Towards Sound and Optimal Leakage Detection Procedure

Liwei Zhang[1], A. Adam Ding[1], Francois Durvaux[2], Francois-Xavier Standaert[2], and Yunsi Fei[3]

[1] Department of Mathematics, Northeastern University, Boston, MA 02115, USA
[2] ICTEAM/ELEN/Crypto Group, Universite catholique de Louvain, Belgium
[3] Department of Electrical and Computer Engineering
Northeastern University, Boston, MA 02115, USA
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. We provide theoretical proof of the optimality of the HC 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 analysis (SCA) has been shown to be a serious threat to modern cryptographic implementations. For more than a decade now, researchers actively invented various side-channel attacks and proposed countermeasures to protect devices against such attacks. As 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 often used for blackbox evaluation of SCA resistance. The TVLA procedure scans the leakage traces (e.g., physical measurements of the power consumption) with a univariate

test, and declares no leakage if the test statistics at all points along the leakage trace falls below a critical value.

It is preferred to use a generic univariate test in the TVLA procedure to avoid dependence on a specific leakage model. The CRI's TVLA proposal runs the Welch's t-test [1, 2] on data sets 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 fixed plaintexts while the random class of measurements come from encryptions of random plaintexts. Recently several extensions of the t-test (e.g., higher order and multivariate leakage detection) have been proposed by researchers [3–6].

Durvaux and Standaert [5] at EuroCrypt2016 proposed a correlation-based test (ρ-test) to detect exploitable leakage aimed at a particular intermediate computation. Such a *specific* test yields sparser leakage relating to this targeted intermediate value, and is better suited for identifying Point-Of-Interest (POI) for exploitable leakage. While this identification is necessary for practical SCA, it is not required for the purpose of leakage detection. Both the specific and non-specific leakage detection tests can be used in the TVLA framework.

In this work we first study the TVLA procedure itself from a theoretical perspective. The TVLA procedure declares a device as 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 $\alpha < 0.00001$ 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 number of leakage points on the trace increases. For long traces, the overall significance level can be quite large, so is the test statistic value, and therefore a non-leaky device can not pass the 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). The issue is caused by the multiple univariate tests at all time points which led [3] to suggest using false discovery rate to decide the detection limit. However, an explicit rigorous way of setting the threshold value would help for sound application of the current TVLA procedure.

In view of this state-of-the-art, we make two contributions in this paper. First, we propose a sound method to set the threshold value according to an overall statistical significance level. The current TVLA procedure makes the decision (leaky versus non-leaky) based on the largest test statistic, hence it is a statistical minimum p-value (mini-p) procedure that decides only with the minimal p-value of all those univariate tests. 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).

Second, we propose to improve the (univariate) leakage detection procedure with a statistically optimal HC metric. For the leakage detection purpose, the evaluator searches for evidence of key-dependent leakages along the trace, without necessarily identifying the POIs exactly. Hence it is very similar as the statistical independence scanning procedure [11–17] widely used in other high-dimensional statistical applications. Depending on 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. We incorporate the "Higher Criticism" (HC), a state-of-art statistical method for detecting sparse and weak signals, into the TVLA 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 method

In the TVLA leakage detection setup, an evaluator collects many traces of physical measurements, and tries to find if some points on the traces leak key information through a key-sensitive intermediate variable V. 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 point, and makes decision by combining the results. That is, we test the null hypothesis (there is no leakage signal):

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

versus the alternative hypothesis (there is leakage):

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

at the i-th time point, where 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 under null hypothesis, i.e., $\mathrm{P}(|S| \geq |\widehat{\boldsymbol{s}}_i|)$ where S denotes a random variable that follows the distribution of

the test statistic under the null hypothesis (1). 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 $\max_{1\le i\le n_L} |\widehat{\boldsymbol{s}}_i| \ge \mathrm{TH}$ 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 method for considering multiple (n_L) testing but utilizing only the test with the minimal p-value. We will propose changing this mini-p multiple testing method later.

While the usage of a particular univariate test is not essential to the framework, we first describe two common univariate tests to use as concrete examples for a better understanding of the overall leakage detection 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 most natural attack is the correlation power analysis (CPA) distinguisher. The CPA uses the Pearson correlation, ρ, which can 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 $N(0,1)$. 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 nonlinear (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 $\widehat{\boldsymbol{s}}_i$:

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

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.

3 Methodology

In this section, we first discuss how to set the threshold for the 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. We summarize the HC leakage detection framework the step by step.

3.1 Threshold Setting in the 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 < 0.00001 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% Type I error rate) 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 the 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 $\alpha_{\text{TH}} = 1 - (1-\alpha)^{1/n_L}$ for an overall significance level α. Then for the t-test, the threshold is $\text{TH} = \text{CDF}_t^{-1}(1 - \alpha_{\text{TH}}/2, \nu_s)$ 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 $\text{CDF}_{N(0,1)}^{-1}(1 - \alpha_{\text{TH}}/2)$. 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

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

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

The current mini-p procedure ignores the information on all other p-values except for the minimal p-value $\min_{1 \le i \le n_L} p_i$. 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.

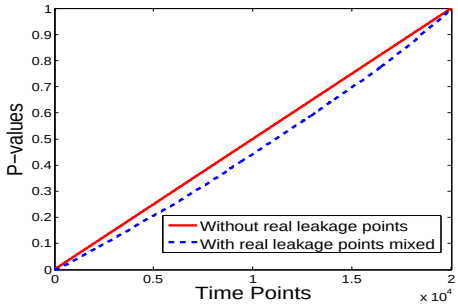


Fig. 1: Comparison of the distributions of ordered p-values under the null hypothesis and under the alternative hypothesis.

The leakage detection problem can now be restated as comparing the distribution of the obtained p-values p_1, ..., 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 $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(n_L)}$ 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,

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

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 $p_{(1)}$ 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 $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(n_L)}$.

(2) Calculate $\widehat{\mathbf{HC}}_{n_L,i}$, $i = 1, ...n_L$, 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 $b^{\mathrm{HC}}_{n_L,\alpha}$ at a given 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 large n_L, 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].

A matlab program for using HC procedure on leakage detection are provided in the appendix. The user provides the n_L test p-values and the type I error rate α, and the program output the detection results.

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	10^2	10^3	10^4	10^5	10^6	10^7	10^8
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 downward 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

Next we present our HC detection framework with salient steps. Fig. 2 gives the flow chart, where the steps within the dash-circled box are common in the current TVLA procedure as well.

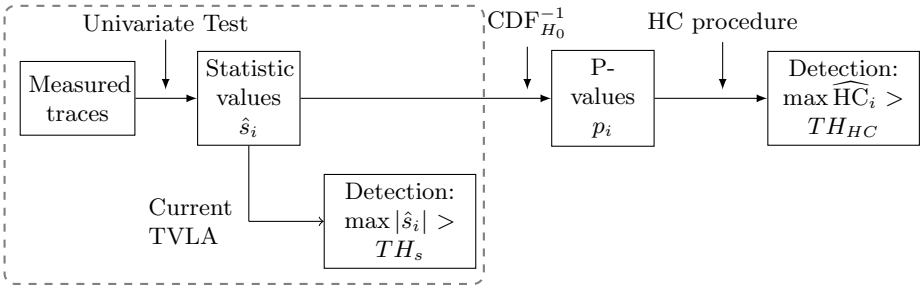


Fig. 2: HC leakage detection flow chart.

(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 detection step (III), we write a computation module to efficiently calculate the thresholds of HC (provided in the appendix). The user provides the n_L test p-values and the type I error rate α as inputs, and the leakage detection module 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, which can be made sound by choosing the threshold as in Section 3.1. The proposed approach conduct the HC procedure in step (III) instead.

4 Theory on optimal leakage procedure using HC.

This section introduces the theory on optimality of the HC procedure in high-dimensional statistical testing, and apply it to the leakage detection setting.

4.1 Optimality of the HC Procedure in Mixture Gaussian Testing

We first describe the statistical theory on the optimality of the HC procedure for the common mixture Gaussian setting in literature. That is, we test

$$H_0 : x_i \sim N(0,1), \text{ versus } H_1 : x_i \sim (1-q)N(0,1) + qN(\Delta, 1), \tag{12}$$

for observations x_i, $i = 1, ..n_L$. We then show how such theory can be used in the leakage detection testing of (8) versus (9) in the next subsection.

This mixture Gaussian distribution setting can be considered as testing the q proportion of signals with strength Δ in a sample of n_L dimension. The high-dimensional statistical theory indicates how strong (Δ) a signal can be detected for any given sparsity level as $n_L \to \infty$. The common notations in literature re-express the sparsity factor and the signal strength as two parameters $\beta = -\log(q)$ and $\gamma = \Delta^2/[2\log(n_L)]$. On the Euclidean space constructed by these two factors, statistical theory indicates that there is an undetectable region where no statistical tests can distinguish H_0 and H_1 well. More precisely, we first introduce the following definition.

Definition 1. *A statistical test procedure is asymptotically powerless (or asymptotically powerful) if the sum of its type I and type II error rates converges to 1 (or 0) as n_L goes to infinity.*

Theoretical Boundary For the mixture Gaussian distribution testing problem (12), all statistical procedures are asymptotically powerless in the region below the detection boundary given by equation (1.6) in [11] (proofs in [18]),

$$g(\beta) = \begin{cases} \beta - 1/2 & 1/2 < \beta < 3/4, \\ (1-\sqrt{1-\beta})^2 & 3/4 \le \beta < 1. \end{cases} \tag{13}$$

Detection Boundaries of HC and mini-p Procedures The HC procedure is optimal [11] for testing (12) because the HC procedure is asymptotically powerful when $\gamma > g(\beta)$, i.e., for all parameters (β, γ) above the theoretical minimum detection boundary (13). In contrast, a mini-p procedure is not optimal since it is asymptotically powerless for all parameters (β, γ) below the following boundary, according to the Theorem 1.4 of [11],

$$g_{max}(\beta) = (1-\sqrt{1-\beta})^2, 1/2 \le \beta < 1. \tag{14}$$

Fig. 3 draws these two detection boundaries (13) and (14). The solid red line is the detection boundary for HC procedure which coincides with the theoretical minimum detection boundary. Below this line (the yellow shade area) is the undetectable region, and above this line is the detectable region. The mini-p procedure's detection boundary curve is plotted as the black dash line, higher than the red line. In the next subsection, we show that these optimality theory do apply to the leakage detection setting.

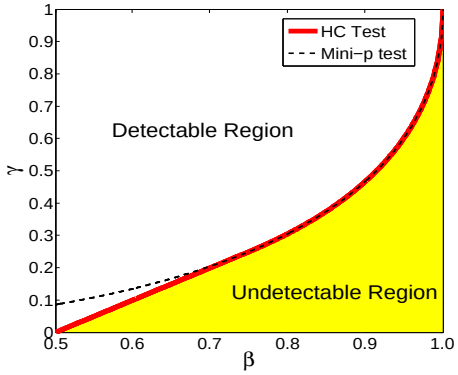


Fig. 3: The undetectable/detectable regions for mini-p test and HC test.

4.2 Leakage Detection Boundaries and Optimal Procedures

For any test statistic $\hat{s}_i$ based on L_i and for any linear transformation $f(L_i)$, there is always an equivalent test statistics based on $f(L_i)$ that gives exactly the same p-value. Therefore, without loss of generality, we consider the noise and the intermediate variables in (1) and (2) are normalized so that

$$L_i = \tilde{V}\delta_i + r_i, \quad i = 1, \cdots, n_L \tag{15}$$

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.

Theoretical Boundary For simplicity, we consider the simple model where there are $n_0 = qn_L$ POIs with same SNR Δ^2. That is, q proportion of δ_i taking a common non-zero value Δ (and the rest of $\delta_i = 0$). There are n_{tr} observations for each L_i: $L_{i,1}, ..., L_{i,n_{tr}}$. The most powerful test for the i-th univariate hypothesis test must be based on the sufficient statistic [25] for (15): $U_i = (1/n_{tr})\sum_{j=1}^{n_{tr}} \tilde{V}_j L_{i,j}$. Clearly U_i follows the $N(\delta_i, 1/n_{tr})$ distribution, and hence $\sim N(0, 1/n_{tr})$ at time points with no leakage($\delta_i = 0$). Let $x_i = \sqrt{n_{tr}}U_i$. Then the leakage detection problem becomes a mixture Gaussian distribution testing problem, using the sample x_1, ..., x_{n_L}, for

$$H_0 : x_i \sim N(0,1), \text{ versus } H_1 : x_i \sim (1-q)N(0,1) + qN(\sqrt{n_{tr}}\Delta, 1). \tag{16}$$

This is same as the problem (12) except the factor $\sqrt{n_{tr}}$. Therefore, the theoretical minimum detection boundary is given by (13), but with

$$\gamma = n_{tr}\Delta^2/[2\log(n_L)]. \tag{17}$$

Detection Boundaries of HC and mini-p Procedures We now compare the detection boundaries of HC and mini-p Procedures for the earlier concrete examples in section 2.2. Given a test statistic $\widehat{\boldsymbol{s}}_i$, either the t-test in (6) or the ρ-test in (4), we can consider its test SNR $|\mathrm{E}(\widehat{\boldsymbol{s}})|^2/\mathrm{Var}(\widehat{\boldsymbol{s}})$ which grows linearly in n_{tr}. Thus we denote its test SNR as $n_{tr}\delta_{s_i}^2$. Next we should that $\widehat{\boldsymbol{s}}_i$ converges to a $N(\sqrt{n_{tr}\delta_{s_i}^2}, 1)$ distribution as $n_{tr} \to \infty$, and hence can be related to the results for (16) above.

First, at non-leaky time points ($\delta_i = 0$), the test SNR also equals zero, and $\widehat{\boldsymbol{s}}_i \sim N(0, 1)$ as described in section 2.2.

Second, at POIs with $\delta_i = \Delta \neq 0$, we summarize the asymptotic distribution of $\widehat{\boldsymbol{s}}_i$ in the following Theorem, with proofs in Appendix 7.2.

Theorem 1. *When the total number of traces $n_{tr} \to \infty$:*
For the fixed-vs-fixed data partition, the t-test statistic in (6)

$$\widehat{\boldsymbol{s}}_i \to N(\sqrt{n_{tr}\Delta^2}, 1); \tag{18}$$

For the fixed-vs-random data partition and a constant $\tilde{V}_{cons} < 1$, the t-test statistic in (6)

$$\widehat{\boldsymbol{s}}_i \to N(\sqrt{n_{tr}\Delta^2}\tilde{V}_{cons}, 1)[1 + O(\Delta^2)]; \tag{19}$$

In all settings, the ρ-test statistic in (4)

$$\widehat{\boldsymbol{s}}_i \to N(\sqrt{n_{tr}\Delta^2}, 1)[1 + O(\Delta^2)]. \tag{20}$$

From this theorem, at POIs with $\delta_i \neq 0$, $\delta_{s_i} = \delta_i$ for the ρ-test statistic (4) in all cases and for the t-test statistic (6) in the fixed-vs-fixed test setup. In the fixed-vs-random setting, $\delta_{s_i} = \delta_i\tilde{V}_{cons} < \delta_i$.

Therefore, for the ρ-test statistic (4) in all cases and for the t-test statistic (6) in the fixed-vs-fixed test setup, $\{\widehat{\boldsymbol{s}}_i : i = 1, ..., n_L\}$ consists a data sample of size n_L for (16). Hence the HC-based leakage procedure achieves the theoretical minimum detection boundary above. But the current TVLA (mini-p) procedure is not optimal with boundary (14).

Remark 1. For the fixed-vs-random t-test, $\{\widehat{\boldsymbol{s}}_i : i = 1, ..., n_L\}$ consists a data sample for a problem similar to (16) but with SNR reduced by a factor $\tilde{V}_{cons}^2$. Thus the HC-procedure's detection boundary $g(\beta)/\tilde{V}_{cons}^2$ is the theoretical minimum detection boundary given $\{\widehat{\boldsymbol{s}}_i : i = 1, ..., n_L\}$. In contrast, the mini-p procedure's detection boundary is $g_{max}(\beta)/\tilde{V}_{cons}^2$. Our proposed HC-based leakage procedure is optimal in combining the n_L given univariate tests in this case too. We do not claim that the univariate test itself (such as the fixed-vs-random t-test, a generic test) is optimal, but rather claim that the procedure framework is optimal in combining the given univariate tests.

Remark 2. When there is only a single POI ($n_0 = 1$, corresponding to sparsity $\beta = 1$), the detection efficiencies are the same for the HC procedure and for the mini-p procedure. As more POIs exist on the trace (i.e., as β decreases),

the detection of leakage existence also becomes much easier using HC procedure than using mini-p procedure, which is reflected by the smaller n_{tr} needed to raise γ in (17) above the detection boundary $g(\beta)$ than $g_{max}(\beta)$.

Remark 3. To find exploitable leakage about a particular V, [5] sampled random plaintexts with varying V values for the ρ-test, to detect and locate sparse signals for this targeted V. The fixed-versus-fixed and fixed-versus-random t-tests, being nonspecific, would find more leakage signals along the trace. The choice of sampling scheme and tests affects both the SNR and the sparsity of the leakage signals. The HC procedure can lead to better detection than the mini-p procedure when there are multiple leakages, as likely in the fixed-versus-fixed and fixed-versus-random settings. Note that the identification of the exploitable leakage is a harder question than simply detecting leakage existence. For certification of non-leaky device, the optimal leakage detection procedure proposed here should be applied. To identify exploitable leakage, after leakage detection, specific test such as the ρ-test should be conducted and possibly more traces need to be collected for identification.

Remark 4. The HC procedure above assumed that the noise are independent among different time points along the trace. However, the performance of HC procedure is not affected under the likely short-range dependence [26] (i.e., the dependence among noises is concentrated to nearby time points) in practice. Extending generalized HC procedure [27] for strongly dependent noise for leakage detection can be considered in the future work.

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.

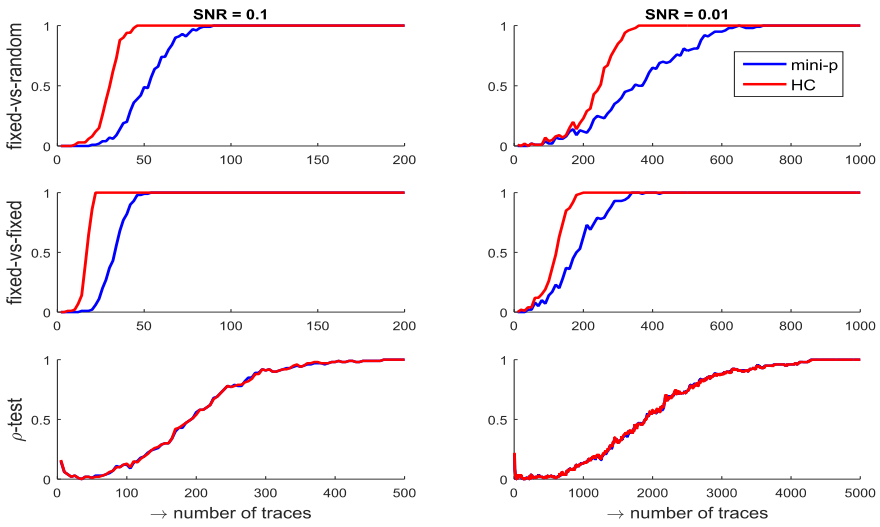


Fig. 4: Success rates curves for the HC (red) and mini-p (blue) procedures applied on the fixed-vs-random, fixed-vs-fixed, and ρ leakage detection tests.

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 [28], 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.

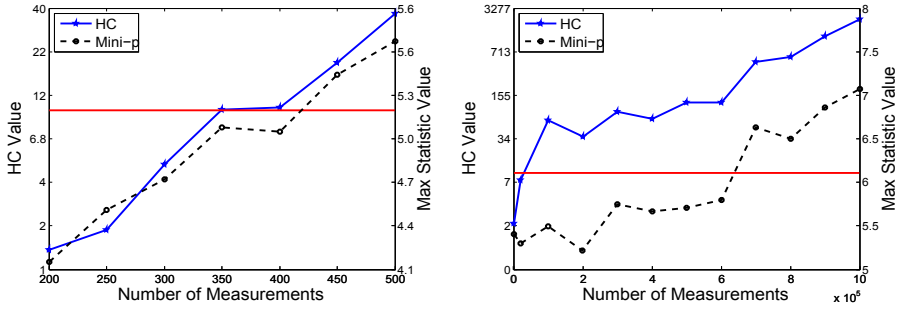
To summarize, these preliminary results mostly show that there is a clear practical improvement of the HC procedure over the mini-p in cases where *(i)* the signal is not sparse, and *(ii)* the SNR is low. In the next experiments, we focus on the ρ leakage detection test.

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).



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

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

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 [29]. The masked implementation of the AES follows the scheme described in [30] and runs on the Virtex-5 FPGA embedded on a SASEBO-GII board. This set of traces contains $N = 1,400,000$ 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 [31–33], 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.

Acknowledgment: This work has been funded in parts by National Science Foundation grants CNS-1314655, CNS-1337854 and CNS-1563697, and by the European Commission through the H2020 project 731591 (acronym REASSURE) and the ERC project 724725 (acronym SWORD). Franois-Xavier Standaert is a senior research associate of the Belgian Fund for Scientific Research (FNRS-F.R.S.).

References

1. G. Goodwill, B. Jun, J. Jaffe, and P. Rohatgi, "A testing methodology for side-channel resistance validation," in *NIST Non-Invasive Attack 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," in *Constructive Side-Channel Analysis and Secure Design: 7th International Workshop, COSADE 2016, Graz, Austria, April 14-15, 2016, Revised Selected Papers*, F.-X. Standaert and E. Oswald, Eds. Cham: Springer International Publishing, 2016, pp. 163–183.
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," in *Cryptography and Information Security in the Balkans: First International Conference, BalkanCryptSec 2014, Istanbul, Turkey, October 16-17, 2014, Revised Selected Papers*. 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," in *International Workshop on Constructive Side-Channel Analysis and Secure Design*, 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. E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses*. Springer Science & Business Media, 2006.
26. P. Hall and J. Jin, "Properties of higher criticism under strong dependence," *The Annals of Statistics*, pp. 381–402, 2008.

27. I. Barnett, R. Mukherjee, and X. Lin, "The generalized higher criticism for testing snp-set effects in genetic association studies," *Journal of the American Statistical Association*, vol. 112, no. 517, pp. 64–76, 2017.
28. 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.
29. "Testbed for side channel analysis and security evaluation," 2014. [Online]. Available: http://tescase.coe.neu.edu
30. M.-L. Akkar and C. Giraud, "An implementation of DES and AES, secure against some attacks," in *International Workshop on Cryptographic Hardware and Embedded Systems*, 2001, pp. 309–318.
31. 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.
32. K. Schramm and C. Paar, "Higher order masking of the aes," in *Cryptographers Track at the RSA Conference*. Springer, 2006, pp. 208–225.
33. 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 HC_THRESHOLD] = HCTEST(X,ALPHA) returns the results and the threshold
%     of HC test at the significance level (100*ALPHA)%.
%   X is a vector of p values.
%   ALPHA must be a scalar. If missing, then default value of 1% is used.
%
%   The hc-test test 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 alpha significance level.
%   D=1 indicates that the null hypothesis can be rejected at the alpha level.
%     That is, the corresponding data contains some secret information.
%

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 fixed values with equal probability of $1/2$, for $j = 1, ..., n_{tr}$. Since $\tilde{V}_j$ is already normalized in (15) to have mean zero and variance of one, the two fixed values have to be transformed into 1 and -1 here. Hence with $\delta_i = \Delta$, $E[\bar{L}_{\tilde{V}=1} - \bar{L}_{\tilde{V}=-1}] = [\Delta(1) - \Delta(-1)] = 2\Delta$, $Var[\bar{L}_{\tilde{V}=1} - \bar{L}_{\tilde{V}=-1}] = n_{tr}Var(r_j)/(n_{tr}/2)^2 = 4/n_{tr}$. We have, from Central Limit Theorem, the t-test statistic

$$\widehat{\boldsymbol{s}}_i \to \frac{\bar{L}_{\tilde{V}=1} - \bar{L}_{\tilde{V}=-1}}{\sqrt{4/n_{tr}}} \to N(\Delta\sqrt{n_{tr}}, 1). \tag{21}$$

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

$$\hat{\rho}_i = \frac{Cov(L_i, \tilde{V})}{\sqrt{Var(L_i)Var(\tilde{V})}} \to \frac{(1/n_{tr})\sum_{j=1}^{n_{tr}}(\delta_i \tilde{V}_j + r_{i,j})\tilde{V}_j}{\sqrt{(1+\delta_i^2)}}. \tag{22}$$

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 $\tilde{V}_j$ is already normalized to have mean zero, $0 = E(\tilde{V}) = (1/2)\tilde{V}_{cons} + (1/2)E(\tilde{V}_{rand})$, we have

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

Also $\tilde{V}_j$ is already normalized to have variance one, $1 = Var(\tilde{V}) = E(\tilde{V}^2) = (1/2)\tilde{V}_{cons}^2 + (1/2)\{[E(\tilde{V}_{rand})]^2 + Var(\tilde{V}_{rand})\} = \tilde{V}_{cons}^2 + (1/2)Var(\tilde{V}_{rand})$, we have $Var(\tilde{V}_{rand}) = 2(1-\tilde{V}_{cons}^2)$. Clearly this implies that the constant $\tilde{V}_{cons} < 1$. Hence under the alternative hypothesis of $\delta_i = \Delta$, then $E[\bar{L}_{\tilde{V}_{cons}} - \bar{L}_{\tilde{V}_{rand}}] = \Delta\tilde{V}_{cons} - \Delta E(\tilde{V}_{rand}) = 2\Delta\tilde{V}_{cons}$, and $Var[\bar{L}_{\tilde{V}_{cons}} - \bar{L}_{\tilde{V}_{rand}}] = [(n_{tr}/2)Var(r_j) + (n_{tr}/2)(\Delta^2 Var(\tilde{V}_{rand}) + Var(r_j))]/(n_{tr}/2)^2 = (2/n_{tr})[1 + \Delta^2 2(1-\tilde{V}_{cons}^2) + 1]$. Hence the t-test statistic

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

Omitting the smaller order term Δ^2, this is

$$\widehat{\boldsymbol{s}}_i \to N(\Delta\sqrt{n_{tr}}\tilde{V}_{cons}, 1)[1 + O(\Delta^2)]. \tag{23}$$

Using the same calculations under equation (22), the mean and variance of the ρ-test statistic are

$$E(\hat{\rho}_i) = \frac{\Delta}{\sqrt{1+\Delta^2}} \qquad Var(\hat{\rho}_i) = \frac{1}{n_{tr}}\frac{1+\Delta^2 E(\tilde{V}^4)}{1+\Delta^2}. \tag{24}$$

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 (22) apply similarly to get that ρ-test statistic follows the $N(\Delta\sqrt{n_{tr}}, 1)$ distribution approximately.