When enough is enough: early stopping of biometrics error rate testing

Michael E. Schuckers
Department of Mathematics,
Computer Science and Statistics
St. Lawrence University and
Center for Identification Technology Research (CITeR)
schuckers@stlawu.edu

Emily H. Sheldon
Department of Mathematics,
Computer Science and Statistics
St. Lawrence University

Hilary A. Hartson
Department of Statistics
North Carolina State University

Abstract

Testing of biometric devices to determine whether or not their error rates meet a specified threshold is often an important criterion for determining whether or not a device will be deployed. One consideration for this is how much testing is to be completed. In this paper we apply sequential testing methodology to testing whether or not a biometric device has an error rate below a certain bound. This approach has the advantage that the testing can be stopped at any time. A decision can be made once there is enough evidence to support or reject the hypothesis of interest. Application is made to biometric error rate data for three different modalities - facial recognition, hand geometry and fingerprints. Our results show that this approach can produce more efficient and less costly tests of biometric devices.

1. Introduction

One of the most common performance metrics for biometric devices (BD's) is their matching performance. This is usually expressed as false match and false non-match rates: FMR's and FNMR's respectively. These are similar though not identical to false accept and false reject rates: FAR's and FRR's respectively. See Mansfield and Wayman [5] for an explanation of the differences between these two sets of statistical summaries. Another commonly used summary for these devices is the equal error rate or EER. Testing a BD to determine if the observed error rates meet certain bounds on those rates is a common reason for device testing. The traditional approach to testing is that a sample size is determined *a priori* and testing is done until that sample size is met. In this paper we present alternative to this that allows for the cessation of testing when there is enough evidence to accurately do so. This is a common statistical problem in many areas including clinical trials. See, e.g. [4]. The general approach we will take falls under a group of statistical methods known as sequential analysis. The goal of this paper is to describe this methodology for making the testing of these device more efficient by allowing the possibility of stopping the testing early.

For the statistical evaluation of BD's and their error rates, the primary focus of the work in this area has been on confidence intervals and in some cases on sample size calculations. Schuckers [10] proposed a confidence interval based upon the Beta-binomial distribution. This work on confidence intervals was extended by the same author in Schuckers [9] to include two methods that did *not* depend on the Beta-binomial distribution. That work also included sample size calculations. Several approximate methods have been proposed for sample size calculations. "Doddington's Rule" [2] states that one should collect data until there are 30 errors. Likewise the "Rule of 3" [5] is that $3/(\text{the number of attempts})$ is an appropriate upper bound for a 95% CI for the overall error rate when zero errors are observed. However, both of these methods make use of the binomial distribution which is often an unacceptable choice for biometric data [12] because of the assumption of independence between decisions. Recently, Dass *et al* [1] proposed a simulation-based approximate method for calculating sample sizes for a confidence region around errors rates on a receiver operating characteristic (ROC) curve. This work is important in that it, like Schuckers [9], implicitly

utilizes a correlation structure. Thus, both of these methods imply that the assumptions of the binomial distribution - specifically the assumption of independent trials - are not appropriate for data from a BD. This is a direct consequence of making multiple decisions from a single comparison pair. In this paper we will build the work of Schuckers [9] because of the explicit nature of the sample size calculations given there. To that end we will extend those calculations to include statistical power calculations maintaining the correlation structure between decisions(match/no match). It will be these power calculations upon which we want our sequential methods to improve.

This article is organized in the following manner. Section 2 lays out the process of testing and the mathematical notation that will be used throughout. Sequential methodology for more efficient testing and the sequential likelihood ratio test are discussed in Section 3. The next section, Section 4, discussion our application of this approach to non-synthetic biometric decision data. Our conclusions and a discussion of future work in this area can be found in Section 5.

2. Testing Framework and Notation

Since our focus in this paper is error rate estimation, we focus on errors in the matching process. To that end, we begin by establishing the testing framework and the accompanying notation. We will begin by assuming that we are testing an error rate δ and we would like to test whether or not we can conclude that $\delta < \delta_0$ for some bound on the error rate γ_0. For example, one possibility would be that we want to certify a BD to have an error rate below $\delta_0 = 0.01$. Formally we are doing a statistical hypothesis test with

$$\begin{aligned} H_0 &: \delta \geq \delta_0 \\ H_1 &: \delta < \delta_0. \end{aligned} \tag{1}$$

We then decide whether to reject H_0 or to fail to reject H_0 in traditional hypothesis testing. It is important to remind the reader of the two relevant quantities: the level of a hypothesis test, α, and the power of a hypothesis test, $1-\beta$. See [8] for more details. The level of a test, α, represents the probability of a Type I error which is $\alpha = Pr(\text{Reject } H_0 \mid H_0 \text{ is true})$. β is then the probability of a Type II error $\beta = Pr(\text{Fail to reject } H_0 \mid H_0 \text{ is false})$ and the power of a hypothesis test is $1-\beta$ which is $Pr(\text{Reject } H_0 \mid H_0 \text{ is false})$ Note that these errors differ, in this context from matching errors - false matches and false non-matches.

To estimate δ and subsequently make inference regarding δ, we need to define the decision data that is involved in its estimation. Assume first that through time t there are n_t comparison pairs to be tested. Since each matching decision is based upon a comparison of images we will treat each pair separately and build our correlation model around these pairs. A comparison pair is a collection of $m_i^{(t)}$ comparisons collected from individuals where $i = 1, 2, \ldots, n_t$ through time t. These comparisons can be either decision data or dichotomized match score data. See Ma *et al* [6] for a detailed discussion of data categories for biometrics data. For FMR data, the individuals in each comparison pair are different; for FNMR data, the individuals are the same. Let

$$Y_{ij} = \begin{cases} 1 & \text{if the } j^{th} \text{ decision from} \\ & \text{the } i^{th} \text{ comparison pair is an error,} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

We will refer to the Y_{ij}'s as decision data [6]. Let $\mathbf{Y}^{(t)} = (Y_{11}, \ldots, Y_{1m_1^{(t)}}, \ldots, Y_{n_t 1}, \ldots, Y_{n_t m_n^{(t)}})^T$ be the collection of all observed decisions through time t. Let $X_i^{(t)} = \sum_{i=1}^{m_i^{(t)}} Y_{ij}$ represent the total number of errors for the i^{th} individual through time t. We propose the following explicit correlation structure for the decision data $\mathbf{Y}^{(t)}$. That structure is

$$Corr(Y_{ij}, Y_{i'j'}) = \begin{cases} 1 & \text{if} \quad i = i', j = j' \\ \rho & \text{if} \quad i = i', j \neq j' \\ 0 & otherwise. \end{cases} \tag{3}$$

Here ρ represents the correlation between different decisions made on the same comparison pair. We will refer to it as an intra-comparison correlation. This is the same quantity that Schuckers [9] uses to model correlation; we have just made the correlation structure for that model explicit. Thus, we assume that decisions from the same comparison pair are correlated but that decisions from different comparison pairs are uncorrelated. For the sequential testing we describe below, we will consider the data observed through time t. Thus, n_t will be the total number of observed comparison pairs through time t and $\mathbf{Y}_t$ will be collection of observed decisions through time t.

3. Sequential Methodology

Sequential statistical methods were first developed during World War I by Abraham Wald [11]. There have been numerous advances subsequently in this area. (The interested reader is directed to Ghosh and Sen [3] for more recent work.) Sequential methodology differs from the traditional statistical hypothesis testing framework to allow for three possible conclusions at any point in data collection rather than the traditional two conclusion (reject/fail to reject) format. Under sequential methodology at time t, conclusions are made to either

1. Accept H_0,
2. Accept H_1,

3. Continue testing.

The first two outcomes indicate that enough data has been collected to make a statistically sound decision. This last outcome is an indication that not enough evidence has accumulated to produce an informed decision. The basis for these decisions is the likelihood function.

The sequential probability ratio test (SPRT) approach that we will use is based on a simple vs. simple hypothesis test. For example, for testing a parameter δ we would test $H_0 : \delta = \delta_0$ vs. $H_1 : \delta = \delta_1$ where $\delta_1 < \delta_0$ is chosen to be an acceptable value for the conclusion you would like to draw. The goal here is to reject H_0 if the BD meets the requirement of a sufficiently small error rate. In the case of testing for error rates, we would choose $\delta_1 < \delta_0$ at a value that is acceptable for the purposes of testing. For example, in biometric testing we might choose a value of $\delta_1 = 0.025 < \delta_0 = 0.05$. By selecting $\delta_1 = 0.025$ we are saying that an error rate of 0.025 would be acceptably below our bound of $\delta = 0.05$ so that we could confidently conclude that the error rate is below our bound. Having selected δ_0, δ_1, we must also choose values for the level of the test, α, and the power of the test, $1 - \beta$. The test at time t is then

$$LR_t = \frac{L(\delta_0, \hat{\rho}_0 \mid \mathbf{Y}_t)}{L(\delta_1, \hat{\rho}_1 \mid \mathbf{Y}_t)} \tag{4}$$

where and $L(\delta, \rho \mid \mathbf{Y}_t)$ is the likelihood function and $\hat{\rho}$ is the maximum likelihood estimate based on $\hat{\rho}_i = \arg\max_\rho L(\delta_i, \rho)$. For decision data from BD's, we will follow Schuckers [9] who showed that the Beta-binomial distribution was a reasonable fit to both FMR and FNMR decision data. The likelihood is then

$$\begin{aligned} L(\delta, \rho \mid \mathbf{Y}_t) &= \prod_{i=1}^{n} \left\{ \binom{m_i^{(t)}}{X_i^{(t)}} \frac{\Gamma((1-\rho)\rho^{-1})}{\Gamma(\delta(1-\rho)\rho^{-1})} \right. \\ &\times \frac{\Gamma(\delta(1-\rho)\rho^{-1} + X_i^{(t)})}{\Gamma((1-\delta)(1-\rho)\rho^{-1})} \\ &\times \left. \frac{\Gamma((1-\delta)(1-\rho)\rho^{-1} + m_i^{(t)} - X_i^{(t)})}{\Gamma((1-\rho)\rho^{-1} + m_i^{(t)})} \right\} \end{aligned}$$

Under the logic of the SPRT, we will reject if LR is small enough and we will accept if LR is large enough. Following Wald [11], at time t the following decisions are made:

1. Accept H_0 if $LR > B$,
2. Accept H_1 if $LR < A$, and
3. Continue collecting data if $A < LR < B$

where $A = (1 - \beta)/\alpha$ and $B = \beta/(1 - \alpha)$ with $P(\text{Accept } H_0 \mid \delta = \delta_1) = \beta$ and $P(\text{Reject } H_0 \mid \delta = \delta_0) = \alpha$. A and B are derived from Wald's original formulation of the SPRT. Alternatively since the natural logarithm is a monotone function, we can decide based on $ln(A)$ and $ln(B)$ against

$$ln(A) < \ell(\gamma_0, \hat{\rho}) - \ell(\gamma_1, \hat{\rho}) < ln(B). \tag{5}$$

where $\ell(\gamma, \rho) = ln(L(\gamma, \rho))$. Since our focus is on the error rate, δ, we will consider ρ a nuisance variable. As stated above the motivation for this work is to expedite decision making regarding a BD's error rate. In the next section we evaluate whether or not the proposed procedure meets these claims.

4. Application of Sequential Methods to BD data

In this section we apply the above methods to BD data, both synthetic and real. We will begin by applying this methodology to data from three different biometric modalities. This data was collected by Ross and Jain [7] and includes data from facial recognition, hand geometry and fingerprint BD's. We will consider both FNMR and FMR sequential testing for this data. For all of these modalities we will treat the data as collected over time as one would do in a sequential testing situation.

For their paper on multibiometrics Ross and Jain [7] collected data from 50 individuals from three biometric modalities – face, fingerprint and hand geometry – and recorded the match scores for ten comparisons for each within individual comparison pair of 50 individuals and for five comparisons for each between individual comparison pair composed of crosscomparing those same 50 individuals. Note that the between individual cross comparisons here are not symmetric and thus there were $49 \times 50 = 2450$ comparison pairs in the sense we are using here. Thus there are 500 decisions to compare an individual to themselves and 12250 decisions comparing one individual to another. To simulate sequential testing at each time t we sampled with replacement from among the comparison pairs. Data for FMR and FNMR were generated and tested separately. We continued this process until the SPRT decision rule indicated that enough information had been gathered to make a decision.

To evaluate how well the sequential decision making methodology performed, we considered several hypothesis tests for each modality and each type of error: false match and false non-match. We found thresholds for each device that yielded error rates – both FMR and FNMR – close to the traditional values of 0.10, 0.05 and 0.01. We then considered SPRT's around this data. In particular, we were interested in the Type I and Type II error rates that the SPRT achieved compared to the nominal levels of $\alpha = 0.05$ and $\beta = 0.10$. The hypothesized alternative for δ was chosen to be $\delta_1 = \frac{1}{2}\delta_0$. For each combination of parameters we

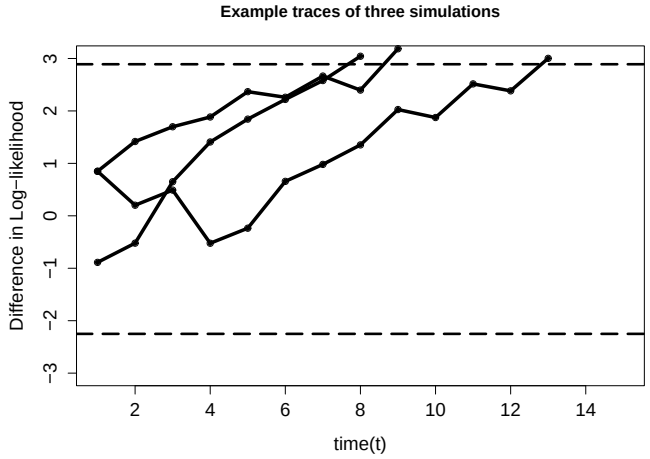

Figure 1.

sampled (with replacement) a comparison pair and the corresponding decisions until there was enough information to stop the test. That is, after each 'new' set of data is added we reran the SPRT to determine if stopping was appropriate based on the test. We recorded the number of comparison pairs added until the test was complete and we will refer to this as the stopping times. For each hypothesis test and each modality we replicated the SPRT 1000 times, each time generating different data $\mathbf{Y}_t$, to gain an understanding of the distribution of stopping times for each test. Figure 1 illustrates graphically this process for three separate simulations. The dashed horizontal lines represent the boundaries at which the SPRT will conclude that enough data has been collected to make a statistically appropriate conclusion.

Tables 1, 2 and 3 contain the results of the simulations. In these tables, the 'true error rate' is the error rate for the full data from each modality and 'ρ' is the intra-class correlation for that same data. n^* is the fixed sample size power calculation for determining the number of comparison pairs to sample with a one-sided alternative hypothesis. That calculation is

$$\begin{aligned} n^* &= \lceil m^{-1}(\delta_0 - \delta_1)^{-2} \\ &\times \Big(z_{1-\alpha}\sqrt{\delta_0(1-\delta_0)(1+\rho(m-1))} \\ &+ z_{1-\beta}\sqrt{\delta_1(1-\delta_1)(1+\rho(m-1))}\Big)^2\Big\rceil \end{aligned} \quad (6)$$

where $\lceil \bullet \rceil$ represent the 'ceiling' function and z_k is the k^{th} percentile of a standard normal distribution. We use the estimates of δ and ρ calculated from the data for calculating n^*. n_p in the tables of results represents the $100 * p^{th}$ percentile from the distribution of stopping times.

Overall these tables show that the SPRT performs extremely well. When the correct decision is H_0, we should expect an error rate of $\alpha = 5\%$. With one exception, the '% incorrect decisions' are within 2% of the nominal level. In that case, the percentage was much lower than would be anticipated. We should expect the '% incorrect decisions' when the correct decision is to H_1 to be $\beta = 10\%$. The '% incorrect decisions' for these decisions exhibits far more variability than the Type I error rate, α, simulations. Most of these percentages are either around the nominal level, 10%, or below. Only one of these error rates is significantly higher than that value: FNMR for Hand Geometry, $H_0 : \delta = 0.200$. Moving to the stopping times, the percentiles of the distribution for each combination of parameters suggests that the SPRT does better than the fixed sample size, n^*. The fixed sample size falls above the 75^{th} percentile for all but one of the simulations. Thus, we can conclude that with probability greater than 75% the SPRT will result in fewer comparison pairs collected than the traditional testing approach. In most cases the fixed sample size, n^* falls between the 75^{th} and 97.5^{th} percentiles for the distribution of stopping times. Because of the stochastic nature of the sequential approach, it is possible that the stopping time will be larger than the fixed sample size. If we compare n^* to the median of the stopping times, $n_{0.50}$ then we see that median savings due to using the SPRT is significant. On average, the ratio of median stopping times to fixed sample size is approximately 0.495. Thus the typical savings from using the SPRT rather than the traditional sample size approach is 50%. This is a significant amount given the cost of testing of a device.

5. Conclusions

In this paper we have presented an alternative method for testing whether or not a BD meets a specific bound for its error rate. This approach is based upon the sequential probability ratio test developed by Wald [11]. Traditionally biometric testing has been done by fixing the number of individuals to be tested and collecting data until that number has been reached. The sequential alternative allows a determination about the fitness of a device's matching performance *early*, while data is being collected. This can allow testing to be stopped earlier and has the potential to result in larger savings both in time and money. In particular, we find that for the three datasets that we observed in this paper, the savings one could expect from this are in the neighborhood of 25-50%. This is a significant improvement over the fixed sample size approach. These results are consistent across all three of the modalities tested here. To confirm and extend on these results we need to further study several issues. First, the impact of choosing different values for δ_1. In this study we limited ourselves to $\delta_1 = \frac{1}{2}\delta_0$. Other values need to be considered to assess the robustness of the results here. Next, we want to explore other values for α and β to look at their impact sequential testing. Another area of interest is other error rates beyond the $0.10, 0.05$ and 0.01 tested here. In particular, there is interest in being able to test values of δ that are orders of magnitude smaller than those tested here. Finally, we would like to consider a hybrid testing approach that would involve sequential testing until the fixed sample size is achieved. This would guarantee a worst case sample size but it will also likely have implications for the Type I and Type II error rates. These issues are worthy of further study. We are confident that the changes to application will not alter the overall conclusions we have described above. It is possible to improve biometric testing by stopping the testing when enough (data) is enough (to make valid inference).

Acknowledgements

Generous funding for this work comes from a grant from the Center for Identification Technology Research. Additional funding for this work comes from NSF grants CNS-0325640 and CNS-0520990. CNS-0325640 is cooperatively funded by the National Science Foundation and the United States Department of Homeland Security. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Center for Identification Technology Research.

References

[1] S. C. Dass, Y. Zhu, and A. K. Jain. Validating a biometric authentication system: Sample size requirements. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):19–30, 2007.

[2] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds. The NIST speaker recognition evaluation: overview methodology, systems, results, perspective. *Speech Communication*, 31(2-3):225–254, 2000.

[3] B. K. Ghosh and P. K. Sen, editors. *Handbook of Sequential Analysis*. Marcel Dekker, Inc, 1991.

[4] C. Jennison and B. W. Turnbull. *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC, 1999.

[5] T. Mansfield and J. L. Wayman. Best practices in testing and reporting performance of biometric devices. on the web at www.cesg.gov.uk/site/ast/biometrics/media/BestPractice.pdf, 2002.

[6] Y. Ma, M. Schuckers, and B. Cukic. Guidelines for appropriate use of simulated data for bio-authentication research. In V. Bhagavatula and V. Govindaraju, editors, *Proceedings of the Fourth IEEE Workshop on Automatic Identification Advanced Technologies*, pages 251–256, 2005.

[7] A. Ross and A. K. Jain. Information fusion in biometrics. *Pattern Recognition Letters*, 24(13):2115–2125, September 2003.

[8] M. J. Schervish. *Theory of Statistics*. Springer-Verlag, New York, 1995.

[9] M. E. Schuckers. Estimation and sample size calculations for correlated binary error rates of biometric identification rates. In *Proceedings of the American Statistical Association: Biometrics Section [CD-ROM]*, Alexandria, VA, 2003. American Statistical Association.

[10] M. E. Schuckers. Using the beta-binomial distribution to assess performance of a biometric identification device. *International Journal of Image and Graphics*, 3(3):523–529, 2003.

[11] A. Wald. *Sequential Analysis*. Dover Publications, 1947.

[12] J. L. Wayman. Confidence interval and test size estimation for biometric data. In *Proceedings of IEEE AutoID '99*, pages 177–184, 1999.

Table 1. Facial Recognition Sequential Testing of $H_0 : \delta = \delta_0$ vs $H_1 : \delta < \delta_0$, $\alpha = 0.05, \beta = 0.10$

Error Type	δ_0	δ_1	True error rate	ρ	Correct decision	n^*	$n_{0.50}$	$n_{0.75}$	$n_{0.975}$	% incorrect decisions
FMR	0.100	0.050	0.0984	0.0000	H_0	48	18	31	81	0.042
FMR	0.050	0.025	0.0510	0.0000	H_0	100	32	57	144	0.030
FMR	0.010	0.005	0.0098	0.0000	H_0	517	186	321	742	0.049
FMR	0.200	0.100	0.0984	0.0000	H_1	22	16	23	44	0.065
FMR	0.100	0.050	0.0510	0.0000	H_1	48	31	43	92	0.108
FMR	0.020	0.010	0.0098	0.0000	H_1	257	128	197	394	0.068
FNMR	0.100	0.050	0.1000	0.0000	H_0	24	10	18	39	0.030
FNMR	0.050	0.025	0.0500	0.0290	H_0	63	17	31	78	0.029
FNMR	0.010	0.005	0.0100	0.0000	H_0	259	90	159	371	0.044
FNMR	0.200	0.100	0.1000	0.0000	H_1	11	11	14	26	0.039
FNMR	0.100	0.050	0.0500	0.0290	H_1	31	16	24	47	0.095
FNMR	0.020	0.010	0.0100	0.0000	H_1	129	72	106	210	0.072

Table 2. Hand Geometry Sequential Testing of $H_0 : \delta = \delta_0$ vs $H_1 : \delta < \delta_0$, $\alpha = 0.05, \beta = 0.10$

Error Type	δ_0	δ_1	True error rate	ρ	Correct decision	n^*	$n_{0.50}$	$n_{0.75}$	$n_{0.975}$	% incorrect decisions
FMR	0.100	0.050	0.1025	0.0321	H_0	54	17	29	75	0.031
FMR	0.050	0.025	0.0504	0.0091	H_0	104	32	55	140	0.035
FMR	0.010	0.005	0.0099	0.0000	H_0	517	175	315	768	0.061
FMR	0.200	0.100	0.1025	0.0321	H_1	25	16	22	45	0.157
FMR	0.100	0.050	0.0504	0.0091	H_1	50	29	43	84	0.092
FMR	0.020	0.010	0.0099	0.0000	H_1	257	128	197	405	0.067
FNMR	0.100	0.050	0.1020	0.0514	H_0	35	10	17	38	0.052
FNMR	0.050	0.025	0.0500	0.0222	H_0	60	17	33	91	0.049
FNMR	0.010	0.005	0.0100	0.0000	H_0	259	89	149	425	0.035
FNMR	0.200	0.100	0.1020	0.0514	H_1	16	11	15	31	0.125
FNMR	0.100	0.050	0.0500	0.0222	H_1	29	18	26	61	0.083
FNMR	0.020	0.010	0.0100	0.0000	H_1	129	72	106	231	0.064

Table 3. Fingerprint Sequential Testing of $H_0 : \delta = \delta_0$ vs $H_1 : \delta < \delta_0$, $\alpha = 0.05, \beta = 0.10$

Error Type	δ_0	δ_1	True error rate	ρ	Correct decision	n^*	$n_{0.50}$	$n_{0.75}$	$n_{0.975}$	% incorrect decisions
FMR	0.100	0.050	0.0991	0.0142	H_0	51	17	30	76	0.032
FMR	0.050	0.025	0.0496	0.0000	H_0	100	35	62	160	0.042
FMR	0.010	0.005	0.0102	0.0000	H_0	517	176	309	793	0.036
FMR	0.200	0.100	0.0991	0.0142	H_1	23	16	22	44	0.096
FMR	0.100	0.050	0.0496	0.0000	H_1	48	29	41	92	0.079
FMR	0.020	0.010	0.0102	0.0000	H_1	257	142	211	433	0.077
FNMR	0.100	0.050	0.1000	0.0000	H_0	24	11	19	39	0.016
FNMR	0.050	0.025	0.0520	0.0000	H_0	50	17	30	69	0.031
FNMR	0.010	0.005	0.0100	0.0000	H_0	259	90	162	372	0.038
FNMR	0.200	0.100	0.1000	0.0000	H_1	11	11	14	24	0.011
FNMR	0.100	0.050	0.0520	0.0000	H_1	24	18	25	50	0.082
FNMR	0.020	0.010	0.0100	0.0000	H_1	129	72	106	225	0.058