right) + K_Y\left(\hat{s}_0\right)\right) - \left(m\hat{r}_0 + k\hat{s}_0\right)\right\} - \left\{K\left(\tilde{r}, \tilde{s}, \tilde{t}\right) - \left(m\tilde{r} + k\tilde{r} + (z - 0.5)\,\tilde{t}\right)\right\}\right]} \\ \tilde{u}_2 &= 2\sinh\left(\tilde{t}/2\right)\sqrt{\left|K''\left(\tilde{r}, \tilde{s}, \tilde{t}\right)\right| / \left(K''_X\left(\hat{r}_0\right) K''_Y\left(\hat{s}_0\right)\right)},\end{aligned}$$

where $|K''|$ is the determinant of the 3×3 Hessian matrix K'', $(\hat{r}_0, \hat{s}_0)$ are defined by (12) and $\left(\tilde{r}, \tilde{s}, \tilde{t}\right)$ through (13). Note that for numerical reasons the term $1 - \Phi(\tilde{w}_2)$ in (14) should be computed as $\Phi(-\tilde{w}_2)$ for $\tilde{w}_2 > 0$.

For the two-sided test we need to find the values l for which the inequality (3) holds. Here we make the simplifying assumption that the pmf is monotone as we move in both directions away from its mode. Hence, assuming that the observed z is larger than the mode (the case where z is less than the mode is handled analogously), we only need to find the point z_0 defined as

$$z_0 = \max\left\{l < z \,:\, P(Z = l \mid A_{mk}) \leq P(Z = z \mid A_{mk})\right\}. \tag{15}$$

The 2-sided p-value (4) is then approximated by $P(Z \geq z) + P(Z \leq z_0)$ where the two terms in the sum can be computed from (14).

Finding z_0 can be done using a binary search where for each considered value l we approximate $P(Z = l \mid A_{mk})$ using the saddlepoint approximation of the pmf given by (4.7) in [2]:

$$\begin{aligned}P(Z = l \mid A_{mk}) \approx \frac{1}{\sqrt{2\pi}}&\left\{\frac{\left|K''\left(\hat{r}, \hat{s}, \hat{t}\right)\right|}{K''_X\left(\hat{r}_0\right) K''_Y\left(\hat{s}_0\right)}\right\}^{-1/2} \\ &\times \exp\left[\left\{K\left(\hat{r}, \hat{s}, \hat{t}\right) - \left(m\hat{r} + k\hat{s} + l\hat{t}\right)\right\} - \left\{\left(K_X\left(\hat{r}_0\right) + K_Y\left(\hat{s}_0\right)\right) - \left(m\hat{r}_0 + k\hat{s}_0\right)\right\}\right],\end{aligned}$$

where $(\hat{r}_0, \hat{s}_0)$ are defined by (12) and $\left(\hat{r}, \hat{s}, \hat{t}\right)$ is the solution of (13) with the RHS replaced by (m, k, l).

[7] In fact they coincide with K, K', and K'' where $t = 0$.