Towards Sound and Optimal Leakage Detection Procedure

Liwei Zhang, A. Adam Ding, Francois Durvaux, Francois-Xavier Standaert, and Yunsi Fei

[1] Department of Mathematics, Northeastern University, Boston, MA 02115
[2] ICTEAM/ELEN/Crypto Group, Universite catholique de Louvain, Belgium
[3] Department of Electrical and Computer Engineering
Northeastern University, Boston, MA 02115
zhang.liw@husky.neu.edu,a.ding@northeastern.edu,
francois.durvaux@gmail.com,fstandae@uclouvain.be,y.fei@northeastern.edu

Abstract. Evaluation of side channel leakage for the embedded crypto systems requires sound leakage detection procedures. We relate the test vector leakage assessment (TVLA) procedure to the statistical minimum p-value (mini-p) procedure, and propose a sound method of deciding leakage existence in the statistical hypothesis setting. To improve detection, an advanced statistical procedure Higher Criticism (HC) is applied. The detection of leakage existence and the identification of exploitable leakage are separated when there are multiple leakage points. For leakage detection, the HC-based procedure is shown to be optimal in that, for a given number of traces with given length, it detects existence of leakage at the signal level as low as possibly detectable by any statistical procedure. Numerical studies show that the HC-based procedure perform as well as the mini-p based procedure when leakage signals are very sparse, and can improve the leakage detection significantly when there are multiple leakages.

Keywords: Side channel analysis, leakage detection, higher criticism

1 Introduction

Side-channel attacks (SCA) has been shown to be a serious threat for modern cryptographic implements. For more than a decade now, researchers have actively studied the side-channel attacks and proposed countermeasures to protect devices against such attacks. As various countermeasures are integrated into commercial customer devices, evaluating the resistance of devices against SCA becomes an important issue. A *leakage detection* test procedure, Cryptography Research (CRI)'s test vector leakage assessment (TVLA) [1, 2] is generally adopted as the standard methodology for blackbox evaluation of SCA resistance. The TVLA procedure scans the physical measurements leakage trace (e.g., a power trace) with a univariate test, and no leakage is detected if the test statistics fail to exceed a critical value at all points along the leakage trace.

It is preferred to use a generic univariate test in TVLA procedure to avoid relying on a specific leakage model. To achieve this, the test commonly runs on data sampled according to a nonspecific partition, usually the fixed-vs-fixed sampling or the fixed-vs-random sampling, where the fixed class of measurements come from encryptions of a fixed plaintext while the random class of measurements come from encryptions of random plaintexts. The natural test to detect the difference in power consumptions between two classes of such a nonspecific partition is the Welch's t-test [1, 2] in CRI's TVLA proposal. Extensions of the t-test (e.g., higher order and multivariate leakage detection) have been proposed by various researchers [3–6].

Durvaux and Standaert [5] in EuroCrypt2016 proposed a correlation-based test (ρ-test) to detect exploitable leakage aimed at a particular intermediate computation. Their ρ-test runs on data that are partitioned specifically using this targeted intermediate value. The exact leakage model on this intermediate value is avoided by profiling the power traces. Such a specific test yields sparser leakage signals relating to this specific targeted intermediate value, and is better suited at identification of Point-Of-Interest (POI) for this exploitable leakage. While this identification is a necessary first step in launching practical side-channel attacks, it is not required for the purpose of leakage detection. For detection of leakage existence, the nonspecific partition methods (such as the fixed-vs-fixed sampling and the fixed-vs-random sampling) generally find more leakage signals and perform well. Both the specific and non-specific leakage detection tests can be used in the TVLA framework.

Here we study the TVLA procedure itself from a theoretical perspective. The TVLA procedure declares that a device is leaky, if the maximum test statistic (over all points on the trace) exceeds a critical value. For the Welch's t-test, current TVLA procedure generally uses the critical value of 4.5 [2, 7–9], which corresponds to a statistical significance level of $\alpha < 0.001$ for the univariate test. However, this significance level does not consider the total number of univariate tests, i.e., the total number of points on the trace. The overall significance level increases as the total number of points on the trace increases. For long traces, the overall significance level can be quite large so that non-leaky devices can not pass TVLA t-testing with the critical value of 4.5. Hence, Balasch et. al [10] suggested raising the critical value to 5 for longer traces based on numerical experiments. However, for even longer traces, the non-leaky devices still can not pass at this higher critical value of 5 (see Section 3.1). Therefore, the current TVLA procedure needs a more rigorous way of setting the criterion.

Based on this state-of-the-art, we make two contributions in this paper as following. We first propose a sound method to set the critical value according to an overall statistical significance level, completing the current TVLA procedure. To make the decision (leaky versus non-leaky) based on the largest test statistic is equivalent to decide using the minimal p-value of all those univariate tests. Hence the current TVLA procedure is a statistical minimum p-value (mini-p) procedure. The threshold can be set through the mini-p procedure at any given statistical significance level, taking account of the trace length. For the t-test

based TVLA, we provide explicit expression of this threshold which also varies with the number of traces (used as the degree of freedoms in the test).

Secondly, we propose to improve the (univariate) leakage detection procedure with a statistically optimal HC metric. The mini-p procedure focuses on the POI with the largest test statistic, and decides existence of leakage along the whole trace based on this selected single POI only. It does not sufficiently use the information provided by test statistics at all other points on the trace. For the leakage detection purpose, the evaluator searches for evidence of key-dependent leakages along the trace, without necessarily needing to identify the POIs exactly. Hence it is very similar as the statistical independence scanning procedure [11–17] which has been widely used in other high-dimensional statistical application. In terms of the signal strength and signal sparsity, there is an *undetectable region* [18] where no statistical test can discern the existence of leakage. An optimal leakage procedure should be able to detect any leakage outside this minimal theoretical undetectable region. The current TVLA (mini-p) procedure is not optimal , as its undetectable region is larger than the minimal theoretical undetectable region for all statistical tests. We introduce the "Higher Criticism" (HC), a state-of-art statistical method for detecting sparse and weak signals, to the TVLA procedure. The HC-based procedure is shown to be an optimal leakage detection procedure.

Our work improves the TVLA procedure to optimally utilize the multiple leakages for detection. This is independent of whether the univariate test itself is optimal. Both specific and nonspecific leakage detection tests above can be used, with their relative advantages and limitations [5] still apply. Our work is also orthogonal to the work of combining multiple leakages for a single attack, e.g., [19–24].

Our proposed procedure optimally combines the *detections* of univariate leakage existence at all points along the leakage trace. It works as good as the mini-p for very sparse leakage signals, and significantly improves the detection in scenarios where there are multiple leakage signals.

2 Background and Model Notations

2.1 TVLA procedure as a mini-p testing framework

In the TVLA leakage detection setup, an evaluator collects many traces of physical measurements, and wants to find if some points on the traces leaking key information through some key-sensitive intermediate values. Let n_{tr} and n_L denote, respectively, the total number of traces and the total number of points on each trace. That is, the evaluator has n_{tr} realizations of the random vector $\boldsymbol{L} = [L_1, \cdots, L_{n_L}]$. The scanning procedure such as TVLA do a univariate statistical test at each time points, and make the decision by combining the results. That is, we test the null hypothesis

$$L_i = r_i \tag{1}$$

versus the alternative hypothesis

$$L_i = V + r_i \tag{2}$$

at the i-th time point, where V is a key-sensitive intermediate variable and r_i is random noise.

The test is usually done with a test statistic $\widehat{\boldsymbol{s}}_i$. Statistically the p-value is the probability that test statistic value can be observed by chance under null hypothesis, i.e., $\mathrm{P}(|S| \geq |\widehat{\boldsymbol{s}}_i|)$ where S denotes a random variable which follows the distribution of the test statistic under the null hypothesis (measurement at the i-th time point is pure random noise). For a single hypothesis test, the null hypothesis is rejected when $|\widehat{\boldsymbol{s}}_i|$ is too big or equivalently when the p-value is too small.

The TVLA procedure decides that leakage exists as long as any one of the tests rejects the null hypothesis. That is, the device is considered leaky when $\max_{1 \leq i \leq n_L} |\widehat{\boldsymbol{s}}_i| \geq \mathrm{TH}$ for a threshold value TH (or equivalently when the minimum p-value is smaller than a threshold value α_{TH}). Therefore, the current TVLA procedure is in fact a mini-p test framework for combining multiple (n_L) testing which utilizes the minimal p-value only. We will propose changing this mini-p multiple testing framework later.

While the usage of a particular univariate test is not essential to the framework, we first do describe two common univariate tests to use as concrete examples that fit in the overall framework.

2.2 Univariate Tests: ρ-test, t-test, Specific versus Nonspecific Tests

Given the leakage model (2) with the known intermediate value V, the nature attack is the correlation power analysis (CPA) distinguisher. The CPA uses the correlation, which can be also be used for leakage detection in *ρ-test*. The correlation is

$$\hat{\rho}_i = \mathrm{Corr}(L_i, V). \tag{3}$$

The test statistic is taken as the Fisher's transformation on $\hat{\rho}_i$ scaled by $\sqrt{n_{tr}}$:

$$\widehat{\boldsymbol{s}}_i = \frac{1}{2} \ln \left(\frac{1 + \hat{\rho}_i}{1 - \hat{\rho}_i} \right) \sqrt{n_{tr}}. \tag{4}$$

Under the null hypothesis (no leakage at the i-th time point), $\widehat{\boldsymbol{s}}_i$ approximately follows the standard normal distribution. So the corresponding p-value is calculated by:

$$p_i = 2 \times (1 - \mathrm{CDF}_{N(0,1)}(|\widehat{\boldsymbol{s}}_i|)), \tag{5}$$

where $\mathrm{CDF}_{N(0,1)}(\cdot)$ is the cumulative distribution function of the standard normal distribution.

The ρ-test can be considered as an ideal test for perfectly modeled power leakage, often Hamming Weight or Hamming Distance of a Sbox output. A more generic version of ρ-test is proposed by Durvaux and Standaert [5] where

they profiled the leakage on the targeted V, thus allowing implementation in a blackbox manner.

Another common generic test is the Welch's t-test [1, 2], where the $\boldsymbol{L}_i$ measurements are partitioned into two sets $L_{i,A}$ and $L_{i,B}$, and compared by the test statistic

$$\widehat{\boldsymbol{s}}_i = \frac{\overline{L}_{i,A} - \overline{L}_{i,B}}{\sqrt{\frac{\hat{\nu}_{i,A}^2}{n_A} + \frac{\hat{\nu}_{i,B}^2}{n_B}}}, \tag{6}$$

where $\overline{L}_{i,A}$ and $\overline{L}_{i,B}$ denote the sample means (average values) in each set, $\hat{\nu}_{i,A}$ and $\hat{\nu}_{i,B}$ denote the sample standard deviations, n_A and n_B denote the numbers of measurements for the set A and B, respectively. The corresponding p-value is calculated as the probability, under a t-distribution with ν_t degree of freedom, that the random variable exceeds the observed statistic value $\widehat{\boldsymbol{s}}_i$:

$$p_i = 2 \times (1 - \mathrm{CDF}_t(\widehat{\boldsymbol{s}}_i, \hat{\nu}_i)), \qquad \tau = 1, \cdots, n_L, \tag{7}$$

where $\mathrm{CDF}_t(\cdot, \hat{\nu}_i)$ is the cumulative distribution function of t-distribution with the degree of freedom

$$\hat{\nu}_i = (\hat{\nu}_{i,A}^2/n_A + \hat{\nu}_{i,B}^2/n_B)^2 / [(\hat{\nu}_{i,A}^2/n_A)^2/(n_A - 1) + (\hat{\nu}_{i,B}^2/n_B)^2/(n_B - 1)].$$

In practice, the degree of freedom $\hat{\nu}_i$ may be big so that the $\mathrm{CDF}_t(\cdot, \hat{\nu}_i)$ can be approximated by $\mathrm{CDF}_{N(0,1)}(\cdot)$. In that case, the p-value for t-test can also be calculated from (5).

Recall that [5] used the ρ-test as a specific test on data partitioned according to the specific intermediate value. The t-test is naturally used on data with two classes with nonspecific partition (fixed-vs-fixed and fixed-vs-random). The data collection methods, specific versus nonspecific, affect how sparse and how strong the leakage signals are in the data. Those are the two critical factors in the theoretical analysis in Section 4.

Notice that the ρ-test statistic can also be used on the nonspecific data partition. In such cases, there could be a difference in the test SNR $|\,\mathrm{E}(\widehat{\boldsymbol{s}}_i)|^2/\,\mathrm{Var}(\widehat{\boldsymbol{s}}_i)$ for the ρ-test statistic and the t-test statistic. But generally the difference in test SNRs due to the choice of test statistics (ρ versus t) is relative small, compared to the differences in model SNRs between data collected by the nonspecific partition (fixed-vs-fixed and fixed-vs-random) versus specific partition(random plaintexts for [5]'s ρ-test). For the leakage model (2), the model SNR is $Var(V)/Var(r)$. The relationships between the test SNR and the model SNR will be quantified in the theory section 4. Generally, we can analyze the multiple testing procedure given the test SNRs without specifying which particular univariate test is used.

3 Methodology

In this section, we first discuss how to set the threshold for mini-p procedure correctly. We then describe the higher criticism (HC) procedure, and roughly compare it with mini-p procedure. In Section 4, we will theoretically show that HC is an optimal leakage detection method.

3.1 Threshold Setting in Mini-p Procedure

The current TVLA procedure declares a device as leaky when $\max_{1\le i\le n_L} |\widehat{\boldsymbol{s}}_i| \ge$ TH. However, the threshold value TH was not set at a given significance level (Type I error rate) as in usual statistical methods. The t-test threshold of TH = 4.5 is suggested originally as it corresponds to a significance level of < 0.001 for each *univariate test* [1, 2]. However, the overall significance level varies with the number of time points n_L on the trace, so that the procedure is not doing a fair testing for traces with different lengths. Particularly, for a long trace, a leakage free device is often declared as leaky. For this reason, Balasch et al. [10] suggested raising the threshold to TH = 5 for long traces. In Table 1(a), we give the type I error rates under both TH = 4.5 and TH = 5 for the current TVLA procedure. As the number n_L increases, the type I error rate increases. Particularly when $n_L = 1,000,000$, a leakage free device will almost always be declared as leaky (99.9% probability) under the threshold TH = 4.5, and still about 44% chance of being declared as leaky with the higher threshold TH = 5. Either way, we observe that for any such fixed threshold for the test statistic, type I error rate varies greatly for different n_L. Thus a more formal way of setting threshold value according to the trace length is needed, to allow fair evaluation across different trace lengths.

Table 1: T-test threshold and Type I error rates for varying trace lengths n_L.

(a) Type I error rates α under fixed threshold values.

n_L	10^2	10^3	10^4	10^5	10^6
TH = 4.5	0.00068	0.0068	0.0661	0.4957	0.9987
TH = 5	0.000057	0.00057	0.0057	0.0557	0.4363

(b) Threshold values TH under fixed type I error rates.

n_L	10^2	10^3	10^4	10^5	10^6	10^7	10^8
$\alpha = 0.001$	4.417	4.892	5.327	5.731	6.110	6.467	6.806
$\alpha = 0.01$	3.889	4.416	4.891	5.326	5.730	6.109	6.466

Realizing that the current TVLA procedure is in fact a mini-p procedure, the threshold for the minimum p-value should be set as $\alpha_{\mathrm{TH}} = 1 - (1-\alpha)^{1/n_L}$ for an overall significance level α. Then for the t-test, the threshold is TH = $\mathrm{CDF}_t^{-1}(1 - \alpha_{\mathrm{TH}}/2, \nu_s)$ where CDF_t^{-1} is the inverse of CDF of t-distribution. This threshold value depends on the number of traces n_{tr} which affects the degrees of freedom ν_s in the t-distribution. When ν_s is big, this can also be calculated as $\mathrm{CDF}_{N(0,1)}^{-1}(1 - \alpha_{\mathrm{TH}}/2)$. In Table 1(b), we list the cutoff values, for the type I error rate of 0.001 and 0.01 under various trace lengths (assuming ν_s is big).

Next, we propose an improved leakage detection method based on the higher criticism (HC) [11, 12] which utilize the information contained in all n_L test statistics more efficiently.

3.2 Higher Criticism

Statistically, the leakage detection can be formulated as testing an overall hypothesis

$$H_0 : \text{Model (1) holds at all time points } (i = 1, ..n_L), \tag{8}$$

$$H_1 : \text{Model (2) holds at some time points..} \tag{9}$$

The current mini-p procedure makes the overall decision based on the minimal p-value $\min_{1 \le i \le n_L} p_i$ only. This ignores the information on all other p-values except the extreme one. The HC method utilizes the information stored in the distribution of p-values. Under the null hypothesis (8), all observed p-values should follow a uniform distribution on the interval $[0, 1]$. For the time points where leakage exists as equation (2), the expected p-values will be smaller than those generated from the uniform distribution. Hence under the alternative hypothesis (9) of some POIs with leakage (a mixture distribution), the obtained p-values trend to be smaller than those generated under the uniform distribution. Fig. 1 draws two curves of the ordered p-values under these two hypothesises. The figure clearly shows the difference of the distributions of the ordered p-values under H_0 and H_1.

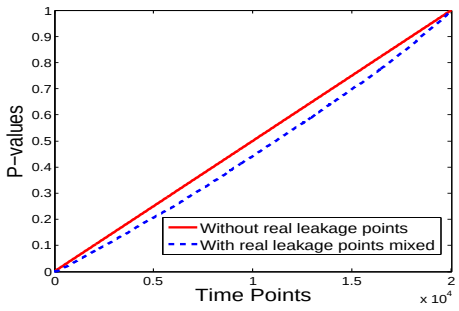


Fig. 1: Comparison of the distributions of ordered p-values under the null hypothesis and under the alternative hypothesis.

The leakage detection problem can now be restated as comparing the distribution of the obtained p-values $\hat{p}_1$, ..., $\hat{p}_{n_L}$ with the uniform distribution, or equivalently as detecting the difference between the two curves in Fig. 1. Naturally, to detect the difference between the two curves, we can compare the ordered p-values $\hat{p}_{(1)} \le \hat{p}_{(2)} \le ... \le \hat{p}_{(n_L)}$ with their expected values $1/n_L, 2/n_L, ..., n_L/n_L$ under the uniform distribution. The HC procedure is based on the normalized distances for these comparisons,

$$\widehat{\mathrm{HC}}_{n_L,i} = \frac{\sqrt{n_L}(i/n_L - \hat{p}_{(i)})}{\sqrt{\hat{p}_{(i)}(1 - \hat{p}_{(i)})}}, \quad i = 1, ...n_L. \tag{10}$$

The HC procedure makes the detection if the maximum of these normalized distance $\widehat{\mathrm{HC}}_{n_L,i}$ exceeds a threshold. In contrast, the mini-p procedure only use

the first distance $\widehat{\mathrm{HC}}_{n_L,1}$ corresponding to the smallest p-value $\hat{p}_{(1)}$ only. That is, the mini-p procedure focused on the difference between the two curves in Fig. 1 at the lower-left corner only. When n_L is big, the maximum normalized distance often does not occur at $\widehat{\mathrm{HC}}_{n_L,1}$. Thus the HC procedure can be more effective in detecting the difference by comparing the whole curves instead of using only the pair of extreme points at the lower-left corner.

Formally, the HC procedure is as follows:

(1) Sort the p-values in ascending order $\hat{p}_{(1)} \leq \hat{p}_{(2)} \leq ... \leq \hat{p}_{(n_L)}$.

(2) Calculate the HC objective function $\widehat{\mathbf{HC}}_{n_L,i}$, $i = 1, ...n_L$, from equation (10).

(3) The HC statistic $\widehat{\mathrm{HC}}_{n_L,max}$ is defined as,

$$\widehat{\mathrm{HC}}_{n_L,max} = \max_{1 \leq i \leq n_L/2} \widehat{\mathrm{HC}}_{n_L,i}. \tag{11}$$

(4) Compare the obtained HC statistic $\widehat{\mathrm{HC}}_{n_L,max}$ with the HC threshold $b^{\mathrm{HC}}_{n_L,\alpha}$ corresponding to some chosen significance level α. When $\widehat{\mathrm{HC}}_{n_L,max} \leq b^{\mathrm{HC}}_{n_L,\alpha}$, we accept the null hypothesis of no leakage. When $\widehat{\mathrm{HC}}_{n_L,max} > b^{\mathrm{HC}}_{n_L,\alpha}$, we reject the null hypothesis and declare that leakage exists.

The HC threshold $b^{\mathrm{HC}}_{n_L,\alpha}$ is set to the $1-\alpha$ quantile of the HC statistic $\widehat{\mathrm{HC}}_{n_L,max}$'s distribution under the null hypothesis. Since each $\widehat{\mathrm{HC}}_{n_L,j}$ asymptotically follows a standard normal distribution $N(0,1)$ under the null hypothesis, this quantile $b^{\mathrm{HC}}_{n_L,\alpha}$ can obtained by simulation from the n_L standard normal random variables. For n_L big, the threshold $b^{\mathrm{HC}}_{n_L,\alpha}$ can be approximated through the connection to Brownian bridge, for example the calculation formula provided in Li and Siegmund [15].

When $n_L \geq 100$, $b^{\mathrm{HC}}_{n_L,\alpha} \approx 10.10$ and 31.65 for $\alpha = 0.01$ and 0.001 respectively. To compare the mini-p procedure and HC procedure, let us assume that the HC threshold is achieved at the max T-statistic (same as mini-p procedure), and translate the HC threshold in terms of the max T-statistic. The thresholds of maximum T-statistics for mini-p and HC procedures are then listed in the following Table 2.

Table 2: Thresholds of maximum t-test statistics for mini-p and HC procedures.

α	n_L	100	1,000	10,000	100,000	1,000,000	10,000,000	100,000,000
0.001	$Tmax_{mini-p}$	4.417	4.892	5.327	5.731	6.110	6.467	6.806
	$Tmax_{HC}$	4.418	4.892	5.327	5.731	6.110	6.468	6.807
0.01	$Tmax_{mini-p}$	3.889	4.416	4.891	5.326	5.730	6.109	6.466
	$Tmax_{HC}$	3.900	4.426	4.899	5.334	5.737	6.116	6.473

In terms of the maximum t-test statistic, we notice that the thresholds for the two procedures are almost the same, with the HC threshold being barely higher. The HC procedure gains more detection power than the mini-p procedure when

$\widehat{\mathrm{HC}}_{n_L,max}$ does not occur at the largest t-test statistic. Particularly for devices with some countermeasures, the remaining hard-to-detect leakage points may not have strong leakage signals. Then the test statistics corresponding to those real leakage points may not become the largest, compared to the test statistics at other noisy points on a long n_L trace. However, they do move the curve in Fig. 1 without becoming the largest one, and these differences can be picked up by the HC procedure but not by the mini-p procedure.

3.3 HC Framework

Here, we summarize the HC detection process step by step. The flow chart in Fig. 2 summarizes the steps in the HC leakage detection procedure, where the steps in the dash-circled box are also executed in the current TVLA procedure.

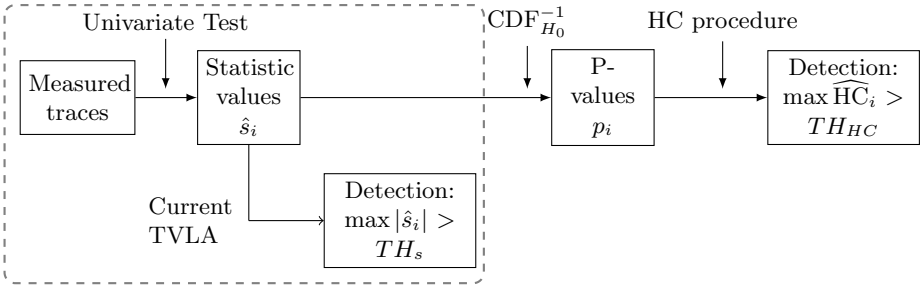


Fig. 2: HC leakage detection flow chart.

(I) The evaluator collects a set of physical measurements, then calculate a selected univariate test (e.g., tests in [3, 5, 4, 6]) at each time point along the measurement traces. Therefore, n_L statistic values are obtained.

(II) The evaluator finds the cumulative distribution function (CDF) of the above statistic under the null hypothesis H_0 (pure noise model). Using the CDF, the n_L statistic values are translated into n_L p-values (which may also be used by mini-p procedure), e.g., as in equations (5) and (7).

(III) Based on the n_L p-values, the HC procedure in Section 3.2 is used to decide if any leakage exists at a given type I error rate α.

For the last step (III), we provide a code to efficiently calculate the thresholds of HC in the Appendix 7.1. The user provides the n_L test p-values and the type I error rate α as inputs, and the code outputs the threshold and the detection results.

The current TVLA does not do step (II) and the threshold is not chosen according to a statistical type I error rate. We have shown that it is equivalent to doing step (II) and then conducting a mini-p procedure, with can be made sound by choosing the threshold as in Section 3.1. The proposed approach would conduct the HC procedure in step (III) instead.

4 Theory

In this section, we theoretically compare the mini-p and HC procedure. When we consider the combination of n_L detection tests using the n_{tr} traces, the results are affected also by two factors: the SNR and the sparsity of real leakage points. The first factor is reflected by the model SNR $\mathrm{Var}(V)/\mathrm{Var}(r)$ in equation (2), or by the related test SNR $|\mathrm{E}(\widehat{\boldsymbol{s}})|^2/\mathrm{Var}(\widehat{\boldsymbol{s}})$. The second factor is reflected by the proportion of POIs following model (2) instead of the pure noise model (1).

Since n_L is usually large, we will consider the high-dimensional statistical asymptotic theory for detectable signals. On the Euclidean space constructed by these two factors, the high-dimensional statistical asymptotic theory separates it into three regions: undetectable regions, detectable but unidentifiable region, and identifiable region. In the undetectable region, the POIs are too sparse and the signal at those POIs are too weak so that no statistical multiple testing procedure may cleanly discern the *existence* of POIs satisfying model (2). In the detectable but unidentifiable region, a statistical multiple testing procedure can detect the existence of POIs but not *identify* their location always. In the identifiable region, a statistical multiple testing procedure can identify all POIs. The POIs identification is important for exploiting the existing side-channel leakage. But for the leakage detection procedure, it is more important to simply detect the existence of any POI. Therefore, we would like to improve the TVLA procedure to have an undetectable region as small as possible.

In this section, we first set up the model notations, representing the sparsity and signal strength by two parameters β and γ. We then discuss, under fixed-vs-fixed and fixed-vs-random setting, the relationship between the model SNR and the test SNR. Then a lower detection boundary for all statistical tests is stated for the multiple leakage model. We call a leakage detection procedure *optimal* if and only if its detection boundary agrees with this theoretical lower detection boundary. We show that HC procedure detects leakage existence above exactly this theoretical detection boundary, while mini-p procedure has a higher detection boundary. Figure 3 shows these regions. Hence the mini-p procedure is not optimal but the HC procedure is optimal .

Model Notations First, we normalize the noise and the intermediate variables in (1) and (2), so that

$$L_i = \tilde{V}\delta_i + r_i, \quad i = 1, \cdots, n_L \tag{12}$$

where $r_i \sim N(0,1)$ is standard Gaussian distributed noise, $\tilde{V}$ is the normalized intermediate variable so that $E(\tilde{V}) = 0$ and $Var(\tilde{V}) = 1$. Hence the model $SNR = Var(\tilde{V}\delta_i)/Var(r_i) = \delta_i^2$ at the i-th time point.

Let n_0 denotes the total number of POIs (where $\delta_i \neq 0$). Then we represent

$$n_0 = n_L^{1-\beta}. \tag{13}$$

The β, called *sparsity parameter*, reflects how sparse the POIs are (bigger β, sparser the POIs). We will focus on the case of sparse signals with $\beta \in (1/2, 1)$.

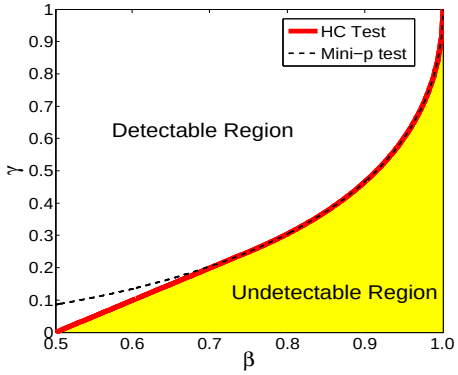


Fig. 3: The undetectable/detectable regions for mini-p test and HC test.

The ability of any statistical procedure to detect the signals is affected by the sparsity and the strength of the signals. For simplicity, let us focus on a simple model where all POIs has the same SNR Δ^2, i.e., all non-zero $\delta_i = \Delta$. Let $\bar{r}_i = \sum_{j=1}^{n_{tr}} r_{i,j}$ be the average of noises in the n_{tr} measurements at i-th time point. Then the expected value of the maximum noise $\max_{1\le i\le n_L} |\bar{r}_i|$ is $\sqrt{2\log(n_L)/n_{tr}}$. Therefore, if the SNR $\Delta^2 > 2\log(n_L)/n_{tr}$, then the leakage and their locations are easy to identify. On the other hand, if we do not need to identify the location, we can detect the existence of leakage for smaller SNR $\Delta^2 < 2\log(n_L)/n_{tr}$. We denote

$$\Delta^2 = \gamma \cdot 2\log(n_L)/n_{tr}, \quad 0 \le \gamma \le 1, \tag{14}$$

calling γ the strength parameter. Then the detectable limit of signals can be represented in terms of the strength γ and the sparsity β. To study the effectiveness of the leakage procedures, we consider the asymptotic case that both n_L and n_{tr} increases to infinity, and $n_{tr} >> \log(n_L)$ so that $\Delta = o(1)$ to concentrate on the detectable boundary for weak leakages. The strength parameter γ ($0 \le \gamma \le 1$) reflects the ratio of the SNR allowing leakage detection versus the SNR allowing exact leakage location identification.

fixed-vs-fixed and fixed-vs-random t-tests and ρ-tests While the leakage detection procedures (mini-p and HC) do not depend on the underlying univariate tests (t-test and ρ-test), we do derive the distributions of the test statistics $\widehat{\boldsymbol{s}}$ for commonly used univariate tests under the model (12) to provide some concrete examples. Particularly, we relate the SNR for these test statistics to the model SNR.

For *fixed-vs-fixed* t-test, the traces are partitioned into two sets with two different fixed values of V. Model (12) normalized $\tilde{V}$ so that $E(\tilde{V}) = 0$ and $Var(\tilde{V}) = 1$. Hence each $\tilde{V}_j$ takes on one of two values 1 or -1 with equal probability of 1/2, for $j = 1, ..., n_{tr}$. For the *fixed-vs-random* t-test, one data set

are fixed to V_{cons} and on the other data set the intermediate value V_{rand} varies randomly. Thus the normalized intermediate value $\tilde{V}$ have half probability being fixed to a constant $\tilde{V}_{cons}$, and half probability being assigned random value $\tilde{V}_{rand}$. Under null hypothesis of $\delta_i = 0$, the test statistic $\widehat{\boldsymbol{s}}_i$ for t-test in (6) and ρ-test in (4) both converge to a $N(0,1)$ distribution as $n_{tr} \to \infty$, for both fixed-vs-fixed and fixed-vs-random settings. We summarize their asymptotic distribution under the alternative hypothesis of $\delta_i = \Delta$ in both settings in the following Theorem, with proofs in Appendix 7.2.

Theorem 1. *When the total number of traces $n_{tr} \to \infty$: the fixed-vs-fixed t-test statistic*

$$\widehat{\boldsymbol{s}}_i \to N(\sqrt{n_{tr}\Delta^2}, 1); \tag{15}$$

The fixed-vs-random t-test statistic

$$\widehat{\boldsymbol{s}}_i \to N(\sqrt{n_{tr}\Delta^2}\tilde{V}_{cons}, 1)[1 + O(\Delta^2)]; \tag{16}$$

In all settings, the ρ-test statistic

$$\widehat{\boldsymbol{s}}_i \to N(\sqrt{n_{tr}\Delta^2}, 1)[1 + O(\Delta^2)]. \tag{17}$$

We can see that, omitting the smaller order term, the test SNR of ρ-test is $n_{tr}\Delta^2$ corresponding to the model SNR Δ^2 in model (12). Thus for any fixed model SNR Δ^2, increasing the number of measurements n_{tr} will lead to a linear increase in the test SNR and allows the leakage detection eventually. The test SNR for both ρ-test and t-test is $n_{tr}\Delta^2$ under the fixed-vs-fixed test setting. This asymptotic equivalence between the two types of distinguishers has been pointed out before in [5, 25]. For the fixed-vs-random setting, the t-test SNR $n_{tr}\Delta^2\tilde{V}_{cons}^2$ is smaller than the ρ-test SNR since $\tilde{V}_{cons} < 1$. Here the ideal ρ-test has a bigger SNR knowing the perfect power model (12). Without knowing the power model, a profiling method [5] may be used. The actual performance of a profiled ρ-test would be worse off than the ideal ρ-test.

The ρ-test and t-test under fixed-vs-random and fixed-versus-fixed samplings are both nonspecific leakage tests as the data partition is nonspecific. [5]'s ρ-test is conducted on samples of random plaintexts and being specific for a targeted intermediate V. The nonspecific data partition generally leads the tests to find more leakage signals along the trace than those found by the specific test. Thus the HC procedure are likely to perform better in the fixed-versus-fixed and fixed-versus-random settings, utilizing the multiple leakage signals.

These results provide some concrete examples for the next step of leakage detection with multiple (n_L) such tests. Next we consider the detection boundary for all statistical procedure for hypothesis testing of (8) versus (9), and the specific detection boundary for HC and mini-p procedures using these t-tests and ρ-tests.

Detection Boundaries The ability to detect leakage existence is affected by the signal strength, signal sparsity, total number of time points n_L and total

number of measurement traces n_{tr}. Statistical theory can tell the detectable limits when n_L and n_{tr} increases to the infinity. To apply the theory, we first introduce two definitions.

Definition 1. *A leakage detection procedure is asymptotically powerless if the sum of its type I and type II error rates converges to 1 for testing* (8) *versus* (9) *as n_L goes to infinity.*

Definition 2. *A leakage detection procedure is asymptotically powerful if the sum of its type I and type II error rates converges to 0 for testing* (8) *versus* (9) *as n_L goes to infinity.*

Clearly an asymptotically powerless leakage detection procedure can not really distinguish the existence of leakage or not. For each univariate hypothesis test we only need to consider the tests based on the sufficient statistic [26] for (12), which is $U_i = (1/n_{tr})\sum_{j=1}^{n_{tr}} \tilde{V}_j L_{i,j}$. U_i follows the $N(0, 1/n_{tr})$ distribution under the null hypothesis and follows $N(\Delta, 1/n_{tr})$ distribution under the alternative hypothesis. Hence the null hypothesis (8) is equivalent to: $U_i \sim N(0, 1/n_{tr})$ for $i = 1, ..., n_L$, and the alternative hypothesis (9) is equivalent to: $U_i \sim (1-q)N(0, 1/n_{tr}) + qN(\Delta, 1/n_{tr})$ for $i = 1, ..., n_L$. Here q is the proportion of real POIs, which is $q = n_L^{-\beta}$ by (13). For this mixture Gaussian distribution testing problem, there is a detection boundary in the (β, γ) plane such that all statistical procedures are asymptotically powerless in the region below this detection boundary. That boundary is given by equation (1.6) in [11] (see [18] for proofs),

$$g(\beta) = \begin{cases} \beta - 1/2 & 1/2 < \beta < 3/4, \\ (1 - \sqrt{1-\beta})^2 & 3/4 \le \beta < 1. \end{cases} \tag{18}$$

Therefore, the generic detection boundary for all leakage detection procedures is summarized as the following.

Theorem 2. *When $\gamma < g(\beta)$, all leakage detection tests are asymptotically powerless. That is*

$$\mathrm{P}_{H_0}(\textit{Reject } H_0) + \mathrm{P}_{H_1}(\textit{Accept } H_0) \to 1, \quad n_L \to \infty, \tag{19}$$

for testing the null hypothesis (8) *versus the alternative hypothesis* (9)*.*

Theorem 2 states that, comparing to the test SNR$= 2\log(n_L)/n_{tr}$ that allows exact leakage identification, no statistical procedure can cleanly detect the existence of leakage if the test SNR is further reduced by a factor $\gamma < g(\beta)$.

For each leakage detection procedure, there is a *specific leakage detection boundary*. Above the specific curve, the corresponding leakage detection procedure is asymptotically powerful, while below this curve, this test is asymptotically powerless. If the specific boundary achieves the theoretical boundary (18) in Theorem 2, then that leakage detection procedure is statistically optimal.

Definition 3. *A leakage detection procedure is optimal if its specific leakage detection boundary coincides with the generic leakage detection boundary* (18)*.*

The HC procedure leads to optimal leakage detection procedures.

Theorem 3. *We consider testing the null hypothesis* (8) *versus the alternative hypothesis* (9)*.*

- *For the ρ-tests and the fixed-vs-fixed t-test, when $\gamma > g(\beta)$,*
- *for the fixed-vs-random t-test, when $\gamma \tilde{V}_{cons}^2 > g(\beta)$*

the HC procedure is asymptotically powerful:

$$\mathrm{P}_{H_0}(\textit{Reject } H_0) + \mathrm{P}_{H_1}(\textit{Anccept } H_0) \to 0, \quad n \to \infty. \tag{20}$$

Notice that the HC leakage detection procedure with the fixed-vs-random t-test has a specific detection boundary higher than (18), due to the fact that the test SNR for this t-test is smaller than allowable by the inherent model SNR. However, among all statistical procedures based on the fixed-vs-random t-test statistics, HC procedure does achieve the theoretical minimum detection boundary. In contrast, the mini-p procedure (current TVLA standard) is not optimal. The detection boundary of mini-p procedure is given by

$$g_{max}(\beta) = (1 - \sqrt{1-\beta})^2, 1/2 \leq \beta < 1. \tag{21}$$

Theorem 4. *We consider testing the null hypothesis* (8) *versus the alternative hypothesis* (9)*.*

- *For the ρ-tests and the fixed-vs-fixed t-test, when $\gamma > g_{max}(\beta)$,*
- *for the fixed-vs-random t-test, when $\gamma \tilde{V}_{cons}^2 > g_{max}(\beta)$*

the mini-p procedure is asymptotically powerful:

$$\mathrm{P}_{H_0}(\textit{Reject } H_0) + \mathrm{P}_{H_1}(\textit{Anccept } H_0) \to 0, \quad n \to \infty. \tag{22}$$

See the proof in Appendix 7.4. According to Theorem 1.4 of [11], the mini-p procedure is asymptotically powerless below this specific detection boundary.

In the Fig. 3, we draw the leakage detection boundaries. The red line is the generic leakage detection boundary which coincides with the specific leakage detection boundary for HC procedure. Below this line (the yellow area) is the undetectable region, and above this line is the detectable region. The mini-p procedure's specific leakage detection boundary curve is plotted as the black dash line, higher than the red line. Hence the mini-p procedure is not optimal, and there are leakages that can be detected by HC procedure but not by the mini-p procedure.

When there is only a single POI ($n_0 = 1$, corresponding to sparsity $\beta = 1$), the detection and location identification of the leakage signal in a long trace is equally difficult (with detection boundary at $\gamma = 1$ in the top-right corner of Fig. 3). That happens because, for detection, the leakage test signal at this single POI has to clearly exceed all the noise test signals at other points on the trace, thus revealing the location of this single POI. In such cases, the detection efficiencies are the same for the HC procedure and for the mini-p procedure.

When there are multiple POIs, we can detect their existence, since they pushes the p-values' distribution away from the uniform distribution, without needing all of their test signals risen to the top (which is required by exact identification of their location). Thus as more POIs exist on the trace (i.e., as β decreases), it become easier to detect than to identify the leakage, as reflected by the decreasing γ along the detection boundary. With enough many POIs, the detection of leakage existence also becomes much easier using HC procedure than using mini-p procedure, which is reflected by the divergence between their detection boundaries as β decreases.

5 Numerical Results

In this section, we investigate the performance of HC procedure and mini-p procedure on synthetic data and real implementations. The results on synthetic data validate the theoretical analysis of Section 4 on the impact of the signal strength and the signal sparsity on the leakage detection performances. Then, the experiments on real traces clearly show the relevance of making use the HC metrics in typical case-studies: *(i)* an unprotected and *(ii)* a masked implementation of the AES.

5.1 Validation on Synthetic Data

Setup description For these experiments, we simulate traces of a complete execution of a 8-bit AES-128 implementation (10 rounds) with a Hamming weight leakage function and Gaussian noise. The 16 Hamming weights corresponding to the 16 intermediate bytes are computed for the plaintext and the result of every `AddRoundKey`, `SubBytes`, and `MixColumns` operation. Each of the 496 calculated values is uniquely reflected in one time sample (hence dictating the traces length) on which random noise following a Gaussian distribution is added. We consider two cases with levels of noise corresponding to SNRs of 0.1 and 0.01. For both cases, the three detection tests discussed in Section 4 are applied: *(i)* non-specific t-test with fixed-vs-random plaintexts, *(ii)* non-specific t-test with fixed-vs-fixed plaintexts, and *(iii)* specific ρ-test with random plaintexts.

In order to test the performance of the HC and mini-p procedures, we oberve their evolution when adding more and more traces. If a statistic is greater than its respective threshold, we consider that a leakage is detected (returning 1), and that there is no detected leakage otherwise (returning 0). This experiment is repeated 100 times on independent trace sets. The 100 obtained vectors are then averaged to build success curves. Fig. 4 shows the success rates of the HC (red curve) and mini-p (blue curve) procedures that are applied on the p-values output by these three detection tests.

Note: the purpose of these experiments is not to directly compare non-specific and specific leakage detection tests. They are rather chosen because of the different signals they exploit. In the first case, a non-specific detection test aims

at finding leakages in a non-sparse signal with a larger amplitude: every sample can lead to detection regardless of its actual usability (i.e. to retrieve the key). In the second case, a specific detection test aims at finding leakages in a sparse signal with lower amplitude: it only spots the useful points-of-interest.

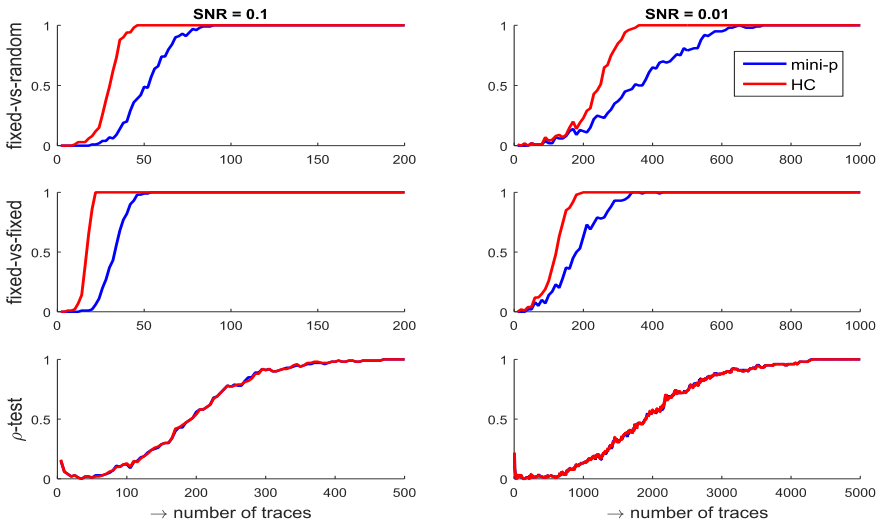


Fig. 4: Success rates curves for the HC (red) and mini-p (blue) procedures applied on the fixed-vs-random, fixed-vs-fixed, and ρ leakage detection tests.

Results interpretation The results depicted in Fig. 4 allow us to make the following observations:
(I) On the signal sparsity: the detection based on the HC procedure performs better than the one based on the mini-p only with the non-specific tests, i.e. when the signal is not sparse (all data-dependent samples can be spotted by the test, independent of their exploitability). Conversely, the specific test targets a very specific value. Therefore, the signal is very sparse (there is a single point-of-interest) and the HC and mini-p success rate curves completely match. This first observation support the detectable region boundaries depicted in Section 4. The single point-of-interest in the specific test here is a simulated extreme case. In practice, a single leaky instruction can also lead to multiple points-of-interest on the measured traces (e.g., due to high sampling rate). Then, even for the specific tests, the HC procedure will detect the leakage faster than the mini-p procedure in practice.
(II) On the impact of the noise: as previously observed in the literature [25], increasing the noise leads to decreasing the detection speed by the same factor for a given procedure. Therefore, the ratio between the detection speed of the HC and mini-p procedures remains constant. However, although it does not

change much for devices with low levels of noise, it can have an impact on the certification outcome for devices with large levels of noise.
(III) Non-specific detection tests: as previously stated by Durvaux et al. [5] one can notice that appropriately choosing the input of a non-specific test can lead to a better detection: the fixed-vs-fixed test performs approximately twice better than the fixed-vs-random test. Due to our Hamming weight leakage function, the maximum distances are twice larger with the fixed-vs-fixed than with the fixed-vs-random test. Similarly to the impact of the noise, the larger is the noise, the bigger is the potential impact on a certification outcome.

To summarize, these preliminary results mostly show that there is a clear practical improvement of the HC procedure over the mini-p in cases where *(i)* the signal is not sparse, and *(ii)* the SNR is low. In the next experiments, we focus on the ρ leakage detection test.

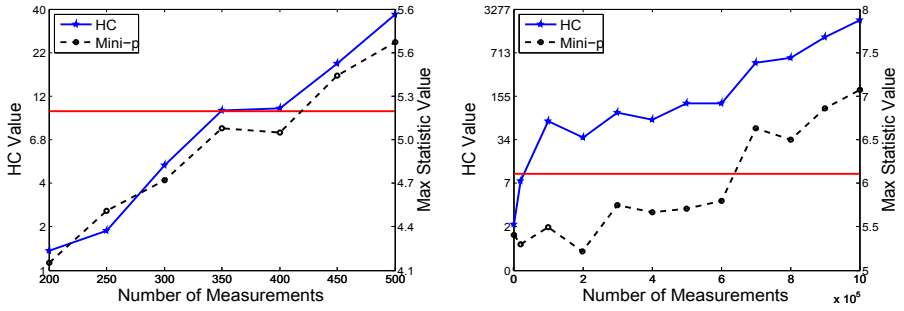
5.2 Leakage Detection on Real Traces: Unprotected AES

Setup description In this section, we investigate the performances of the HC and mini-p procedures on real power traces for non-sparse signal and high SNR. For this purpose, we consider an unprotected AES implementation running on an AVR 8-bit micro-controller embedded on a SASEBO-W board. Power traces are sampled with a LeCroy WaveRunner 640zi oscilloscope that produces $50,000$-sample leakage traces. The results based on a fixed-vs-random ρ leakage detection test are given in Fig. 5 (a). Instead of previous success rate curves, we show that statistical values of HC and mini-p procedures as what evaluators do during the leakage examination. They are displayed with respectively the blue and the black curves (scales are respectively labeled on the left and right sides).

Results interpretation Under the significance level of 0.01, with $n_L = 50,000$, the thresholds of maximum ρ test statistic (Fisher's transformation) and HC statistic are 5.2 and 10.1, respectively. (Note: they can be easily calculated by the code we provided.) In Fig. 5, the red line denotes these cutoffs. Once the obtained statistic value exceeds the red line, evaluators declare that the leakage is detected. The HC procedure detects the existence of leakage with about $n_{tr} = 350$ while the mini-p procedure requires $n_{tr} = 450$. HC procedure is a little more efficient than mini-p procedure, and it coincides with the strong leakage signal strength (estimated SNR around 0.2).

5.3 Leakage Detection on Real Traces: Masked AES

Setup description We then illustrate the application of HC procedure on detecting second order leakage on a masked AES implementation, i.e. low SNR (sparsity in this case is hard to estimate, but the results indicate that there are multiple leakage points for masked values). For this purpose, we make use of traces available on the website of the TeSCASE project [27]. The masked implementation of the AES follows the scheme described in [28] and runs on the Virtex-5 FPGA embedded on a SASEBO-GII board. This set of traces contains



(a) Unprotected AES data (b) Masked AES data

Fig. 5: Statistic Values of Mini-p and HC procedures on two AES implementations.

$N = 1,400,000$ power traces of $n_w = 3125$ samples. It was previously verified that the traces embed no first-order leakage. Then, HC and mini-p procedures are compared for detecting the second-order leakage existence for this protected implementation. Since the centered-product is the natural candidate when attacking second-order leakages [29–31], we use it to combine all pairs of leakages. The result is then used as observations for leakage detection [4]. That is, for a n_w long trace, we examine correlations of the $n_L = n_w^2$ centered-product leakages with the Hamming distance of a targeted SBox (1st SBox byte in last round). The detection results based on ρ test are given in the Fig. 5 (b).

Results interpretation Under the significance level of 0.01, with $n_L = n_w^2$, the thresholds of maximum ρ test statistic and HC statistic are 6.1 and 10.1, respectively. Compared to unmasked AES, its leakage signal strength is lower, both mini-p and HC procedure require much more measurements to detects the existence of second-order leakages. The HC procedure requires about $n_{tr} = 40,000$ measurements while the mini-p procedure requires $n_{tr} = 620,000$. In other words, in this case-study, the HC procedure allows detecting the leakages more than 15 times faster than the mini-p procedure.

6 Conclusions

We put the leakage detection procedure on a sound footing by proposing detection criterions based on the overall statistical Type I error rate. The proposed HC-based leakage detection procedure is shown to be theoretically optimal at combining detections from multiple leakage points, and can greatly improve the leakage certification process in practice.

References

1. G. Goodwill, B. Jun, J. Jaffe, and P. Rohatgi, "A testing methodology for side-channel resistance validation," in *NIST Non-Invasive At-*

tack Testing Workshop, Sept. 2011. [Online]. Available: http://csrc.nist.gov/news_events/non-invasive-attack-testing-workshop/papers/08_Goodwill.pdf
2. J. Cooper, E. DeMulder, G. Goodwill, J. Jaffe, G. Kenworthy, and P. Rohatgi, "Test vector leakage assessment (tvla) methodology in practice," in *International Cryptographic Module Conference*, 2013. [Online]. Available: http://icmc-2013.org/wp/wp-content/uploads/2013/09/goodwillkenworthtestvector.pdf
3. L. Mather, E. Oswald, J. Bandenburg, and M. Wójcik, "Does my device leak information? an a priori statistical power analysis of leakage detection tests," in *International Conference on the Theory and Application of Cryptology and Information Security*. Springer, 2013, pp. 486–505.
4. T. Schneider and A. Moradi, "Leakage assessment methodology," in *International Workshop on Cryptographic Hardware and Embedded Systems*. Springer, 2015, pp. 495–513.
5. F. Durvaux and F.-X. Standaert, "From improved leakage detection to the detection of points of interests in leakage traces," in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 2016, pp. 240–262.
6. A. A. Ding, C. Chen, and T. Eisenbarth, "Simpler, faster, and more robust t-test based leakage detection," 2016.
7. B. Bilgin, B. Gierlichs, S. Nikova, V. Nikov, and V. Rijmen, "Higher-order threshold implementations," in *International Conference on the Theory and Application of Cryptology and Information Security*. Springer, 2014, pp. 326–343.
8. E. Nascimento, J. López, and R. Dahab, "Efficient and secure elliptic curve cryptography for 8-bit avr microcontrollers," in *International Conference on Security, Privacy, and Applied Cryptography Engineering*. Springer, 2015, pp. 289–309.
9. T. De Cnudde, B. Bilgin, O. Reparaz, and S. Nikova, *Higher-Order Glitch Resistant Implementation of the PRESENT S-Box*. Cham: Springer International Publishing, 2015, pp. 75–93.
10. J. Balasch, B. Gierlichs, V. Grosso, O. Reparaz, and F.-X. Standaert, "On the cost of lazy engineering for masked software implementations," in *International Conference on Smart Card Research and Advanced Applications*. Springer, 2014, pp. 64–81.
11. D. Donoho and J. Jin, "Higher criticism for detecting sparse heterogeneous mixtures," *Ann. Statist*, pp. 962–994, 2004.
12. ——, "Higher criticism thresholding: Optimal feature selection when useful features are rare and weak," *Proceedings of the National Academy of Sciences*, pp. 14 790–14 795, 2008.
13. J. Fan and J. Lv, "Sure independence screening for ultra-high dimensional feature space," *J. Royal Statistical Society: Series B*, vol. 70, pp. 1–35, 2008.
14. J. Fan, Y. Feng, and R. Song, "Nonparametric independence screening in sparse ultra-high-dimensional additive models," *Journal of the American Statistical Association*, vol. 106, no. 494, pp. 544–557, 2011.
15. J. Li, D. Siegmund *et al.*, "Higher criticism: p-values and criticism," *The Annals of Statistics*, vol. 43, no. 3, pp. 1323–1350, 2015.
16. D. Donoho, J. Jin *et al.*, "Higher criticism for large-scale inference, especially for rare and weak effects," *Statistical Science*, vol. 30, no. 1, pp. 1–25, 2015.
17. Z. Wu, Y. Sun, S. He, J. Cho, H. Zhao, and J. Jin, "Detection boundary and higher criticism approach for rare and weak genetic effects," *Ann. Appl. Stat.*, vol. 8, no. 2, pp. 824–851, 06 2014. [Online]. Available: http://dx.doi.org/10.1214/14-AOAS724
18. Y. I. Ingster, "Minimax detection of a signal for i (n)-balls," *Mathematical Methods of Statistics*, vol. 7, no. 4, pp. 401–428, 1998.

19. C. Archambeau, E. Peeters, F.-X. Standaert, and J.-J. Quisquater, "Template attacks in principal subspaces," in *Int. Workshop on Cryptographic Hardware and Embedded Systems*, 2006, pp. 1–14.
20. F.-X. Standaert and C. Archambeau, "Using subspace-based template attacks to compare and combine power and electromagnetic information leakages," in *International Workshop on Cryptographic Hardware and Embedded Systems*. Springer, 2008, pp. 411–425.
21. M. Bär, H. Drexler, and J. Pulkus, "Improved template attacks," 2010.
22. M. Elaabid, O. Meynard, S. Guilley, and J.-L. Danger, "Combined side-channel attacks," in *Information Security Applications*, 2011, pp. 175–190.
23. O. Choudary and M. G. Kuhn, "Efficient template attacks," in *Smart Card Research and Advanced Applications*. Springer, 2013, pp. 253–270.
24. N. Bruneau, S. Guilley, A. Heuser, D. Marion, and O. Rioul, "Less is more," in *International Workshop on Cryptographic Hardware and Embedded Systems*. Springer, 2015, pp. 22–41.
25. S. Mangard, E. Oswald, and F. X. Standaert, "One for all - all for one: unifying standard differential power analysis attacks," *IET Information Security*, vol. 5, no. 2, pp. 100–110, June 2011.
26. E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses*. Springer Science & Business Media, 2006.
27. "Testbed for side channel analysis and security evaluation," 2014. [Online]. Available: http://tescase.coe.neu.edu
28. M.-L. Akkar and C. Giraud, "An implementation of DES and AES, secure against some attacks," 2001, pp. 309–318.
29. S. Chari, C. S. Jutla, J. R. Rao, and P. Rohatgi, "Towards sound approaches to counteract power-analysis attacks," in *Annual International Cryptology Conference*. Springer, 1999, pp. 398–412.
30. K. Schramm and C. Paar, "Higher order masking of the aes," in *Cryptographers Track at the RSA Conference*. Springer, 2006, pp. 208–225.
31. A. A. Ding, L. Zhang, Y. Fei, and P. Luo, "A statistical model for higher order dpa on masked devices," in *International Workshop on Cryptographic Hardware and Embedded Systems*. Springer, 2014, pp. 147–169.

7 Appendix

7.1 Matlab code for Conducting the HC-test

The following code implements the step (III) of the proposed leakage detection procedure. It takes in the p-values calculated in step (II) and returns the decision of leaky/non-leaky.

```
function [d hc_threshold] = hctest(x,alpha)
%HCTEST  Leakage detection test.
%   D = HCTEST(X) performs a hc-test of the hypothesis that the p vlaues
%   in the vector X come from a uniform distribution U(0,1), i.e., the
%   corresponding data is leakage-free, and returns the result of the test
%   in D.  D=0 indicates that the null hypothesis ("leakage-free") cannot
%   be rejected at the 1% significance level.  D=1 indicates that the null
%   hypothesis can be rejected at the 1% level. The corresponding data
%   contains some secret information.
%
%   X is a vector of p values.
%
%   D = HCTEST(X,ALPHA) performs the test at the significance level
%   (100*ALPHA)%.  ALPHA must be a scalar.
%
%   [D HC_THRESHOLD] = HCTEST(X,ALPHA) returns the threshold of HC test
%   at the significance level (100*ALPHA)%.  ALPHA must be a scalar.

%   References:

if nargin < 2
    alpha = 0.01;
elseif ~isscalar(alpha) || alpha <= 0 || alpha >= 1
    error(message('stats:ttest:BadAlpha'));
end

% Calculate the threshold of HC test at the significance level (100*ALPHA)%.
myfun = @(nl,th) hcpvalue(nl,th);
nl = length(x);
fun = @(th) myfun(nl,th)-alpha;
hc_threshold = 0.5;
x0 = 0.1;
while hc_threshold<1.07
    x0 = x0*10;
    hc_threshold = fzero(fun,x0);
end

% Calculate the value of the HC statistic
```

```
x_sort = sort(x,'ascend');
hc = sqrt(nl)*([1:nl]/nl-x_sort)./sqrt(x_sort.*(1-x_sort+1e-50));
hc_max = max(hc(1:floor(nl/2)));

% Determine if the data is leakage-free.
if hc_max>hc_threshold
   d = 1;
else
   d = 0;
end

%% The following function
function pvalue = hcpvalue(nl,th)
%HCPVALUE  The pvalue of the variable HC under the null hypothesis.
%   PVALUE = HCPVALUE(nL,TH) calculates the pvalue at the value TH
%   for the variable HC under the hypothesis that p values
%   come from a uniform distribution U(0,1),
%   i.e., the corresponding data is leakage-free.
%
%   NL is an interger: the number of leakage points.
%   TH is a value
%
%   References:
%      [1] M. Li and D. Siegmund "Higher criticism: $ p $-values and criticism",
%      The Annals of Statistics, 2015, vol. 43, no. 3, pp. 1323--1350.

f1 = @(x,y) (x + (y^2-y*(y^2+4.*(1-x).*x).^0.5)/2 ) / (1+y^2);
f2 = @(x,y) 1/(1+y^2) - y*(1-2.*x) ./ ((1+y^2) * (y^2+4*x.*(1-x)).^0.5);

k = [1:floor(nl/2)];
c1 = f1(k/nl,th/nl^0.5);
c2 = f2(k/nl,th/nl^0.5);
pvalue = sum(betapdf(c1,k,nl+1-k) .* (c1./k) .* (1-(1-k/nl).*c2./(1-c1)));
```

7.2 Proof of Theorem 1

(1) First we consider the fixed versus fixed setting, where each $\tilde{V}_j$ takes on one of two values 1 or -1 with equal probability of $1/2$, for $j = 1, ..., n_{tr}$. Hence under the alternative hypothesis of $\delta_i = \Delta$, $E[\bar{L}_{\tilde{V}=1} - \bar{L}_{\tilde{V}=-1}] = 2\Delta$, $Var[\bar{L}_{\tilde{V}=1} - \bar{L}_{\tilde{V}=-1}] = 2(2/n_{tr})$. We have, from Central Limit Theorem, the t-test statistic

$$\widehat{\boldsymbol{s}}_i \to \frac{\bar{L}_{\tilde{V}=1} - \bar{L}_{\tilde{V}=-1}}{\sqrt{4/n_{tr}}} \to N(\Delta\sqrt{n_{tr}}, 1). \tag{23}$$

The correlation in ρ-test of equation (3) in the main text becomes

$$\hat{\rho}_i = \frac{Cov(L_i, \tilde{V})}{\sqrt{Var(L_i)Var(\tilde{V})}} \to \frac{(1/n_{tr})\sum_{j=1}^{n_{tr}}(\delta_i \tilde{V}_j + r_{i,j})\tilde{V}_j}{\sqrt{(1+\delta_i^2)}}. \tag{24}$$

Therefore, under the alternative hypothesis of $\delta_i = \Delta$,

$$E(\hat{\rho}_i) = \frac{\Delta E(\tilde{V}^2) + E(r_i \tilde{V})}{\sqrt{(1+\Delta^2)}} = \frac{\Delta}{\sqrt{(1+\Delta^2)}},$$

and $Var(\hat{\rho}_i) = (1/n_{tr})E[\Delta^2 \tilde{V}^4 + r^2 \tilde{V}^2]/(1+\Delta^2) = 1/n_{tr}$. For small $\Delta = o(1)$ and $n_{tr} \to \infty$, omitting the smaller order term Δ^2 from $1+\Delta^2$, the ρ-test statistic $\widehat{\boldsymbol{s}}_i$ in equation (4) also follows $N(\Delta\sqrt{n_{tr}}, 1)$ distribution.

(2) Now we consider the fixed versus random setting, where $\tilde{V}$ has half probability being fixed to a constant $\tilde{V}_{cons}$, and half probability being assigned random value $\tilde{V}_{rand}$. Since $0 = E(\tilde{V}) = (1/2)\tilde{V}_{cons} + (1/2)E(\tilde{V}_{rand})$, we have

$$E(\tilde{V}_{rand}) = -\tilde{V}_{cons}.$$

Since $1 = Var(\tilde{V}) = E(\tilde{V}^2) = (1/2)\tilde{V}_{cons}^2 + (1/2)\{[E(\tilde{V}_{rand})]^2 + Var(\tilde{V}_{rand})\} = \tilde{V}_{cons}^2 + (1/2)Var(\tilde{V}_{rand})$, we have $Var(\tilde{V}_{rand}) = 2(1 - \tilde{V}_{cons}^2)$. Hence under the alternative hypothesis of $\delta_i = \Delta$, then $E[\bar{L}_{\tilde{V}=1} - \bar{L}_{\tilde{V}=-1}] = 2\Delta$, $Var[\bar{L}_{\tilde{V}=1} - \bar{L}_{\tilde{V}=-1}] = 2(2/n_{tr})$. Hence the t-test statistic

$$\widehat{\boldsymbol{s}}_i \to \frac{\bar{L}_{\tilde{V}_{cons}} - \bar{L}_{\tilde{V}_{rand}}}{\sqrt{2[1 + (1+\Delta^2 2(1-\tilde{V}_{cons}^2)]/n_{tr}}} \to N(\Delta\sqrt{n_{tr}}\frac{\tilde{V}_{cons}}{\sqrt{1+\Delta^2(1-\tilde{V}_{cons}^2)}}, 1).$$

Omitting the smaller order term Δ^2, this is

$$\widehat{\boldsymbol{s}}_i \to N(\Delta\sqrt{n_{tr}}\tilde{V}_{cons}, 1)[1 + O(\Delta^2)]. \tag{25}$$

Using the same calculations under equation (24), the mean and variance of the ρ-test statistic are

$$E(\hat{\rho}_i) = \frac{\Delta}{\sqrt{1+\Delta^2}} \qquad Var(\hat{\rho}_i) = \frac{1}{n_{tr}}\frac{1+\Delta^2 E(\tilde{V}^4)}{1+\Delta^2}. \tag{26}$$

Thus the ρ-test statistic, omitting the smaller order term, follows the $N(\Delta\sqrt{n_{tr}}, 1)$ distribution.
(3) For the specific data partition setting. Notice now, $\tilde{V}$ have mean zero and variance Δ^2. The calculations under equation (24) apply similarly to get that ρ-test statistic follows the $N(\Delta\sqrt{n_{tr}}, 1)$ distribution approximately.

7.3 Proof of Theorem 3

Here, we only prove for the HC procedure using the t-test under the fixed versus fixed setting, as the proof for all other tests are very similar. As shown in Theorem 1,

– on the non-leaky time points,

$$\widehat{\boldsymbol{s}} \sim N(0,1), \tag{27}$$

– on the leaky POIs,

$$\widehat{\boldsymbol{s}} \sim N(\Delta\sqrt{n_{tr}},1), \tag{28}$$

where $\Delta = \sqrt{2\gamma\log(n_L)/n_{tr}}$ as in equation (14).

Let $N(a)$ denote the number of test statistics ($\widehat{\boldsymbol{s}}_i$) exceeding $\sqrt{2a\log(n_L)}$ for some $a \geq 0$:

$$N(a) = \sharp\left\{i : \widehat{\boldsymbol{s}}_i \geq \sqrt{2a\log(n_L)}\right\}. \tag{29}$$

Then the HC statistic $\widehat{\mathrm{HC}}_{n_L,max}$ has some relationship with $N(a)$, $0 \leq a \leq 1$. Under null hypothesis,

$$N(a) \sim \text{ Binomial}(n_L, p_0), \tag{30}$$

where $p_0(a) = \mathrm{P}(N(0,1) \geq \sqrt{2a\log(n_L)}) = \Phi(\sqrt{2a\log(n_L)})$ with $\Phi(\cdot)$ denoting the cumulative density function of the standard normal distribution $N(0,1)$. Let

$$W(a) = \sqrt{n_L}\frac{N(a)/n_L - p_0(a)}{\sqrt{p_0(a)(1-p_0(a))}}. \tag{31}$$

Then $W(a)$ is an empirical process defined on $0 \leq a \leq 1$. For each fixed a, $W(a)$ converges to $N(0,1)$. As $n_L \to \infty$, $p_0(a)$ varies from 0 to 1/2 for $0 \leq a \leq 1$. Recall from equation (11),

$$\widehat{\mathrm{HC}}_{n_L,max} = \max_{1\leq i\leq n_L/2}\frac{\sqrt{n_L}(i/n_L - \hat{p}_{(i)})}{\sqrt{\hat{p}_{(i)}(1-\hat{p}_{(i)})}}$$

is in fact taking maximum of the $W(a)$ at $n_L/2$ discrete values of $p_0(a)$ between 0 to 1/2. So

$$\widehat{\mathrm{HC}}_{n_L,max} = \max_{0\leq a\leq 1} W(a) \tag{32}$$

when $n_L \to \infty$. So the distribution of $N(a)$ can help us derive probability of events involving $\widehat{\mathrm{HC}}_{n_L,max}$.

Under the alternative hypothesis,

$$N(a) \sim \text{ Binomial}(n_L, p_1), \tag{33}$$

where

$$p_1(a) = \mathrm{P}((1-\frac{n_0}{n_L})\cdot N(0,1) + \frac{n_0}{n_L}\cdot N(\sqrt{2\gamma\log(n_L)},1) \geq \sqrt{2a\log(n_L)}). \tag{34}$$

We now analyze the order of p_0 and p_1 for later derivations. Recall for the standard normal variable Z,

$$\phi(z)(\frac{1}{z}-\frac{1}{z^3}) \leq \Phi(z) \leq \frac{\phi(z)}{z}, \tag{35}$$

where $\phi(\cdot) = exp(-z^2/2)/\sqrt{2\pi}$ is the probability density function of the standard normal distribution $N(0,1)$. Using the inequality (35), we have,

$$p_0(a) = \Phi(\sqrt{2a\log(n_L)}) = \mathcal{O}(\frac{n_L^{-a}}{\sqrt{\log(n_L)}}); \tag{36}$$

and when $a > \gamma$,

$$\begin{aligned} p_1(a) &= (1-\tfrac{n_0}{n_L})\Phi(\sqrt{2a\log(n_L)}) + \tfrac{n_0}{n_L}\Phi(\sqrt{2a\log(n_L)} - \sqrt{2\gamma\log(n_L)}) \\ &= (1-\tfrac{n_0}{n_L})p_0(a) + \tfrac{n_0}{n_L}\mathcal{O}(\tfrac{n_L^{-(\sqrt{a}-\sqrt{\gamma})^2}}{\sqrt{\log(n_L)}}). \end{aligned} \tag{37}$$

When $n_0 = n_L^{1-\beta}$ as specified in the main text equation (13), this becomes

$$p_1(a) = p_0(a) + \mathcal{O}(\frac{n_L^{-\beta-(\sqrt{a}-\sqrt{\gamma})^2}}{\sqrt{\log(n_L)}}). \tag{38}$$

Now to show the HC procedure is asymptotically powerful:

$$\mathrm{P}_{H_0}(\text{Reject } H_0) + \mathrm{P}_{H_1}(\text{Accept } H_0) \to 0, \quad n \to \infty, \tag{39}$$

we show that both terms can go to zero at the same time. We first choose a threshold value so that the first term goes zero. Since the HC statistic $\widehat{\mathrm{HC}}_{n_L,max}$ converges to a limit value of $\sqrt{2\log(\log(n_L))}$ as $n_L \to \infty$, we reject H_0 when $\widehat{\mathrm{HC}}_{n_L,max} > \sqrt{2(1+b)\log(\log(n_L))}$ for a small fixed value of $b > 0$. Then

$$\mathrm{P}_{H_0}(\text{Reject } H_0) \to 0.$$

To show that $\mathrm{P}_{H_1}(\text{Accept } H_0) \to 0$ also at the same time, we note that

$$\mathrm{P}_{H_1}(\text{Accept } H_0) \leq \mathrm{P}_{H_1}(\widehat{\mathrm{HC}}_{n_L,max} \leq \sqrt{4\log(\log(n_L))})$$

since $b < 1$. Then using (32), this probability is also bounded above by $\mathrm{P}_{H_1}(W(a) \leq \sqrt{4\log(\log(n_L))})$ for any fixed a value between 0 and 1. That is,

$$\begin{aligned} &\mathrm{P}_{H_1}(\text{Reject } H_0) \\ &\leq \mathrm{P}_{H_1}(\sqrt{n_L}\tfrac{N(a)/n_L - p_0(a)}{\sqrt{p_0(a)(1-p_0(a))}} \leq \sqrt{4\log(\log(n_L))}) \\ &= \mathrm{P}_{H_1}(N(a) \leq n_L p_0(a) + \sqrt{n_L p_0(a)(1-p_0(a))4\log(\log(n_L))}) \\ &\leq \mathrm{P}_{H_1}(N(a) \leq n_L p_0(a) + \sqrt{n_L p_0(a)4\log(\log(n_L))}). \end{aligned} \tag{40}$$

On the other hand, under alternative hypothesis, $\mathrm{E}[N(a)] = n_L p_1(a)$ and $\mathrm{Var}[N(a)] = n_L p_1(a)(1-p_1(a))$, by the Chebyshev's inequality,

$$\begin{aligned} &\mathrm{P}_{H_1}(N(a) \le n_L p_0(a) + \sqrt{n_L p_0(a) 4 \log(\log(n_L))}) \\ &\le \tfrac{n_L p_1(a)(1-p_1(a))}{[n_L p_0(a) + \sqrt{n_L p_0(a) 4 \log(\log(n_L))} - n_L p_1(a)]^2}. \end{aligned} \tag{41}$$

Now we finish the proof by showing that the quantity in (41) converges to zero when $\gamma > g(\beta)$. For this, we need to further separate into two cases of $\gamma \ge 1/4$ and $\gamma < 1/4$.

For the first case of $\gamma \ge 1/4$ and $\gamma > g(\beta)$, we use $a = 1$ to bound the probability. From (36) and (38)

$$p_0(1) = \mathcal{O}(\frac{n_L^{-1}}{\sqrt{\log n_L}}), \qquad p_1(1) = p_0(1) + \mathcal{O}(\frac{n_L^{-\beta-(1-\sqrt{\gamma})^2}}{\sqrt{\log(n_L)}}).$$

When $\gamma \ge 1/4$ and $\gamma > g(\beta)$, we have $\beta + (1-\sqrt{\gamma})^2 < 1$, so

$$p_1(1) = \mathcal{O}(\frac{n_L^{-\beta-(1-\sqrt{\gamma})^2}}{\sqrt{\log(n_L)}}), \qquad p_0(1) = o[p_1(1)].$$

Hence the quantity in (41) is of the order as $1/[n_L p_1(1)]$. That is,

$$\mathrm{P}_{H_1}(\text{Reject } H_0) \le \mathcal{O}(n_L^{-1+\beta+(1-\sqrt{\gamma})^2}) \to 0. \tag{42}$$

For the second case of $\gamma < 1/4$ and $\gamma > g(\beta)$, we use $a = 4\gamma$ to bound the probability. From (36) and (38)

$$p_0(a) = \mathcal{O}(\frac{n_L^{-4\gamma}}{\sqrt{\log n_L}}), \qquad p_1(a) = p_0(a) + \mathcal{O}(\frac{n_L^{-\beta-\gamma}}{\sqrt{\log(n_L)}}).$$

Furthermore, if $\beta > 3\gamma$, then $p_1(a) = p_0(a)[1+o(1)]$. Hence the quantity in (41) is of the order as $n_L p_0(a)/[n_L(p_1(a)-p_0(a))]^2$. That is,

$$\mathrm{P}_{H_1}(\text{Reject } H_0) \le \mathcal{O}(n_L^{-1} \tfrac{n_L^{-4\gamma}}{\sqrt{\log n_L}} [\tfrac{\sqrt{\log(n_L)}}{n_L^{-\beta-\gamma}}]^2) = \mathcal{O}(\sqrt{\log(n_L)} n_L^{-1+2\beta-2\gamma}). \tag{43}$$

When $\gamma \ge 1/4$ and $\gamma > g(\beta)$, we have $\beta - \gamma < 1/2$ always, so the above quantity converges to zero.

On the other hand, if $\beta \le 3\gamma$, then $p_1(a) - p_0(a)$ dominates $p_0(a)$, and the quantity in (41) is of the order as $n_L p_1(a)/[n_L(p_1(a)-p_0(a))]^2$. That is,

$$\mathrm{P}_{H_1}(\text{Reject } H_0) \le \mathcal{O}(n_L^{-1} \tfrac{n_L^{-\beta-\gamma}}{\sqrt{\log n_L}} [\tfrac{\sqrt{\log(n_L)}}{n_L^{-\beta-\gamma}}]^2) = \mathcal{O}(\sqrt{\log(n_L)} n_L^{-1+\beta+\gamma}). \tag{44}$$

When $\gamma \ge 1/4$ and $\gamma > g(\beta)$, we have $\beta + \gamma < 1$ always, so the above quantity converges to zero.

Now we have shown that for every case of $\gamma > g(\beta)$, the HC procedure is indeed asymptotically powerful:

$$\mathrm{P}_{H_0}(\text{Reject } H_0) + \mathrm{P}_{H_1}(\text{Accept } H_0) \to 0, \quad n \to \infty. \tag{45}$$

7.4 Proof of Theorem 4

Here, we only prove for the HC procedure using using the t-test under the fixed versus fixed setting, as the proof for other tests are very similar. To show that the mini-p procedure is asymptotically powerful:

$$\mathrm{P}_{H_0}(\text{Reject } H_0) + \mathrm{P}_{H_1}(\text{Accept } H_0) \to 0, \quad n \to \infty, \tag{46}$$

we first select a choose a threshold value so that the first term goes zero. To do this, we reject H_0 if the mini-p value is less than $1 - \Phi(\sqrt{2a\log(n_L)})$, or equivalently when $\max_{1\le i\le n_L} \widehat{\boldsymbol{s}}_i > \sqrt{2a\log(n_L)}$ for some $a > 1$. Notice that this rejection occurs when $N(a) \ge 1$ for the $N(a)$ defined in equation (29) of last section. Then clearly

$$\mathrm{P}_{H_0}(\text{Reject } H_0) = \mathrm{P}_{H_0}(N(a) \ge 1) = 1 - [1 - p_0(a)]^{n_L}.$$

To judge the limit of this expression, notice that for any sequence of $x_n \to 0$,

$$(1 - x_n)^n = e^{n\log(1-x_n)} \to e^{-nx_n}, \qquad \text{as } n \to \infty. \tag{47}$$

Hence when $x_n = o(1/n)$, $(1 - x_n)^n \to e^0 = 1$; when $x_n = o(1)$ but $nx_n \to \infty$, $(1 - x_n)^n \to e^{-\infty} = 0$.

Recall equation (36),

$$p_0(a) = \mathcal{O}(\frac{n_L^{-a}}{\sqrt{\log(n_L)}}) = o(\frac{1}{n_L}).$$

We have

$$\mathrm{P}_{H_0}(\text{Reject } H_0) = 1 - [1 - o(\frac{1}{n_L})]^{n_L} \to 0, \text{ as } n_L \to \infty. \tag{48}$$

Now to ensure the second term $\mathrm{P}_{H_1}(\text{Accept } H_0) \to 0$ also, we have to be more careful in selecting the a value. When $\gamma > g_{max}(\beta) = (1 - \sqrt{1-\beta})^2$, then $1 - \beta > (1 - \sqrt{\gamma})^2$. For each (γ, β) value in this region, we can find an $a > 1$ value such that $1 - \beta > (\sqrt{a} - \sqrt{\gamma})^2$. Then under the alternative hypothesis, using equations (36) and (38), we have

$$p_1(a) = \mathcal{O}(\frac{n_L^{-a} + n_L^{-\beta-(\sqrt{a}-\sqrt{\gamma})^2}}{\sqrt{\log(n_L)}}) = \mathcal{O}(\frac{n_L^{-\beta-(\sqrt{a}-\sqrt{\gamma})^2}}{\sqrt{\log(n_L)}}). \tag{49}$$

Hence as $n_L \to \infty$, $p_1(a) \to 0$ and

$$n_L p_1(a) = \mathcal{O}(\frac{n_L^{1-\beta-(\sqrt{a}-\sqrt{\gamma})^2}}{\sqrt{\log(n_L)}}) \to \infty,$$

due to $1 - \beta > (\sqrt{a} - \sqrt{\gamma})^2$. By (47), we have

$$\mathrm{P}_{H_1}(\text{Accept } H_0) = \mathrm{P}_{H_1}(N(a) = 0) = (1 - p_1(a))^{n_L} \to 0, \text{ as } n_L \to \infty. \tag{50}$$

This finishes the proof that mini-p procedure is asymptotically powerful when $\gamma > g_{max}(\beta)$.