Multi-Tuple Leakage Detection and the Dependent Signal Issue

Olivier Bronchain, Tobias Schneider and François-Xavier Standaert

Université catholique de Louvain, B1348 Louvain-la-Neuve, Belgium
{olivier.bronchain,tobias.schneider,fstandae}@uclouvain.be

Abstract. Leakage detection is a common tool to quickly assess the security of a cryptographic implementation against side-channel attacks. The Test Vector Leakage Assessment (TVLA) methodology using Welch's t-test, proposed by Cryptography Research, is currently the most popular example of such tools, thanks to its simplicity and good detection speed compared to attack-based evaluations. However, as any statistical test, it is based on certain assumptions about the processed samples and its detection performances strongly depend on parameters like the measurement's Signal-to-Noise Ratio (SNR), their degree of dependency, and their density, i.e., the ratio between the amount of informative and non-informative points in the traces. In this paper, we argue that the correct interpretation of leakage detection results requires knowledge of these parameters which are a priori unknown to the evaluator, and, therefore, poses a non-trivial challenge to evaluators (especially if restricted to only one test). For this purpose, we first explore the concept of multi-tuple detection, which is able to exploit differences between multiple informative points of a trace more effectively than tests relying on the minimum p-value of concurrent univariate tests. To this end, we map the common Hotelling's T^2-test to the leakage detection setting and, further, propose a specialized instantiation of it which trades computational overheads for a dependency assumption. Our experiments show that there is not one test that is the optimal choice for every leakage scenario. Second, we highlight the importance of the assumption that the samples at each point in time are independent, which is frequently considered in leakage detection, e.g., with Welch's t-test. Using simulated and practical experiments, we show that *(i)* this assumption is often violated in practice, and *(ii)* deviations from it can affect the detection performances, making the correct interpretation of the results more difficult. Finally, we consolidate our findings by providing guidelines on how to use a combination of established and newly-proposed leakage detection tools to infer the measurements parameters. This enables a better interpretation of the tests' results than the current state-of-the-art (yet still relying on heuristics for the most challenging evaluation scenarios).

Keywords: Side-Channel Analysis · Leakage Detection · Security Evaluations

Introduction

State-of-the-Art

Leakage detection has become a de facto standard for the fast preliminary assessment of cryptographic implementations against side-channel attacks. In contrast to attack-based evaluations which quantify the difficulty to recover a key, detection methodologies try to answer a much simpler question: *do the measurements obtained by an adversary contain data-dependent information, independent of whether it can be efficiently exploited?* Cryptography Research's non-specific (fixed vs. random) t-test is the most popular example of this trend [CMG+, GJJR11]. It is commonly used to evaluate the security order of

IACR Transactions on Cryptographic Hardware and Embedded Systems ISSN 2569-2925, Vol. 2019, No. 2, pp. 318–345
DOI:10.13154/tches.v2019.i2.318-345

block cipher implementations by comparing leakages obtained for a fixed plaintext (and key) to leakages obtained with a random plaintext (and fixed key).

In general, leakage detection is thought to accelerate the evaluation process by avoiding the need to conduct numerous different attacks which require expert knowledge [Wag12]. Furthermore, it helps to reduce the evaluations' data complexity[1], since it relies on a simple statistical test to answer the aforementioned question, instead of conducting a complete key recovery. As discussed in [MOBW13, SM16, DS16], this gain largely stems from the fact that it is easier to compare two classes of leakages than 2^k ones for a side-channel attack targeting a k-bit key.[2] This reduced data complexity naturally comes at the cost of false positives, i.e., falsely detecting leakages, and false negatives, i.e., not detecting existing leakages, that can make the correct interpretation of detection tests difficult.

On the one hand, "false positives" are caused by detecting informative samples that are either data-dependent but not sensitive (e.g., corresponding to plaintext leakages), or sensitive but hard to exploit (e.g., corresponding to the middle rounds of a block cipher that are out of reach of a standard DPA [MOS11]). Note that strictly speaking, those are true positive from the statistical viewpoint, but "false positives" with respect to the goal of the detection test. Those "false positives" are usually less critical since (*i*) when applying leakage detection on full traces, they are usually accompanied by true positives, and (*ii*) if necessary they are easy to prevent by applying more expensive specific tests, e.g., based on the Signal-to-Noise Ratio (SNR) of some sensitive variables.

On the other hand, and more critically, various types of false negatives can also happen, possibly leading to a "false sense of security", i.e., an absence of detection despite the possibility to mount powerful side-channel attacks.

A first example of false negative is the case where the mean leakages of the fixed and random classes used in the non-specific test are identical (or so close that their detection is hard) [DS16]. This, however, is usually avoided by running the detection on long enough traces to ensure that there are some samples which do not fulfill this equality.[3]

A second and more challenging scenario happens when the strategy used by the leakage detection becomes suboptimal. This can for example take place with masked implementations, of which the security increases exponentially in the number of shares assuming independent and sufficiently noise leakages. As discussed in [Sta17], when the number of shares in the masking scheme is large and the noise is too low, the detection strategy of estimating a higher-order statistical moment becomes extremely data-expensive, even though the target can be attacked after the observation of a single (essentially noise-free) trace. The natural solution to mitigate this risk is to analyze the leakage distributions rather than only their statistical moments. Proposals in this direction include the χ^2 test recently analyzed in the context of leakage detection [MRSS18], or previously proposed tests based on the estimation of the mutual information [MOBW13].[4]

Eventually, a third and so far less discussed issue is that current leakage detection tests typically focus on whether a single leakage sample is data-dependent. For long measurement traces, this strategy is commonly extended by the so-called min-p approach in which the detection is based on the sample of the trace leading to the minimum p-value. More sophisticated methods have been proposed (e.g., Higher Criticism in [DZD+17]), but even they remain far from a detection combining all the leaking samples.

Contribution

Based on this state-of-the-art, our contribution is threefold.

[1] The number of measurements needed to conclude the evaluation.

[2] The fixed vs. fixed test discussed in [DS16] allows marginal improvements in this direction.

[3] Which (e.g.,) occurs with block ciphers for which computations are pseudo-random after some rounds.

[4] If working in a setting where the internal randomness used by the countermeasures of the leaking devices is available to the evaluator, a simpler solution is to directly estimate the shares' SNR.

First, we show that the data complexity of non-specific tests can be further reduced thanks to multi-tuple detection. To this end, we rely on the well-established Hotelling's T^2-test which is the natural extension of the t-test for multiple variables. We propose a general instantiation of such a multi-tuple detection using Hotelling's T^2-test, together with a simplification that we denote as D-test, working under the assumption that the leakage samples of the measurement traces are independent.

Second, we observe that this simplifying independence assumption is implicitly or explicitly used in all current detection methodologies to set the detection threshold in function of the trace length.[5] We demonstrate that dependencies within the traces can strongly affect the results of such detection tests, which rely on the independence assumption. This makes simple conclusions such as "no leakages were detected with up to N measurements" difficult to fairly compare statistically without a proper assessment of the dependencies within the traces. Since Hotelling's T^2-test naturally captures such dependencies, it therefore becomes a tool of choice whenever applicable. Yet, it comes at the cost of a significantly higher computational complexity, since it essentially requires inverting the leakage traces' covariance matrix.

Third, we provide experiments with both simulated data and actual measurements obtained from software and hardware implementations to confirm our findings.

We conclude the paper by describing a general framework to guide evaluators in their selection of the appropriate leakage detection tests in function of a few questions regarding the evaluation/attack setting. On the one hand, what are the evaluator/adversary's assumptions in terms of implementation knowledge (i.e., does the evaluation take place in an open-source or closed-source setting) and randomness knowledge (i.e., can the evaluator access the randomness used in countermeasures such as masking or shuffling)? On the other hand, what are the evaluator/adversary's assumptions regarding the measurements (e.g., are the traces short enough so that their covariance matrix can be estimated, and can we assume the samples to be – sufficiently – independent otherwise)?

We show that (*i*) based on these questions, a good combination of leakage detection tests can lead to meaningful conclusions about the security order of a target implementation, the noise level of its measurements, and the density of informative samples in the traces, and (*ii*) these recommendations and conclusions range from formal to more heuristic, mostly depending on the implementation and randomness knowledge available.

Related works and cautionary remarks. Hotelling's T^2-test has already been mentioned as a potential tool for leakage detection in [RBG+16] but under a different viewpoint. They investigate the connections (and lack thereof) between leakage detection and more comprehensive evaluation metrics reflecting the success of an attack rather than analyzing the data complexity gains that it can provide. The risks due to dependencies within leakage traces have also been mentioned in a recent work by Bache et al. [BPG18], who left their analysis as a challenging open problem. The focus of this last reference is also quite different from ours. It mostly deals with the interpretation of negative detection outcomes (i.e., what can be concluded in the absence of detection?), while our primary focus is on positive detection results, for example in order to assess the "security order" of a masked implementation in a multi-model approach such as [JS17].

1 Background

In this section, we introduce the notations used in the paper and explain the current state of leakage detection based on Welch's t-test [CMG+, GJJR11] including its extension

[5] As previously discussed in [DZD+17], this threshold has to increase with longer traces, since the probability of having a detection by chance (i.e., false positive) increases in the number of samples.

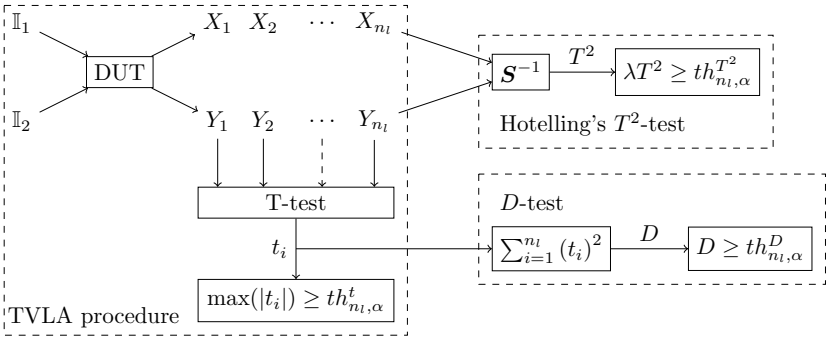


Figure 1: Leakage detection framework.

from [DZD+17]. It will serve as a frame of reference in our experiments to judge the potential of the newly-proposed methodologies.

1.1 Notations

Random variables are denoted as capital letters X while a random vector is written as $\boldsymbol{X} := [X_1, X_2, \ldots, X_{n_l}]$ with a density of informative points $\phi_{\boldsymbol{X}}$ (later defined in Subsection 3.1). Lower cases indicate an observation of a random variable such that $\boldsymbol{x} := [x_1, x_2, \ldots, x_{n_l}]$ is a sample of $\boldsymbol{X}$. For side-channel analysis, the random vectors represent measured traces where each X_i corresponds to a fixed point in time within an encryption. In this context, n_l and N are the trace length and the number of available traces (i.e., measurements).

We denote the mean and variance of a random variable X as μ_X and σ^2_X, while their sampled estimates are written as $\bar{X}$ and s^2_X. For random vectors, Σ and $\boldsymbol{S}$ represent the covariance matrix and its estimate. For statistical tests, α (resp., β) denotes the false positive (resp., negative) probability.

1.2 Leakage Detection with Welch's t-Test and Extensions

Test Vector Leakage Assessment (TVLA) describes an efficient detection methodology based on Welch's t-test initially proposed by Cryptography Research [CMG+, GJJR11]. In its core their approach tests the means of two different sets of measurements for equality and concludes the existence of leakages if a difference can be detected with a certain level of confidence. This basic procedure is split into two phases as depicted in Figure 1.

First, the device under test (DUT) is fed with two sets of inputs $\mathbb{I}_1$ and $\mathbb{I}_2$ while the leakage behavior during the computation is recorded. This results in a set of $N_{\boldsymbol{X}}$ (resp., $N_{\boldsymbol{Y}}$) measurements $\boldsymbol{X}$ (resp., $\boldsymbol{Y}$) for inputs $\mathbb{I}_1$ (resp., $\mathbb{I}_2$). In the widely-used non-specific fixed vs. random test, $\mathbb{I}_1$ consists of only one fixed plaintext and key while $\mathbb{I}_2$ corresponds to a fixed key with random plaintexts. In the following, our experiments are conducted in a *fixed vs. fixed* manner. As detailed in Appendix D, it generally provides better detection rates. Yet, our findings are generic and transfer to the *fixed vs. random* strategy.

Second, Welch's t-test [Wel47] is applied to these measurements for each pair of random variables (X_i, Y_i) corresponding to one point in time separately. It tries to distinguish between the null hypothesis H_0 and its corresponding alternative hypothesis H_a with

$$H_0 : \quad \mu_{X_i} = \mu_{Y_i}, \qquad H_a : \quad \mu_{X_i} \neq \mu_{Y_i}. \tag{1}$$

To this end, it is required to estimate the mean and variance of each random variable and

use them to derive the test statistic t_i and degrees of freedom v_i for each point in time as

$$t_i = \frac{\bar{X}_i - \bar{Y}_i}{\sqrt{\frac{s^2_{X_i}}{N_{X_i}} + \frac{s^2_{Y_i}}{N_{Y_i}}}}, \qquad v_i = \frac{\left(\frac{s^2_{X_i}}{N_{X_i}} + \frac{s^2_{Y_i}}{N_{Y_i}}\right)^2}{\frac{\left(\frac{s^2_{X_i}}{N_{X_i}}\right)^2}{N_{X_i}-1} + \frac{\left(\frac{s^2_{Y_i}}{N_{Y_i}}\right)^2}{N_{Y_i}-1}}.$$

Under H_0, t_i is assumed to follow a Student's t-distribution parameterized with v_i, where $\mathsf{F}_{\mathcal{T}}(\cdot, v_i)$ denotes the corresponding cumulative density function (CDF, cf. Appendix A). If the observed t_i significantly differs from the expected distribution, H_0 is rejected and the alternative hypothesis H_a is accepted. To ensure that this rejection is done with sufficient confidence, the p-value can be computed from t_i and v_i as

$$p_i = 2(1 - \mathsf{F}_{\mathcal{T}}(|t_i|, v_i)),$$

where a small p value indicates a high confidence to reject H_0. In this case, the evaluator concludes the existence of detectable leakages in the measurements.

Setting a Threshold. Usually, the p-value is compared to α, which is set a priori depending on the desired detection accuracy, and if smaller H_0 is rejected. This can be simplified for large numbers of observations for which $\mathsf{f}_t(\cdot, v_i)$ tends to a standard normal distribution. In this case, it is sufficient to directly set a threshold th^t_α for the t_i value, avoiding the computation of the degrees of freedom and p-values. For TVLA, the authors recommend a threshold of 4.5 for $|t_i|$, which if exceeded translates to $p_i < 0.00001$ assuming $v_i > 1000$. TVLA is usually employed using the min-p strategy to extend the test to traces with multiple points in time, i.e., the threshold $th^t_{10^{-5}} = 4.5$ is not affected by the number of points in each trace. However, as noted in [GJJR11] and more recently in [DZD+17], this approach can lead to erroneous results for measurements consisting of large numbers of points. Instead, the authors of TVLA propose to run additional, independent tests and only conclude leakage if it appears multiple times at the same time instance. We evaluate this strategy in Appendix D for our setting, and find that it is not superior to a single test (with twice the amount of measurements) for our purposes. Therefore, we do not discuss it any further. Alternatively, in [DZD+17] it is put forward to set an adjusted threshold based not only on the desired false positive rate α but also on the number of points per trace n_l (i.e., the number of separate Welch's t-tests) as

$$th^t_{n_l,\alpha} = \mathsf{F}^{-1}_{\mathcal{T}}(1 - \frac{1 - (1-\alpha)^{1/n_l}}{2}, v),$$

where F_t refers to the cumulative distribution function (CDF) of Student's t-distribution. In this way, the authors ensure that the desired false positive rate is achieved even for large n_l assuming independence between the random variables. We use this approach to set the threshold for the combination of min-p and Welch's t-test in our experiments. By inverting the previous equation, the p-value is computed according to

$$p = 1 - (1 + 2(\mathsf{F}_{\mathcal{T}}(\max(|t_i|)) - 1))^{1/n_l}.$$

Eventually, the false negative rate β which is the probability to not reject H_0 while $\mu_{\boldsymbol{X}} \neq \mu_{\boldsymbol{Y}}$ is given by

$$\begin{aligned}\beta &= \Pr[th^t_{n_l,\alpha} > \max(|t_i|)] \\ &= \Pi_{i=1}^{n_l} \Pr[th^t_{n_l,\alpha} > |\mathcal{T}(v_i; \delta_i)|],\end{aligned}$$

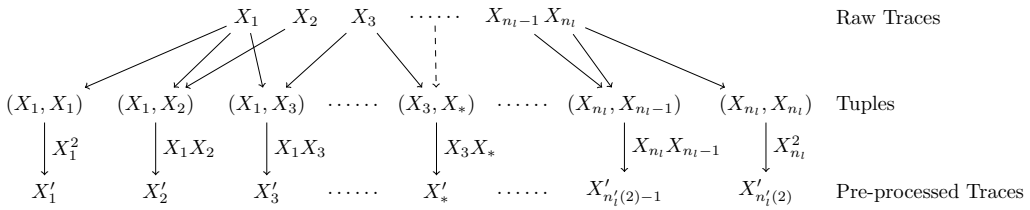


Figure 2: Exemplary multiplicative pre-processing of $\boldsymbol{X}$ for order $d = 2$.

where

$$\delta_i = (\mu_{X_i} - \mu_{Y_i}) / \sqrt{\frac{\sigma_{X_i}^2}{N_{X_i}} + \frac{\sigma_{Y_i}^2}{N_{Y_i}}}$$

is the non-central parameter. We observe that the false negative probability depends on the number of measurements as well as the samples' variance and the difference in mean between the two vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$. Namely, by decreasing the noise variance or by increasing the number of samples, the probability β decreases.

Higher Criticism. Detection can also be performed by recombining the p-values outputted by all the independent Welch's t-test's [DZD+17]. This higher criticism method is a first step in the direction of multi-tuple detection. In their paper, the authors showed that it provides an advantage over Welch's t-test for dense traces (i.e., traces with a lot of informative points) and is equivalent to it in case of sparse traces.

Extension to Higher Orders. The previously-described procedure tests for a difference in the means between the random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$. While for non-masked implementation this is sufficient to detect in most cases, masking schemes (such as [ISW03]) may remove any difference in the means if implemented correctly. For these cases, the leakages are hidden in a higher (mixed) statistical moment and the traces need to be pre-processed before applying TVLA [GJJR11]. A usual heuristic for this purpose is to combine every d-tuple of random variables multiplicatively, as shown in Figure 2. This produces the pre-processed trace sets $\boldsymbol{X'}$, $\boldsymbol{Y'}$ which we define as

$$\boldsymbol{X'} = \left\{ \prod_{i \in \mathbb{T}} X_i, \quad \forall\, \mathbb{T} \overset{d}{\subset} \{1, \ldots, n_l\} \right\}, \qquad \boldsymbol{Y'} = \left\{ \prod_{i \in \mathbb{T}} Y_i, \quad \forall\, \mathbb{T} \overset{d}{\subset} \{1, \ldots, n_l\} \right\},$$

for all considered sets of tuple indices $\mathbb{T}$ with cardinality d. Leakages can only be detected if such a d-tuple contains information about all d shares of the targeted sensitive variable. If there are dth-order leakages, the means of $\boldsymbol{X'}$ and $\boldsymbol{Y'}$ (resp., the dth-order moments of $\boldsymbol{X}$ and $\boldsymbol{Y}$) are different and TVLA will conclude the existence of leakages given enough samples. Note that sometimes, the centralized (mean-free) or standardized (normalized by standard deviation) moments are used [SM16]. Note also that samples pre-processed in this way are not Gaussian, and therefore the heuristic of testing the sample means relies on the central limit theorem in this case (which we briefly discuss in conclusions).

In our simulated experiments, we will only generate first-order leakages and rather indirectly emulate the effect of this pre-processing by adapting the simulation parameters (e.g., the number of samples and density of informative points) – which allows simpler interpretation. In our practical experiments, we will rely on the centered-product to map the higher-order leakages to the mean for all considered tests.

Note that for closed-source designs, this pre-processing quickly becomes very complex, since the evaluator does not know which d variables need to be combined to detect leakages.

Thus, it is necessary to check all possible $n'_l(d)$ d-tuples[6], with

$$n'_l(d) = \binom{n_l + d - 1}{d} = \frac{(n_l + d - 1)!}{d!(n_l - 1)!}, \tag{2}$$

which we can bound with

$$\frac{n_l^d}{d!} \leq n'_l(d) \leq \frac{(n_l + d)^d}{d!}.$$

This exponential increase strongly affects the density (i.e., the ratio of informative vs. non-informative tuples) which is an important parameter in our later evaluations and decreases exponentially in d due to the pre-processing.

2 Multi-Tuple Detection

Welch's t-test as applied in the TVLA methodology exploits only the difference in the means of two paired random variables (X_i, Y_i) as stated in its testing hypotheses (cf. Equation (1)). Therefore, it does not take full advantage of jointly testing all the variables in the vectors $(\boldsymbol{X}, \boldsymbol{Y})$, and instead runs n_l separate tests in the hope that at least one of them gives large enough confidence to reject the null hypothesis. By taking an analogy with distance between two vectors in an n_l-dimensional space, this procedure only looks at the distances between all the dimensions separately.

As a solution to this problem, we propose to use *multi-tuple* tests for leakage detection. Instead of n_l separate tests, the multi-tuple test statistics are derived from the difference of multiple paired variables jointly and, therefore, they can potentially detect leakages with fewer measurements than Welch's t-test. For example, if the joint difference of multiple paired variables shows leakages while there is not enough confidence for each pair individually. Instead of Equation (1), such a multi-tuple detection methodologies consider the following testing hypotheses

$$H_0: \quad \mu_{\boldsymbol{X}} = \mu_{\boldsymbol{Y}}, \qquad H_a: \quad \mu_{\boldsymbol{X}} \neq \mu_{\boldsymbol{Y}}. \tag{3}$$

H_0 assumes that there is no difference of means between the two vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, whereas H_a corresponds to the presence of leakages and states that there is at least one set of paired variables that exhibits a difference of means. In the following, we first introduce Hotelling's T^2-test, before proposing a D-test that, under some assumptions, provides the same testing power as Hotelling while reducing its computational complexity.

2.1 Hotelling's T^2-Test

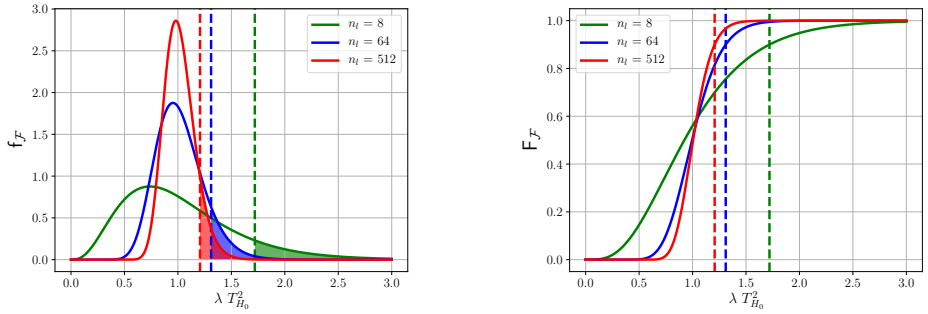
Hotelling's T^2-test is a natural candidate for our use case because it is inherently multi-tuple [Hot31] and corresponds to the statistical hypotheses described in Equation (3). The test statistic T^2 is computed according to

$$T^2 = \frac{N_{\boldsymbol{X}} N_{\boldsymbol{Y}}}{N_{\boldsymbol{X}} + N_{\boldsymbol{Y}}} (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})^\intercal \boldsymbol{S}^{-1} (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}),$$

and requires the estimated means $(\bar{\boldsymbol{X}}, \bar{\boldsymbol{Y}})$ of both random vectors and their estimated pooled covariance matrix $\boldsymbol{S}$ defined as

$$\boldsymbol{S} = \frac{\sum_{i=1}^{N_{\boldsymbol{X}}} (\boldsymbol{x_i} - \bar{\boldsymbol{X}})(\boldsymbol{x_i} - \bar{\boldsymbol{X}})^\intercal + \sum_{i=1}^{N_{\boldsymbol{Y}}} (\boldsymbol{y_i} - \bar{\boldsymbol{Y}})(\boldsymbol{y_i} - \bar{\boldsymbol{Y}})^\intercal}{(N_{\boldsymbol{X}} + N_{\boldsymbol{Y}} - 2)}. \tag{4}$$

[6] Other approaches than this pre-processing exist to deal with masked implementations, such as projection pursuits or dimensionality reductions [DSV+15, CDP16], but they are more applicable in an attack-based evaluation than in a detection-based one we consider here.



(a) Probability density function (b) Cumulative distribution function

Figure 3: The PDF and CDF of different Fisher distributions with varying n_l and their corresponding threshold $th_\alpha^{T^2}$ (dashed lines) for $\alpha = 0.1$ and $N_{\boldsymbol{X}} = N_{\boldsymbol{Y}} = 64$.

Hotelling's T^2-test assumes that the covariance of both random vectors $\boldsymbol{X}$, $\boldsymbol{Y}$ is the same and pools their estimates to achieve better accuracy. However, small differences in the covariances can be tolerated by the test without strongly affecting its performance (see next).[7] Under the null hypothesis and up to some multiplicative factor λ, this test statistic follows a Fisher distribution $\mathcal{F}$ parameterized with $(n_l, N_{\boldsymbol{X}} + N_{\boldsymbol{Y}} - 2)$:

$$\frac{(N_{\boldsymbol{X}} + N_{\boldsymbol{Y}} - 1 - n_l)}{(N_{\boldsymbol{X}} + N_{\boldsymbol{Y}} - 2)n_l} T^2_{H_0} = \lambda\, T^2_{H_0} \sim \mathcal{F}(n_l, N_{\boldsymbol{X}} + N_{\boldsymbol{Y}} - 2).$$

In contrast to Welch's t-test, the form of this distribution does not only depend on the number of measurements $(N_{\boldsymbol{X}}, N_{\boldsymbol{Y}})$, but also on the number of samples in a trace n_l. Based on this, the p-value and a detection threshold for T^2 can be computed as

$$p = 1 - \mathsf{F}_{\mathcal{F}}(\lambda T^2), \qquad th_\alpha^{T^2} = \mathsf{F}^{-1}_{\mathcal{F}}(1-\alpha)/\lambda,$$

where $\mathsf{F}_{\mathcal{F}}$ denotes the CDF of the Fisher distribution (cf. Appendix A). The threshold is, therefore, a function of the number of available measurements, the trace length, and the desired false positive probability α. This dependency is depicted in Figure 3. For fixed (exemplary) parameters $\alpha = 0.1$ and $N_{\boldsymbol{X}} = N_{\boldsymbol{Y}} = 64$, we computed the PDF's and CDF's of the Fisher distribution for varying n_l. It is noticeable that the shape of the distributions is strongly affected by the length of the traces n_l and, therefore, the threshold $th_\alpha^{T^2}$ (represented by the dashed lines) also changes accordingly.

Under the alternative hypothesis, the test statistic follows a non-central Fisher distribution

$$\lambda\, T^2_{H_a} \sim \mathcal{F}(n_l, N_{\boldsymbol{X}} + N_{\boldsymbol{Y}} - 2; \delta),$$

with the effect size δ as an additional parameter which is defined as

$$\delta = \frac{N_{\boldsymbol{X}} N_{\boldsymbol{Y}}}{N_{\boldsymbol{X}} + N_{\boldsymbol{Y}}} (\mu_{\boldsymbol{X}} - \mu_{\boldsymbol{Y}})^\intercal \Sigma^{-1} (\mu_{\boldsymbol{X}} - \mu_{\boldsymbol{Y}}).$$

Here, δ represents the distance between the null and the alternative hypothesis for any covariance, means and number of measurements. Based on this distribution, we can compute the false negative probability β as [Ren98]

$$\beta = \Pr[\lambda\, th_\alpha^{T^2} > \mathcal{F}(n_l, N_{\boldsymbol{X}} + N_{\boldsymbol{Y}} - 2; \delta) | \mu_{\boldsymbol{X}} \neq \mu_{\boldsymbol{Y}}].$$

This gives the probability that the observed test statistic T^2 is smaller than the detection threshold even though the mean difference between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is unequal zero (i.e., H_a).

[7] Note that there are versions of Hotelling's T^2-test which do not rely on this assumption. However, they have strongly increased data complexity, and thus are not considered in the following.

Leakage Detection with Hotelling's T^2-Test. We propose to follow the same procedure as TVLA (and all of its instantiations) to generate the measurements, and replacing the n_l Welch's t-test with one Hotelling T^2-test as depicted in Figure 1. This comes with two advantages that increase the detection rate in certain scenarios. First, Hotelling is multi-tuple and, second, it is not based on the independent signal assumption. Nevertheless, this improved detection rate comes at the cost of the estimation and inversion of the pooled covariance matrix $\boldsymbol{S}$ (Equation (4)). Since $\boldsymbol{S}$ is an $n_l \times n_l$ matrix, storing and computing on it is roughly quadratic in the trace length. Taking into account that the trace length may grow exponentially with the statistical order of the test, performing a T^2-test might not be computationally feasible for very long traces at high orders.

Note that despite Hotelling's T^2-test returns a single statistic, it does not lead to a complete loss of intuition on the leaking points in a measurement trace. Indeed, it still requires computing mean vectors $\bar{X}$ and $\bar{Y}$ which provide such an indication, yet with less confidence due to the faster detection.

Note also that even though the standard Hotelling's T^2-test assumes equal covariances between the sets, it can still be applied in the context of leakage detection for which this is not necessarily true for every scenario (e.g., in the case of masked implementations). As pointed out in [Wy92], the test is robust for unequal covariances under two conditions: *(i)* (N_X, N_Y) are large and equal, and *(ii)* N/n_l is large. The first condition (i.e., the equality in size of the two sets) can be trivially forced by evaluators. The second condition is typically fulfilled for protected designs which are the main target for leakage detection, for which (hundreds of) thousands of traces are necessary to detect. Indeed, n_l is anyway limited by the computational capabilities of the evaluator. For example, in our following experiments, we restricted it 1000 for computational reasons. Under this restriction, we argue that Hotelling's T^2-test can be utilized for leakage detection in a statistically sound manner. We will detail later in the paper how such a detection test can be combined with a min-p strategy in order to deal with longer traces.

2.2 D-Test for Independent Signals

As noted, Hotelling's T^2-test can offer improved detection rates at increased computational cost. To overcome the covariance issue, we propose a specialization of Hotelling's T^2-test denoted as *Diagonal-Test* (D-test). It is based on Hotelling but does assume independence between the signals, i.e., all points in a trace are independent. Therefore, we trade one of the two advantages (multi-tuple + no dependency assumption) of the T^2-test to remove its main efficiency drawback.

In particular, we assume that the covariance matrix is diagonal ($\forall i \neq j,\ \sigma_{X_i,Y_j} = 0$) which is true if the signals are independent. Thus, instead of computing the complete $n_l \times n_l$ matrix $\boldsymbol{S}$, our test requires only the estimation of the n_l diagonal entries. This allows us to rewrite the test statistic as

$$\begin{aligned} D &= \frac{N_{\boldsymbol{X}} N_{\boldsymbol{Y}}}{N_{\boldsymbol{X}} + N_{\boldsymbol{Y}}} \sum_{i=1}^{n_l} \frac{(\bar{X}_i - \bar{Y}_i)^2}{\boldsymbol{S}_{i,i}} \\ &= \sum_{i=1}^{n_l} \frac{(\bar{X}_i - \bar{Y}_i)^2}{\frac{s^2_{X_i}}{N_{\boldsymbol{X}}} + \frac{s^2_{Y_i}}{N_{\boldsymbol{Y}}}} \\ &= \sum_{i=1}^{n_l} t_i^2, \end{aligned} \tag{5}$$

by also replacing the equal covariance assumption of Hotelling with the extension from Welch's t-test to allow different variances. Therefore, just like classical TVLA, our test requires only the estimation of the mean and variance of each variable pair (X_i, Y_i)

separately. It is noticeable, that the D-test is equivalent to TVLA for $n_l = 1$ and to Hotelling for independent signals and equal covariances.

For a sufficient number of measurement, the t-statistics (Equation (5)) follow a normal distribution meaning that under this condition D follows a χ^2-distribution with n_l degrees of freedom. As before, the p-value and detection threshold can be computed as

$$p = 1 - \mathsf{F}_{\chi^2}(D), \qquad th_\alpha^D = F_{\chi^2}^{-1}(1-\alpha),$$

where F_{χ^2} denotes the CDF of the χ^2-distribution (cf. Appendix A). Note that a similar approach was proposed in [WGS06] targeting small sample sizes ($N_{\boldsymbol{X}}$ and $N_{\boldsymbol{Y}}$) which avoids the hypothesis that the t_i's are normally distributed.

Leakage Detection with the $D-Test$. As in the previous sections, the traces are generated according to the TVLA methodology. The n_l Welch's t-tests are then replaced by one D-test. We note that our test statistic can be written as the sum of squared t_i values obtained by different t-tests (cf. Equation (5)). Therefore, in contrast to Hotelling, our D-test has only a linear complexity in the trace length and can for example be performed in parallel to Welch's t-test with minimal overhead (cf. Figure 1) making it feasible in all scenarios in which the t-test can be applied (assuming independent signals). Based on Equation (5), it is noticeable that the performance of the D-test suffers from adding non-informative sampling points. This is indeed true, but applies to all considered leakage detection tests which process multiple points (including Welch's test with min-p). In our simulations, we will show that the different tests are affected to varying degrees to the addition of noisy and informative points (Figure 6).

3 Simulated Experiments

The underlying idea of multi-tuple detection is to take advantage of multiple joint differences in the traces to reduce the data complexity required for detection. We next use simulated experiments to evaluate the gains that can be expected.

In this respect, our starting observation is that (as already observed with the higher criticism approach), these gains may depend on the target implementation. For example, it may happen that not all samples (or tuples) considered in a detection procedure are informative. These non-informative points might thwart the estimation of the multi-tuple test statistic by increasing the noise and, thus, cancel the aforementioned advantage.

Besides, both Welch's t-test with min-p and the D-test assume that the signals are independent, which can further impact the detection rate.

In order to fairly compare the different approaches, our simulations will therefore assess the influence of two main parameters on the detection performances:

1. *The density* refers to the ratio of (pre-processed) informative points in the traces.
2. *The dependency* is the level of deviation of the covariance matrix from diagonal.

As usual, simulations are useful since they allow us to thoroughly assess the influence of these parameters in a controlled environment. Besides the density and dependency (which carry the most important intuitions), we also consider the SNR, β, and trace length n_l as additional parameters influencing our results (in a way that is essentially similar to what was observed in previous works on leakage detection).

In the following, we first introduce our simulation framework and a methodology to estimate the success rate of Welch's t-test, Hotelling's T^2-test, and the D-test. Secondly, we use simulations with independent leakages to show that for dense traces (i.e., when a large proportion of the points in the traces shows evidence of leakages) Hotelling and

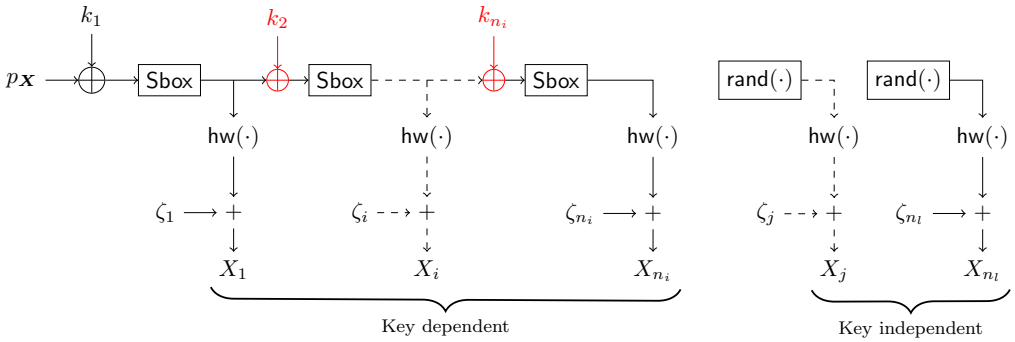


Figure 4: Simulation framework.

the D-test provide significantly faster detection than Welch's t-test using the min-p approach. By contrast, we show the superiority of Welch's t-test with min-p in low-density settings. Thirdly, we highlight the dependent signal issue and its effect on leakage detection methodologies. Our results indicate that it is a critical aspect to interpret detection results in a statistically sound manner. Concretely, they imply that only Hotelling's T^2-test reliably provides the expected detection rate, while the other tests are either too conservative or optimistic in case signal independence is incorrectly assumed.

3.1 Simulation Framework

In our simulations, we use the common assumption of a Hamming weight leakage model with additive Gaussian noise. To this end, we initially pick two fixed 8-bit inputs $p_{\boldsymbol{X}}$ and $p_{\boldsymbol{Y}}$.[8] These are run through a round-based structure consisting of the addition of an independent round key k_i followed by a bijective function $\mathsf{Sbox} : \{0,1\}^8 \mapsto \{0,1\}^8$ as depicted in Figure 4. The value after each function is considered as an informative leakage sample, to which we apply the Hamming weight function $\mathsf{hw}(.)$ and add noise $\zeta \sim \mathcal{N}(\mathbf{0}, \Sigma)$. The noise is sampled from a multivariate Gaussian distribution with the zero vector $\mathbf{0}$ as means and Σ as its covariance matrix. Depending on the scenario, this matrix is chosen to be either diagonal or with varying levels of dependency between its variables (cf. Figure 7). The diagonal is set to σ^2 depending on the desired SNR [Man04], which is equal to $2/\sigma^2$ assuming an 8-bit Hamming weight leakage model and equal covariances for both sets. In order to ensure the independence of the manipulated values independently of the Sbox, a key addition with independent keys $k_i \leftarrow \{0,1\}^8$ is performed at each round.

Density. The density in our simulations is controlled by repeatedly adding random values to the trace, which generates leakage samples that are independent of the key and plaintext and are therefore non-informative. In total, we generate traces of length $n_l = n_i + n_o$, where n_i denotes the number of informative points, that are padded with n_o points of noise. Based on this setting, we define the density of our simulated measurements as

$$\phi = \frac{n_i}{n_l},$$

where $\phi = 1$ denotes that every point is sensitive while $\phi = 0$ indicates no leakages. Note that concretely, a low density typically corresponds to the case of a masked implementation, as per remark in Section 1.2 (e.g., the masked AES implementation in our software case study has a density of $\phi = 0.0001$, while the masked hardware design exhibits a density of $\phi = 0.027$). As already mentioned in Section 1.2, for simplicity our simulations use

[8] Note that these are picked as $p_{\boldsymbol{X}} \neq p_{\boldsymbol{Y}}$ to ensure leakage.

this parameter together with the SNR to reflect more or less protected (possibly masked) implementations. We do not directly generate higher-order leakages nor pre-process the traces in this section which allows easier interpretation (since all samples then remain Gaussian-distributed). By contrast, the real experiments in the next section consider concrete masked implementations for which we pre-process the traces to map higher-order leakages to the mean. From the detection viewpoint, it is mostly the SNR and density of the traces that respectively impact the data and time complexity of the detection task.

Detection Rate. In order to compare the performances of the various detection methods in a given setting, the probability to correctly conclude the presence of leakages while the device is actually leaking is used. This detection rate is directly linked to the false negative probability according to the equation

$$\Pr[T^2_{H_a} > th^{T^2}_{alpha}] = 1 - \beta = 1 - \Pr[\lambda\ th^{T^2}_{\alpha} > \mathcal{F}(n_l, N_{\boldsymbol{X}} + N_{\boldsymbol{Y}} - 2; \delta)]$$

for Hotelling T^2-test (resp., for D-test with diagonal Σ), and

$$\Pr[\max(|t_i|) > |th^t_{n_l,\alpha}|] = 1 - \beta = 1 - \Pi^{n_l}_{i=1} \Pr[th^t_{n_l,\alpha} > |\mathcal{T}(v_i; \delta_i)|]. \tag{6}$$

for Welch's t-test. While the false negative probability (β) is more common in the statistics community, we sometimes use the detection rate because it is intuitive in the side-channel analysis context. If a method applied to a leaking device concludes that it is leaking in 99% of the cases, then its detection rate is equal to 0.99 and $\beta = 0.01$.

Evaluation Method. Our simulations are performed in three steps. First, the desired false positive probability α is set as well as the other implementation parameters (density ϕ, trace length n_l, covariance matrix Σ and SNR). If not mentioned, α is equal to the frequently assumed 10^{-5}. Secondly, the desired detection rate is shown. In the following, it is either set to 0.999 ($\beta = 10^{-3}$) or to 0.9999 ($\beta = 10^{-4}$). Finally, the inputs for the two sets are selected at random, the mean vectors $\mu_{\boldsymbol{X}}$ and $\mu_{\boldsymbol{Y}}$ are obtained and the number of measurement (N) needed to achieve the targeted detection rate is computed.[9] This last step is repeated and its outputs are averaged to remove input dependencies.

3.2 Simulation with Independent Signal

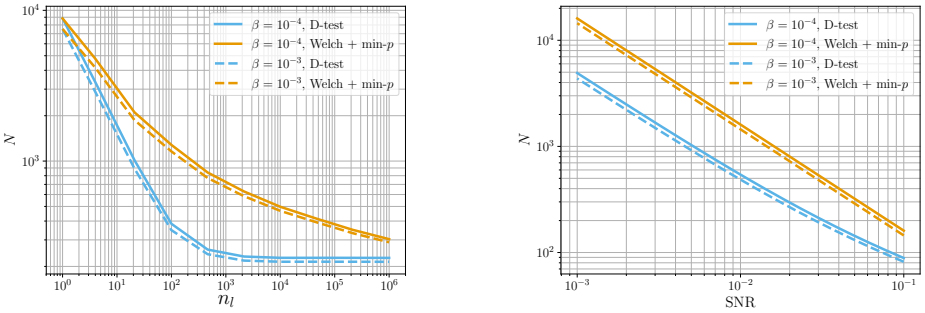
We start by comparing the tests assuming independent signals. To this end, only the diagonal elements of the covariance matrix are set to the desired noise level, while we fix the remaining elements to zero (cf. Figure 7a).

As noted before, Hotelling's T^2-test and the D-test are equivalent in this setting, and we will only consider the latter for the comparison with Welch's t-test and the min-p approach. We primarily focus on the impact of the trace length (n_l) and the density (ϕ) on the data complexity (N) at a fixed detection rate ($1 - \beta$).

Influence of Multiple Tuples. First, we set the density to $\phi = 1$ (i.e., all points in a trace are informative) and observe the influence of the remaining simulation parameters (SNR, n_l) on the detection rate. The results are given in Figure 5 where the vertical axis denotes the number of measurements N, while the horizontal axis shows an increasing n_l for fixed SNR (on the left) and an increasing SNR for fixed n_l (on the right).

The most noticeable influence on the detection complexity stems from the number of points in a trace n_l. While both tests (which are equivalent for $n_l = 1$) benefit from increasing the trace length, the D-test improves more than Welch's t-test with min-p, as shown in Figure 5a. This behavior is expected given that the D-test combines the

[9] Where $N = N_{\boldsymbol{X}} + N_{\boldsymbol{Y}}$ and $N_{\boldsymbol{X}} = N_{\boldsymbol{Y}}$.



(a) SNR $= 0.01$ (b) $n_l = 64$

Figure 5: Number of traces N required to detect for a fixed false negative rate β on simulated traces with a density of $\phi = 1$, and varying trace length n_l and SNR.

differences of all points into a single test statistic while the min-p approach only looks for one worst-case. Therefore, for implementations which leak with such a high density (e.g., unprotected or parallel masked hardware designs), multi-tuple tests can detect (much) faster than Welch's t-test in case of sufficiently long traces. In our example, for $n_l = 10^3$, Welch's t-test with min-p requires 698 measurements to achieve a detection rate of 0.999, while the D-test needs only 229 measurements (which corresponds to an improvement by a factor 3). For $n_l = 10^6$, Welch's t-test requires 303 traces while D-test only 205.

The other parameters affect the detection rate of both tests as expected. Decreasing β basically means that in more experiments the null hypothesis will be rejected, implying the need for more measurements. The same can be observed for changes in the SNR given that more noise/less signal also increases N (as depicted in Figure 5b).

Influence of the Density. We next study the influence of the density at a given noise level as shown in Figure 6. Now, the horizontal axis shows the density ϕ, while the vertical axis is still the number of traces required to detect N.

For a density equal to one, we observe that the D-test requires fewer traces than Welch's t-test with min-p, as expected from Figure 5. However, it is noticeable that both methods suffer from decreasing the density in the measurements.

For the D-test, the decreased density increases the number of zero elements in the vector $\mu_{\boldsymbol{X}} - \mu_{\boldsymbol{Y}}$. Therefore, the effect size δ decreases as well as the distance between the test statistic distribution under the zero and alternative hypothesis. This must be counterbalanced by increasing the number of samples to keep a constant δ and so the desired β. For Welch's t-test with min-p, decreasing the density reduces the number of t_i's which

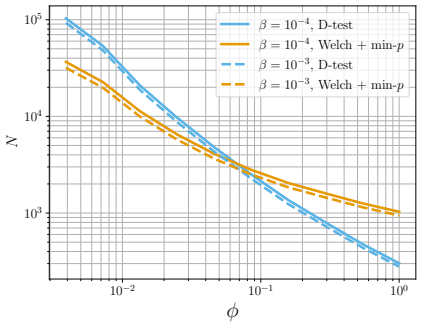
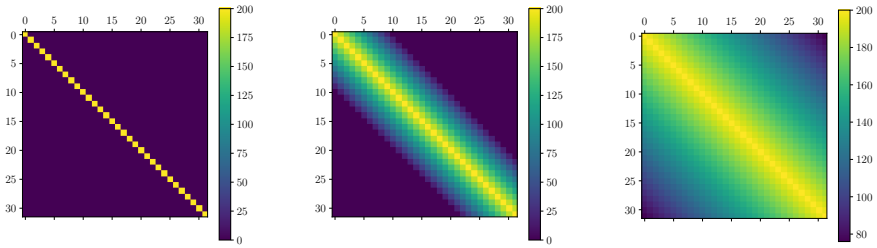


Figure 6: Number of traces N required to detect for a fixed false negative rate β on simulated traces with SNR $= 0.01$, and varying density ϕ.



(a) Σ_1 (b) Σ_2 (c) Σ_3

Figure 7: Covariance matrices considered in our experiments (for $n_l = 32$).

can potentially show leakages and, therefore, reduces the overall chance to detect leakages and increases the number of required measurements. More precisely, in Equation (6), an increasing proportion of δ_i's is equal to zero, making the overall detection rate smaller at a fixed number of measurements. Nevertheless, Figure 6 clearly shows that Welch's t-test with min-p suffers less from the reduced density than the multi-tuple test. For small densities, it even outperforms the D-test. Since the density is a priori not known by the evaluator in a closed-source setting, it is not possible to predict which test performs better a priori. Instead, both Welch's t-test with min-p and the D-test should be run in parallel, with negligible overhead as explained in Subsection 2.2. For high densities, the D-test will typically detect faster. For low densities, the opposite conclusion will hold.

We note that we did not add the results of the higher criticism approach in our figures for readability purposes. However, they essentially detect slower than the multi-tuple test in case of dense traces, and do not beat Welch's t-test with min-p otherwise, making it a less interesting alternative if the two previous tests are run in conjunction.

3.3 Simulation with Dependent Signal

We now explore the dependent signal issue and its effect on Welch's t-test with min-p, Hotelling's T^2-test, and the D-test. In particular, we consider three different noise covariance matrices Σ_1, Σ_2, Σ_3 as depicted in Figure 7. Σ_1 is a diagonal matrix like in the previous section (i.e., it corresponds to independent signals), while Σ_2 and Σ_3 represent increasing rates of dependency between the variables of $\boldsymbol{X}$ and $\boldsymbol{Y}$. Σ_2 maps a situation where a variable X_i (resp., Y_i) is only correlated with a few adjacent variables. In practice, this would relate to the case where the samples inside a clock cycle exhibit a strong dependency on each other, e.g., as in our hardware case study (cf. Figure 10b). For Σ_3 nearly all points are correlated to a certain degree which corresponds to dependencies exceeding clock cycle boundaries, e.g., as in our open-source software case study (cf. Figure 11b).[10] In the following, we first show the influence of a non-diagonal covariance matrix on the distributions of the test statistics. Next, we highlight that it can lead to a wrong estimation of α, hence to a false sense of security based on sub-optimal detection performances. Thirdly, we confirm our intuitions with simulated experiments using Σ_2 and Σ_3.

Impact on the Test Statistics. Given independent signals, the test statistics follow well-defined distributions under the null hypothesis. To assess the effect of signal dependencies on these distributions, we generate $100,000$ sets of $4,000$ traces for each of the aforementioned covariance matrices and apply every leakage detection test. The results are in Figure 8.

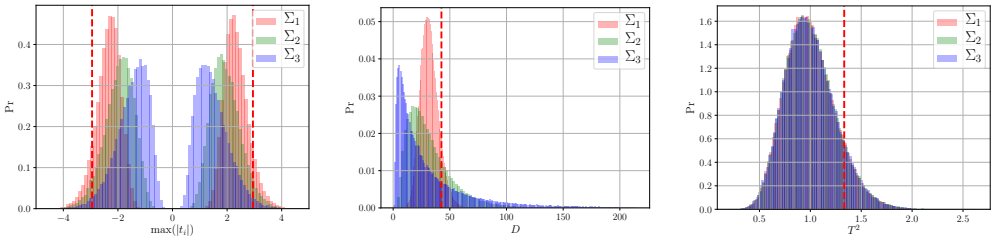
The distribution of Welch's t-test with min-p statistic is composed of two lobes (cf. Figure 8a) since only the largest $|t_i|$ from n_l different Welch's t-test is kept. For Σ_1, we observe the expected distribution (in red) assuming independent signals. However, by

[10] Further descriptions of each matrix are given in Appendix B.

adding slight dependencies in the traces as in the Σ_2 case, the test statistic distribution (in green) is pushed towards the center. This trend is amplified with Σ_3 (in blue).

An equivalent behavior is observed for the D-test test statistic (cf. Figure 8b). The distributions for Σ_2 and Σ_3 differ from the one assuming independence. In particular, stronger dependencies result in a higher spread of the test statistic distribution.

By contrast, Hotelling's T^2-test does not rely on the independence signal assumption and, therefore, the test statistic distributions remain constant in the different settings.



(a) Welch + min-p. (b) D-test. (c) Hotelling's T^2-test.

Figure 8: Null distributions with dependent noise and detection thresholds (dashed lines).

Risks of Incorrect Interpretation. As mentioned in the previous section, the detection thresholds as well as the p-values used in the TVLA methodology are derived from the distributions of the test statistics under the null hypothesis. For the tests relying on an independent signal assumption, these values are derived from the distribution corresponding to Σ_1, e.g., as depicted by the dashed line in Figure 8 for $\alpha = 0.1$.[11] This becomes problematic once this assumption is violated and the observed distributions of the test statistics differ from the one assuming independent signals. It would for example imply that the threshold (the dashed lines) would remain unchanged in the Σ_2 and Σ_3 cases, despite their distributions differing significantly from the Σ_1 one.

Since the detection threshold is used to conclude the presence of leakages, using a faulty threshold (e.g., for Σ_2 and Σ_3) may lead to an incorrect interpretation of the test results. In order to illustrate this risk, we estimated the empirical false positive rate $\hat{\alpha}$ corresponding to detection with the three covariance matrices in simulations. This $\hat{\alpha}$ corresponds to the sampled probability that the test statistic is larger than the threshold chosen for a given α. In our case, a correct test is expected to have an $\hat{\alpha}$ that is close to $\alpha = 0.1$. The results are given in Table 1.

Table 1: Empirical false positive rate $\hat{\alpha}$ estimated from $100,000$ experiments.

	Σ_1	Σ_2	Σ_3
Welch's t-test with min-p	0.099	0.051	0.018
D-test	0.101	0.236	0.237
Hotelling's T^2-test	0.101	0.099	0.102
Hotelling's T^2-test with min-p	0.101	0.095	0.09

In the Σ_1 case, the $\hat{\alpha}$ of all tests converges to the expected value since the independent signal assumptions is indeed fulfilled. By contrast, for Σ_2 and Σ_3, the derived threshold and p-value are no longer correct for Welch's t-test with min-p and the D-test, as illustrated in the second and the third columns of the table.

[11] This value of α is chosen to simplify the later estimation of an empirical false positive rate $\hat{\alpha}$. Further note that Figure 8 depicts the zero distributions of the test statistics (i.e., no leakage) and, thus, only a proportion of α of the test statistics should exceed the threshold.

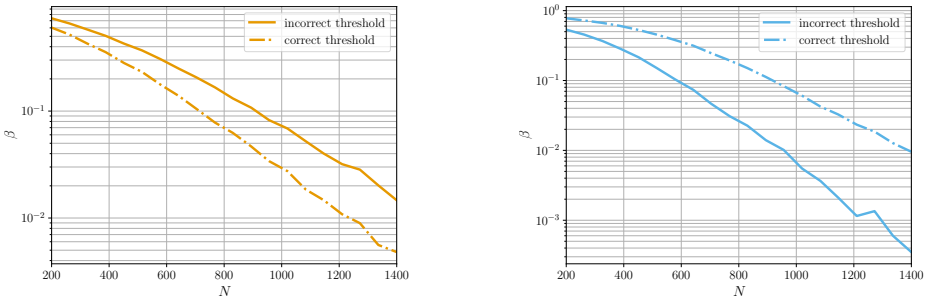
For Welch's t-test with min-p, the estimated $\hat{\alpha}$ converges to values that are smaller than $\alpha = 0.1$ when increasing the dependencies. This confirms the intuition from Figure 8a, where the distributions are pushed towards the center. It results in a smaller fraction of the estimated test statistics exceeding the threshold, hence a lower $\hat{\alpha}$.

An opposite behavior is exhibited by the D-test. For increasingly dependent signals, the $\hat{\alpha}$ becomes larger than expected, which can also be observed in Figure 8b where the area right of the threshold is larger for Σ_2 and Σ_3.

Hence, Hotelling's T^2-test is the only candidate that gives the correct false positive rate even for dependent signals, since it does not rely on the independent signal assumption.

Intuitively, these results indicate that Welch's t-test with min-p may behave too *conservatively* if incorrectly assuming independence. In such cases, it will declare the presence of leakages with a confidence level higher than expected by the evaluator, leading to a reduced detection rate (increased β) as shown in Figure 9a. In this figure, we estimate the correct detection thresholds for Σ_2 by sampling which is feasible since we choose $\alpha = 0.1$, and use them to compare the detection rates with faulty and correct thresholds. It is obvious that the performances of the test suffer significantly from wrongly assuming independence. This effect is easily explained by the most extreme case of fully dependent signal and noise (i.e., if all entries of Σ are the same). In this scenario, all n_l tests would produce the same test statistic and should, therefore, be reduced to only one test. Nevertheless, if this test statistic is still compared to a threshold assuming n_l independent tests, it will result in an overestimation of the security, i.e., the evaluator may not detect existing leakages because of the erroneous threshold.

Interestingly, since the D-test shows the opposite trend (cf. Figure 9b) and may be too *optimistic* (i.e., declare leakages too fast) if incorrectly assuming independence, the combination of these two tests may be used to heuristically bound the detection rate if Hotelling's T^2-test cannot be launched due to computational reasons.[12]



(a) Welch's t-test with min-p. (b) D-test.

Figure 9: Detection rate errors due to incorrect signal independence assumption based on simulated traces with the covariance matrix Σ_2 and SNR = 0.004.

Alternatively, Hotelling's T^2-test can be combined with min-p to allow more flexibility. For example, the evaluation of unprotected asymmetric primitives might result in very long and dense traces. In this case, it is impossible to compute the whole covariance matrix. Instead, the whole trace is partitioned into smaller sections and Hotelling's T^2-test is applied to each separately. This results in multiple test statistics which we combine using min-p. Since the separate tests are not necessarily independent, the false positive rate can be negatively affected as with Welch's t-test. To verify this, we repeat the experiment of Table 1 by generating 10× longer traces and splitting them into 10 sections. The results are given in the last row. It is noticeable that, as for Welch's t-test, the hybrid approach

[12] The latter remains heuristic since the conservative and pessimistic nature of these two tests is only concluded based on experiments in practically-relevant yet limited scenarios.

suffers from a reduced $\hat{\alpha}$, however, to a smaller degree. This strongly depends on the number of considered sections, e.g., the most extreme case would be conducting a test for each sample point separately making it equivalent to the Welch's t-test with min-p. .

4 Practical Experiments

Up to now, we only considered simulated experiments for which we precisely controlled the leakages' density and dependency. To evaluate the performance of the tests in an actual detection scenario, we additionally conducted practical experiments using masked implementations of the AES in hardware and software. For both cases, we initially conduct the tests on the complete (pre-processed) traces before applying filtering to improve detection performances. In particular, we use closed-source filtering (i.e., peak extraction) for the hardware case study and open-source filtering for the software AES.

4.1 First Case Study: Masking in Hardware

We asses the leakages of a protected hardware design. Masking schemes in this scenario usually process the shares of an encoding in parallel, i.e., the leakages are not spread over multiple cycles, but contained in one. Therefore, to perform a dth-order evaluation, it is usually not necessary to apply the full pre-processing (cf. Figure 2), but instead it is sufficient to raise each point of the original traces separately to the dth power:

$$\boldsymbol{X}' = \left\{ (X_i)^d, \quad \forall\, i \in \{1, \ldots, n_l\} \right\}, \qquad \boldsymbol{Y}' = \left\{ (Y_i)^d, \quad \forall\, i \in \{1, \ldots, n_l\} \right\}.$$

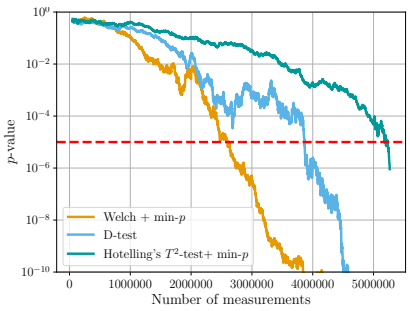
Architecture. As a case study, we use the publicly-available code from [GMK16], which implements AES protected with the Domain Oriented Masking (DOM) scheme.[13] We decided to synthetize the core for two shares which is expected to provide protection against first-order attacks. The main part of the design is the protected Sbox which is implemented in eight pipeline stages. In total, the circuit produces one encryption in 246 cycles and requires 18 bits of randomness per cycle to ensure first-order security. We generate this randomness using an unprotected AES running in counter mode initialized with a random seed. More precisely, before encrypting a RAM FIFO is filled with the total amount of randomness required for one encryption and the PRNG is deactivated during the masked encryption to minimize the noise in the measurements. For more details on the architecture, we refer the reader to the original publications [GMK16, GMK17].

Measurement Setup. The targeted architecture is running at 4[MHz] on a `Sakura-X` board containing a Kintex-7 FPGA. The measurements are taken with a passive probe placed between the power supply and the target FPGA. This power signal is sampled with a Picoscope 5000 oscilloscope at a sampling frequency of 500[MS/s]. In order to reduce the noise, we repeat the experiments for each randomness seed 20 times and average the results. The recorded traces are of length $n_l = 31,250$.

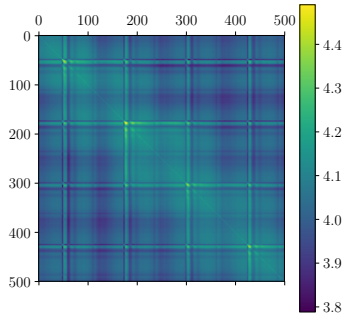
Results. Since the design is protected by masking with two shares, we need to pre-process the traces to evaluate the second-order leakages of the device. To this end, we make the traces mean-free and square each point in time.

We first compare the detection methods on the entire preprocessed traces. Since with $n_l = 31,250$ the traces are too long to invert the covariance matrix, we perform Hotelling's T^2-test with the min-p. The results are given in Figure 10a. As expected, all detection methods successfully declare the presence of leakages, i.e., their p-value exceeds

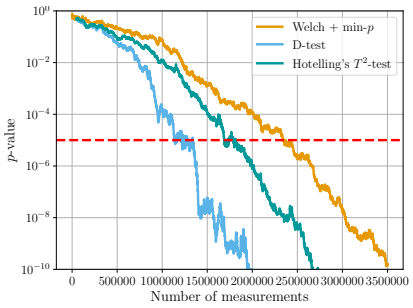
[13] The source code is available at https://github.com/hgrosz/aes-dom.



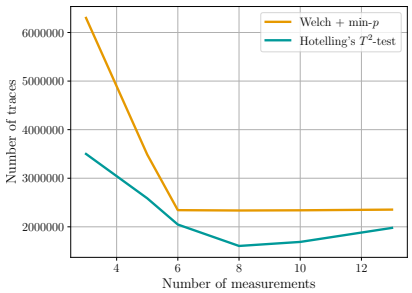
(a) p-values for entire traces.



(b) Estimated Covariance $\log_{10}(\boldsymbol{S})$ over 4 cycles.



(c) p-values with peak extraction ($l = 8$).



(d) Number of traces required.

Figure 10: Leakage detection results using the measurements of a masked hardware AES implementation protected by DOM with two shares.

the detection threshold (dashed line). It is noticeable, that Welch t-test combined with a min-p approach requires around $2.5 \cdot 10^6$ traces to detect leakages, while the D-test reaches the threshold for $4 \cdot 10^6$ measurements and the T^2-test with min-p around $5.2 \cdot 10^6$ traces, which indicates a low density. Indeed, only 824 points out of the $31,250$ exceed the threshold for the Welch's t-test, which means a density of around 0.027.

In order to further improve the detection performances, we then increased the density of the traces by filtering the sample points. One possibility which does not require knowledge of the implementation is peak extraction [MOP07]. It is based on the idea that most of the useful leakage samples are concentrated around the peaks of the clock cycles. In the following, we reduce the traces by choosing only l points within each cycle (selected such that they correspond to the power peak within a cycle).

Figure 10c contains the p-values for the three detection methods on reduced traces with $l = 8$, where Hotelling's T^2-test requires $1.75 \cdot 10^5$ traces to reach the threshold, the t-test $2.4 \cdot 10^6$, and the D-test $1.25 \cdot 10^6$. Interestingly, the t-test with min-p gets more improvements from longer traces than the increased density.

In Figure 10d, we show the average number of traces required in order to reach the $\alpha = 10^{-5}$ threshold with respect to l. On this figure, we observe that by increasing the number of points per cycle, and so n_l, both methods first improve, as expected from Figure 5a. However, by continuously increasing l, Hotelling's T^2-test requires more traces at some point. This can be explained by a reduced density which then results in a less efficient detection (Figure 6). More precisely, in this case the additional leakage samples correspond to lower logic activity and therefore have a lower signal.

In order to better understand our results, we additionally investigated the covariance matrix of our measurements. It is obvious from the heat map given in Figure 10b that it is not diagonal, meaning there are dependencies between the points of the traces. As

discussed in Subsection 3.3, this makes the interpretation of the results for Welch's t-test with min-p, the D-test, and Hotelling's T^2-test with min-p more challenging.

For filtered traces, we can interpret Welch's t-test's reduced performance as a combination of two factors. First, it does not benefit as much from multiple informative points as the multi-tuple tests. Second, it relies on the independent signal assumptions which is not fulfilled. We posit that the latter is also the reason why the D-test detects faster than Hotelling's T^2-test. As shown before, it may underestimate the security in case of dependent signals, leading to faster detection due to an incorrect threshold. Therefore, in this case Hotelling's T^2-test is the only candidate that provides the statistically correct results and detects leakages significantly faster than Welch's t-test with min-p. We also note that it is indeed (heuristically) bounded by the two other methods.

By contrast, for unfiltered traces, the density is significantly reduced. Therefore, the performances of both Hotelling's T^2-test and the D-test suffer, while Welch's t-test is only slightly hampered. In this scenario, Hotelling's T^2-test is less useful (as observed in simulations), while the other two still pose (heuristic) bound on the p-value.

4.2 Second Case Study: Masking in Software

We next validate our methodology in a masked software setting. In contrast with the previous hardware design, the shares are now processed serially, splitting their leakages over multiple points in the traces. In this case, open-source adversaries can use their implementation knowledge to test only the relevant dth-order tuples, while closed-source ones must test all the possible tuples. In the latter case the density is expected to rapidly decrease and the computational complexity increases.

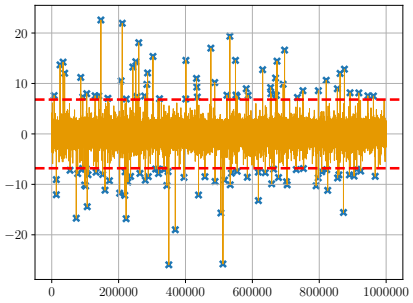
Architecture. The targeted architecture is an Atmel 8-bit design implementing the Rivain-Prouff masking scheme from [RP10] with three shares. More precisely a single Sbox is implemented in 2406[cycles] and requires 144 random bits. The randomness is generated before each encryption thanks to an unprotected AES reduced to 3 rounds. The obtained random bits are stored in memory and accessed during the AES encryption.

Measurement Setup. The measurement setup is a `ChipWhisperer-Lite` by `NewAE Technology Inc.`[14] The power signal is measured with a Picoscope 5000 at frequency of 20[MS/s]. The targeted device is an `ATxMEGA128d4` which is clocked at 7.37[MHz]. We recorded 16 Sboxes for a total of $n_l = 120,000$ leakage points.

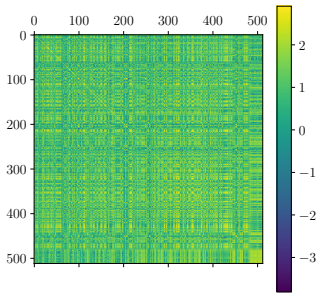
Results. Similar to the hardware case study, the covariance matrix of the measurements as given in Figure 11b is not diagonal and, therefore, indicates a dependency between the points of the traces. In the following, we examine both adversarial scenarios (closed- vs. open-source) and show the limitations of the tests in these settings.

In the first case, the adversary has only very limited knowledge of the target implementation, i.e., she is not able to select points-of-interest. Given that the implementation uses three shares, it should ideally protect against all first- and second-order attacks. However, as noted in [BGG+14], physical effects not considered in the model can reduce the security order. For our measurements, we came to the same conclusion and found second-order leakages. Therefore, in the following, we compare the detection tests at order two instead of three. Since the test of all $n'_l(2) = \binom{120,000+2-1}{2} > 10^{10}$ tuples exceeds the capabilities of our setup, we decided for the unfiltered detection to evaluate only $n'_l = 10^6$ randomly selected tuples within each Sbox. For traces of this size, the computation of the covariance matrix is still out of reach for our setup (cf. Appendix C for a discussion about the time complexity) and even the hybrid of Hotelling with min-p is not viable

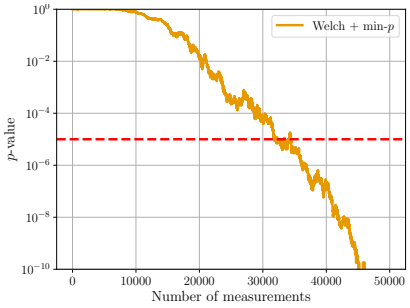
[14] http://newae.com/tools/chipwhisperer/.



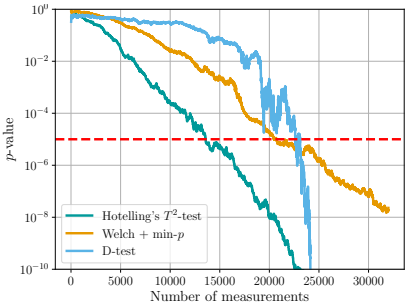
(a) Welch's t-tests' results obtained from pre-processed traces with $N = 100,000$.



(b) Estimated covariance matrix (in $\log_{10}(\boldsymbol{S})$ scale) computed on 512 tuples of interest.



(c) Closed-source p-values.



(d) Open-source p-values.

Figure 11: Leakage detection results: masked software AES implementation.

given the large number of costly matrix inversions needed to be done. Therefore, we only consider Welch's t-test with min-p and the D-test in this case, which benefit from a computational complexity that is linear in the trace length. The detection results are depicted in Figure 11c. It is noticeable that Welch's t-test with min-p outperforms the D-test due to the low density of the traces (the latter was not even able to detect).

Next, we emulated an open-source evaluator by using the Welch's t-test results (on averaged traces) to filter the trace points. The obtained t_i's of a single Sbox for $100,000$ measurements are depicted in Figure 11a, of which only 124 (i.e., the blue crosses in Figure 11a) out of 10^6 are above the detection threshold. This translates to a density of 0.0001 which is in line with the expectation that the density is severely decreased through exhaustive pre-processing. Based on this preliminary experiment, we select the points of interest in each Sbox with the lowest p-values and keep only $n_l = 3,200$ ones.

The results for the three methods with these filtered tuples are given in Figure 11d. As expected, the detection rate of all three tests improves considerably due to the higher density. Furthermore, due to dependencies, Welch's t-test with min-p overestimates the number of traces compared to the Hotelling's T^2-test. An important observation is that, even though the combination of Hotelling's T^2-test with the min-p approach leads to overestimate the number of traces required, we obtain a significant gain compared to a classical t-test based approach. Because of the difference of the test statistic distribution under the independence assumption and real one, the D-test requires more traces to reach the 10^{-5} threshold. However, it remains too optimistic for lower thresholds.

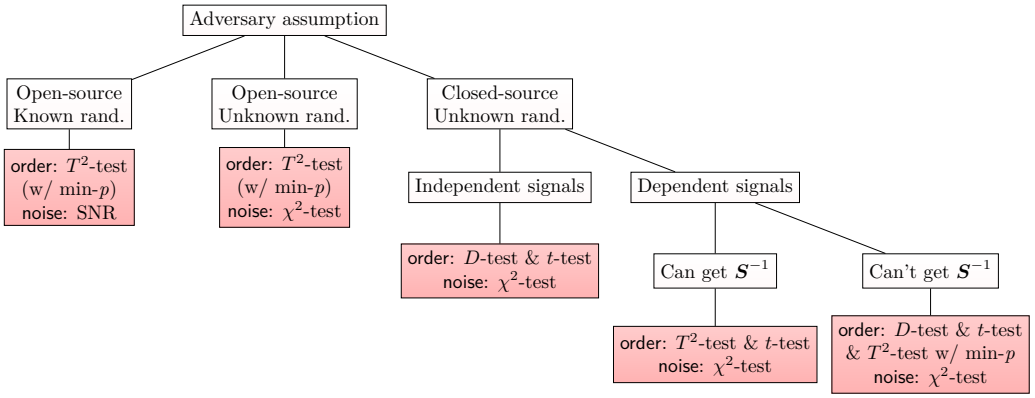


Figure 12: Proposed security assessment framework.

5 Discussion and Conclusion

In this paper, we proposed to use Hotelling's T^2-test and its specialization the D-test for leakage detection. In our experiments, we have shown that multi-tuple detection can lead to improved detection rates over classical Welch's t-test with min-p if the measurements are dense. For low-density traces, the test methodology based on Welch's t-test performed better. Furthermore, we explored the dependent signal issue which may lead to an overestimation of the security order by classical TVLA and an underestimation by the D-test. Only Hotelling's T^2-test behaves in a statistically sound manner with non-diagonal covariances. Nevertheless, its increased computational complexity can be problematic for (e.g.,) masked software implementations, in which case a closed-source evaluator may not be able to apply it even with the min-p extension.

All these observations raise the question of what methodology should be used to perform sound leakage detection. Since we found that not one test is optimal for every setting, we instead propose a generic framework depicted in Figure 12, based on the adversarial assumptions considered in the evaluation. The goal of this framework is to allow evaluators to make founded claims about the main security-influencing parameters of their implementations (i.e., SNR, dependency, density, security order).

First, we consider the implementation knowledge of the adversary. For *open-source* designs, we assume that it is possible to filter non-informative points from the traces and only process points-of-interest in the detection. In this setting, Hotelling's T^2-test is the optimal choice to evaluate the security order. Thanks to the filtering, the evaluator is able to make founded claims about the density of the measurements (i.e., it is expected to be close to $\phi = 1$) and limit the trace length n_l enabling the computation of $\boldsymbol{S}^{-1}$. Therefore, both drawbacks of Hotelling compared to the D-test and Welch's t-test with min-p are eliminated. If even the filtered traces are too long for standard Hotelling, it is advised to apply it with the min-p extension (which may then lead to overestimate the number of traces needed to detect, yet in a more limited manner than by directly using Welch's t-test with the min-p approach. For *closed-source* designs, the evaluator cannot make a founded a priori statement about the density of the measurements. Therefore, in this setting, we recommend running both a multi-tuple test (either Hotelling, Hotelling with min-p or the D-test, depending on whether computing $\boldsymbol{S}^{-1}$ is needed/feasible) and Welch's t-test with min-p in parallel. If the multi-tuple test detects faster, we can conclude that the density is sufficiently high, while the opposite can be concluded if Welch's t-test wins. In both cases, the evaluator estimates the security order with minimum data complexity.

Second, assessing the level of noise in the measurements mostly depends on whether the randomness used in countermeasures is known or unknown to the evaluator. If it is

known, the evaluator can directly estimate the SNR of each share separately as described in [JS17], and from that draw conclusions about the sufficiency of the noise level. By contrast, if it is unknown, the evaluator is limited to heuristics such as the recently-proposed χ^2-test [MRSS18]. Essentially, the heuristic runs the distribution-based χ^2-test in parallel to a moment-based evaluation method like Welch's t-test with min-p, Hotelling, Hotelling with min-p, or the D-test. If the χ^2-test outperforms the moment-based test, the evaluator gets a warning signal that the noise level may be low and (for example) insufficient for the masking countermeasure to provide its exponential security increase.

In general, this framework puts forward that whenever working in a closed-source setting and without known randomness, the tests to launch typically depend on whether the independent signal assumption is acceptable and whether $\boldsymbol{S}^{-1}$ is computable. Hence, it recalls that as the evaluation setting becomes more challenging, additional assumptions and heuristics are increasingly needed, which also implies that the conclusions obtained in such settings need to be considered with care. For the evaluation of the noise level, the heuristic comes from relying on the aforementioned χ^2-test. As for the evaluation of the security order, it comes from the impossibility to compute a sound threshold based on Hotelling's test – yet, in this case using Welch's t-test with min-p, Hotelling with min-p, and the D-test provide an interesting set of (heuristic) lower and upper bounds.

In this respect, a final remark is that strictly speaking, these tests work under an assumption of (close-to) Gaussian leakages. The latter is typically acceptable for unprotected implementations, but not for masked implementations with a multiplicative pre-processing. As mentioned in Section 1.2, the usual assumption in this case is that the sample means that are tested are still close-enough to Gaussian, as per to the central limit theorem. Intuitively, we expect that (*i*) in case of insufficient noise, the assumption will be problematic, but this should be detected thanks to the SNR value (if accessible to the evaluator) or the χ^2-test (otherwise), and (*ii*) in case of large enough noise, the exponential increase of the number of traces to detect should be sufficient for the central limit theorem to be effective. Yet, it is an interesting open problem to to determine the extent to which this assumption can be problematic in practice (in particular in the more challenging to interpret cases where no detections occur).

Supplementary Material. We provide the code for the simulated experiments from which every simulation-based figure can be generated as a complement of the paper [15].

Acknowledgments

François-Xavier Standaert is a senior associate researcher of the Belgian Fund for Scientific Research (FNRS-F.R.S.). This work has been funded in parts by the ERC project 724725 (acronym SWORD) and by the H2020 project REASSURE. The authors thank Itamar Levi for his help in improving the measurement setup used in our experiments, and Carolyn Whitnall for stimulating comments on leakage detection and for putting forward the importance of the independence assumption in this context.

References

[BGG+14] Josep Balasch, Benedikt Gierlichs, Vincent Grosso, Oscar Reparaz, and François-Xavier Standaert. On the cost of lazy engineering for masked software implementations. In *CARDIS*, volume 8968 of *Lecture Notes in Computer Science*, pages 64–81. Springer, 2014.

[15] https://github.com/obronchain/multituple_leakage_detection

[BPG18] Florian Bache, Christina Plump, and Tim Güneysu. Confident leakage assessment - A side-channel evaluation framework based on confidence intervals. In *2018 Design, Automation & Test in Europe Conference & Exhibition, DATE 2018, Dresden, Germany, March 19-23, 2018*, pages 1117–1122. IEEE, 2018.

[CDP16] Eleonora Cagli, Cécile Dumas, and Emmanuel Prouff. Kernel discriminant analysis for information extraction in the presence of masking. In Kerstin Lemke-Rust and Michael Tunstall, editors, *Smart Card Research and Advanced Applications - 15th International Conference, CARDIS 2016, Cannes, France, November 7-9, 2016, Revised Selected Papers*, volume 10146 of *Lecture Notes in Computer Science*, pages 1–22. Springer, 2016.

[CMG+] Jeremy Cooper, Elke De Mulder, Gilbert Goodwill, Josh Jaffe, Gary Kenworthy, and Pankaj Rohatgi. Test vector leakage assessment (TVLA) methodology in practice (extended abstract). ICMC 2013. http://icmc-2013.org/wp/wp-content/uploads/2013/09/goodwillkenworthtestvector.pdf.

[DS16] François Durvaux and François-Xavier Standaert. From improved leakage detection to the detection of points of interests in leakage traces. In Marc Fischlin and Jean-Sébastien Coron, editors, *Advances in Cryptology - EUROCRYPT 2016 - 35th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Vienna, Austria, May 8-12, 2016, Proceedings, Part I*, volume 9665 of *Lecture Notes in Computer Science*, pages 240–262. Springer, 2016.

[DSV+15] François Durvaux, François-Xavier Standaert, Nicolas Veyrat-Charvillon, Jean-Baptiste Mairy, and Yves Deville. Efficient selection of time samples for higher-order DPA with projection pursuits. In Stefan Mangard and Axel Y. Poschmann, editors, *Constructive Side-Channel Analysis and Secure Design - 6th International Workshop, COSADE 2015, Berlin, Germany, April 13-14, 2015. Revised Selected Papers*, volume 9064 of *Lecture Notes in Computer Science*, pages 34–50. Springer, 2015.

[DZD+17] A. Adam Ding, Liwei Zhang, François Durvaux, François-Xavier Standaert, and Yunsi Fei. Towards sound and optimal leakage detection procedure. In Thomas Eisenbarth and Yannick Teglia, editors, *Smart Card Research and Advanced Applications - 16th International Conference, CARDIS 2017, Lugano, Switzerland, November 13-15, 2017, Revised Selected Papers*, volume 10728 of *Lecture Notes in Computer Science*, pages 105–122. Springer, 2017.

[GJJR11] Gilbert Goodwill, Benjamin Jun, Josh Jaffe, and Pankaj Rohatgi. A testing methodology for side channel resistance validation. NIST non-invasive attack testing workshop, 2011. http://csrc.nist.gov/news_events/non-invasive-attack-testing-workshop/papers/08_Goodwill.pdf.

[GMK16] Hannes Groß, Stefan Mangard, and Thomas Korak. Domain-oriented masking: Compact masked hardware implementations with arbitrary protection order. In Begül Bilgin, Svetla Nikova, and Vincent Rijmen, editors, *Proceedings of the ACM Workshop on Theory of Implementation Security, TIS@CCS 2016 Vienna, Austria, October, 2016*, page 3. ACM, 2016.

[GMK17] Hannes Groß, Stefan Mangard, and Thomas Korak. An efficient side-channel protected AES implementation with arbitrary protection order. In *CT-RSA*, volume 10159 of *Lecture Notes in Computer Science*, pages 95–112. Springer, 2017.

[Hot31] Harold Hotelling. The generalization of student's ratio. *The Annals of Mathematical Statistics*, 2(3):360–378, 1931.

[ISW03] Yuval Ishai, Amit Sahai, and David A. Wagner. Private circuits: Securing hardware against probing attacks. In Dan Boneh, editor, *Advances in Cryptology - CRYPTO 2003, 23rd Annual International Cryptology Conference, Santa Barbara, California, USA, August 17-21, 2003, Proceedings*, volume 2729 of *Lecture Notes in Computer Science*, pages 463–481. Springer, 2003.

[JS17] Anthony Journault and François-Xavier Standaert. Very high order masking: Efficient implementation and security evaluation. In Wieland Fischer and Naofumi Homma, editors, *Cryptographic Hardware and Embedded Systems - CHES 2017 - 19th International Conference, Taipei, Taiwan, September 25-28, 2017, Proceedings*, volume 10529 of *Lecture Notes in Computer Science*, pages 623–643. Springer, 2017.

[Man04] Stefan Mangard. Hardware countermeasures against DPA ? A statistical analysis of their effectiveness. In Tatsuaki Okamoto, editor, *Topics in Cryptology - CT-RSA 2004, The Cryptographers' Track at the RSA Conference 2004, San Francisco, CA, USA, February 23-27, 2004, Proceedings*, volume 2964 of *Lecture Notes in Computer Science*, pages 222–235. Springer, 2004.

[MOBW13] Luke Mather, Elisabeth Oswald, Joe Bandenburg, and Marcin Wójcik. Does my device leak information? an a priori statistical power analysis of leakage detection tests. In Kazue Sako and Palash Sarkar, editors, *Advances in Cryptology - ASIACRYPT 2013 - 19th International Conference on the Theory and Application of Cryptology and Information Security, Bengaluru, India, December 1-5, 2013, Proceedings, Part I*, volume 8269 of *Lecture Notes in Computer Science*, pages 486–505. Springer, 2013.

[MOP07] Stefan Mangard, Elisabeth Oswald, and Thomas Popp. *Power analysis attacks - revealing the secrets of smart cards*. Springer, 2007.

[MOS11] Stefan Mangard, Elisabeth Oswald, and François-Xavier Standaert. One for all - all for one: unifying standard differential power analysis attacks. *IET Information Security*, 5(2):100–110, 2011.

[MRSS18] Amir Moradi, Bastian Richter, Tobias Schneider, and François-Xavier Standaert. Leakage detection with the x2-test. *IACR Trans. Cryptogr. Hardw. Embed. Syst.*, 2018(1):209–237, 2018.

[RBG+16] Debapriya Basu Roy, Shivam Bhasin, Sylvain Guilley, Annelie Heuser, Sikhar Patranabis, and Debdeep Mukhopadhyay. Leak me if you can: Does tvla reveal success rate? *IACR Cryptology ePrint Archive*, 2016:1152, 2016.

[Ren98] Alvin C Rencher. *Multivariate statistical inference and applications*, volume 338. Wiley-Interscience, 1998.

[RP10] Matthieu Rivain and Emmanuel Prouff. Provably secure higher-order masking of AES. In Stefan Mangard and François-Xavier Standaert, editors, *Cryptographic Hardware and Embedded Systems, CHES 2010, 12th International Workshop, Santa Barbara, CA, USA, August 17-20, 2010. Proceedings*, volume 6225 of *Lecture Notes in Computer Science*, pages 413–427. Springer, 2010.

[SM16] Tobias Schneider and Amir Moradi. Leakage assessment methodology - extended version. *J. Cryptographic Engineering*, 6(2):85–99, 2016.

[Sta17] François-Xavier Standaert. How (not) to use welch's t-test in side-channel security evaluations. *IACR Cryptology ePrint Archive*, 2017:138, 2017.

[Wag12] Mathias Wagner. 700+ attacks published on smart cards: The need for a systematic counter strategy. In Werner Schindler and Sorin A. Huss, editors, *Constructive Side-Channel Analysis and Secure Design - Third International Workshop, COSADE 2012, Darmstadt, Germany, May 3-4, 2012. Proceedings*, volume 7275 of *Lecture Notes in Computer Science*, pages 33–38. Springer, 2012.

[Wel47] B. L. Welch. The generalization of 'student's' problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35, 1947.

[WGS06] Yujun Wu, Marc G Genton, and Leonard A Stefanski. A multivariate two-sample mean test for small sample size and missing data. *Biometrics*, 62(3):877–885, 2006.

[Wy92] LIN Wen-ying. An overview of the performance of four alternatives to hotelling's t square. *Educational Research Journal*, 7:110–114, 1992.

A Distributions

The distributions used in the paper are defined in Table 2, where Γ is the gamma function, γ the lower incomplete gamma function, B the beta function, I the regularized incomplete beta function, and ${}_2F_1$ is the hypergeometric function.

Table 2: The PDF's and CDF's of the considered distributions.

Distribution	PDF	CDF
$\chi^2(k)$	$\frac{1}{2^{df/2}\Gamma(df/2)}\, x^{df/2-1}e^{-x/2}$	$\frac{1}{\Gamma(k/2)}\,\gamma\left(\frac{k}{2},\,\frac{x}{2}\right)$
$\mathcal{F}(d1,d2)$	$\frac{\sqrt{\frac{(d_1\,x)^{d_1}\;d_2^{d_2}}{(d_1\,x+d_2)^{d_1+d_2}}}}{x\,\mathrm{B}\left(\frac{d_1}{2},\frac{d_2}{2}\right)}$	$I_{\frac{d_1 x}{d_1 x+d_2}}\left(\frac{d_1}{2},\frac{d_2}{2}\right)$
$\mathcal{T}(\nu)$	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$	$\frac{1}{2}+x\Gamma\left(\frac{\nu+1}{2}\right)\times\frac{{}_2F_1\left(\frac{1}{2},\frac{\nu+1}{2};\frac{3}{2};-\frac{x^2}{\nu}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}$

B Investigated Covariance Matrix

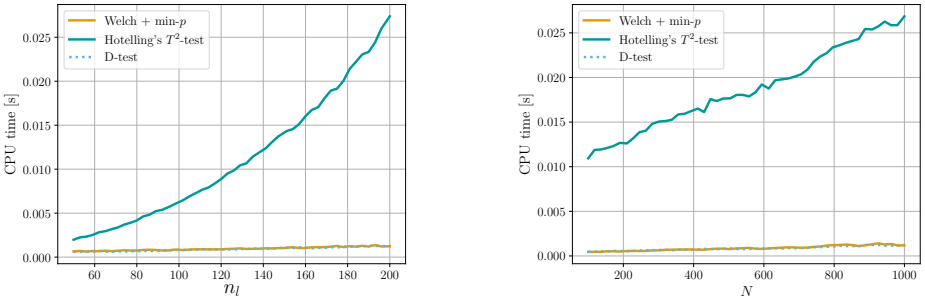
The exact values of the investigated covariance (cf. Figure 7) denoted as $\sigma^2_{i,j}$ are defined according to a dependency coefficient Δ and the SNR. The diagonal elements are given by the SNR such that $\sigma^2_{i,i} = 2/\mathrm{SNR}$, and the non-diagonal elements are defined as

$$\sigma^2_{i,j} = \max(\sigma^2_{i,i} \cdot (1 - \Delta \cdot |i-j|), 0),$$

where Δ measures the independence of one random variable $\boldsymbol{X_i}$ (resp., $\boldsymbol{Y_i}$) with the adjacent ones. Since Σ_1 is diagonal, the coefficient Δ is equal to infinity. For Σ_2 and Σ_3, the coefficients are respectively $\Delta = 0.1$ and $\Delta = 0.02$. For example, this results in the covariances $\sigma^2_{0,1} = 0.9$ and $\sigma^2_{0,2} = 0.8$ for Σ_2 with $\mathrm{SNR} = 1$.

C Computational Complexity Evaluations

Once the traces have been collected, the obtained measurements need to be processed. Since these methods require different knowledge about the data, the time spend in processing is not the same. In the following, we show the CPU time required for the three leakage detection methods. The code is written in `Python3` and is based on `Numpy1.14` library. The time is recorded for a single core of an `Intel(R) Core(TM) i7-5500U CPU @ 2.40GHz`.



(a) $N = 1000$ (b) $n_l = 200$

Figure 13: Required CPU time of the leakage detection algorithms.

First, we observe that the computational complexity of Welch's t-test with min-p and the D-test grows linearly with the number of points in the trace n_l, since they are processed independently (cf. Figure 13a) with only a minor overhead for the D-test (due to the summed t_i^2). However, computing and inverting the pooled covariance matrix is roughly quadratic in n_l, as depicted in Figure 13a. As a result, it rapidly becomes out of reach for large traces (e.g., pre-processed for higher-order detection).

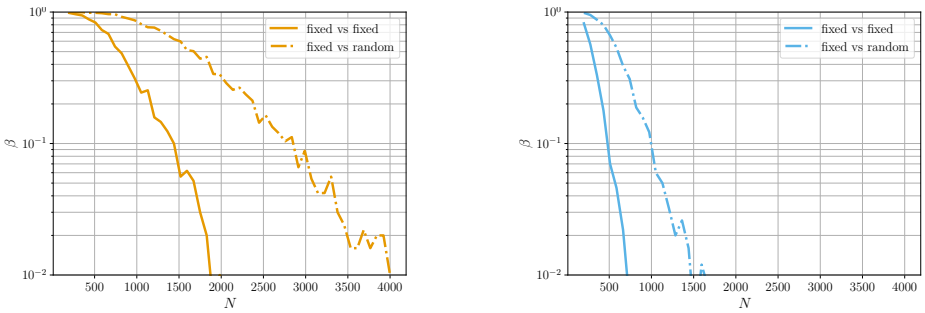
Second, the complexity grows linearly with the number of collected measurements $N_{\boldsymbol{X}}$ and $N_{\boldsymbol{Y}}$ in all the cases. However, the constant is much larger for the Hotelling's T^2-test due to the matrix computation (cf. Figure 13b).

Finally, we mention that even if the processing complexity of the Hotelling's T^2-test is much larger, it significantly reduces the time spent in measurement (in the case of dense traces), which is generally the longest part of a side-channel security evaluation. So in the case of short and dense traces (i.e., open-source design), the Hotelling's T^2-test remains the methodology of choice. In the case of long and low density traces, the D-test should be performed with only a minor overhead in computational complexity and potentially a larger gain in data complexity (cf. Figure 12).

D Comparisons to alternative TVLA

In this section, we provide further observations related to other TVLA-based approaches. First, we conduct some *fixed vs. random* experiments and compare it to our *fixed vs. fixed* results. Second, we discuss the behavior of the original approach from [GJJR11] to reduce the false positive rate (i.e., multiple tests) in our scenario.

Fixed vs. random or fixed vs. fixed? Figure 14 highlights the detection rates for both the t-test and D-test with 500 average experiments. In both cases, we observe that for a fixed number of traces, the false negative probability β is smaller in the case of *fixed vs. fixed* than in the *fixed vs. random* scenario, as expected from [DS16]. This observation justifies the choice of *fixed vs. fixed* for the performed experiments. Nevertheless, it is expected that the observations for *fixed vs. fixed* transfer to *fixed vs. random*.



(a) Welch's t-test with min-p. (b) D-test.

Figure 14: Detection rate for various input sets.

Multiple tests. The approach of [GJJR11, CMG+] consists in running twice the same experiment with the different inputs and declare the presence of leakage if one of the random variables shows leakage in both experiments. In the following, we consider two variants of this approach. The original 2×-Welch's t-tests are evaluated with the same threshold th_1 as a single test. However, as shown later, this may leads to incorrect results. Therefore, we conduct another experiment with 2×-Welch's t-test and an adjusted

Table 3: Empirical false positive rate $\hat{\alpha}$ estimated from 100,000 experiments.

	th	Σ_1	Σ_2	Σ_3
Welch's t-test with min-p	th_1	0.1	0.051	0.017
2×-Welch's t-test with min-p	th_1	0.010	0.002	0
2×-Welch's t-test with min-p	th_2	0.099	0.025	0.002

threshold th_2, which should guarantee the correct fault positive rate in case of independent samples. Table 3 contains the sampled false positive α for various such experiments.

It is noticeable that without an adequately adjusted threshold, the false positive rate is significantly reduced for 2×-Welch's t-test. In particular, it is set to α^2 for Σ_1 which is significantly smaller than α. Only by adjusting the threshold to th_2, the new tests performs as expected. Therefore, in practice it is necessary to consider this effect when trying to correctly interpret the results of a 2×-Welch's t-test.

The false negative probabilities of these approaches are depicted in Figure 15 with N representing the total amount of collected traces.[16] First, by moving from a single Welch's t-test to two by keeping the same threshold th_1, the first method obtains a smaller β with the same number of traces. However, this comparison is not fair since the two approaches do not have the same α. So by adjusting the threshold to th_2, and so keeping the same α for a fair comparison, the single test approach performs better. This is due to the fact that in the second case, a single test is performed on half of the measurements.

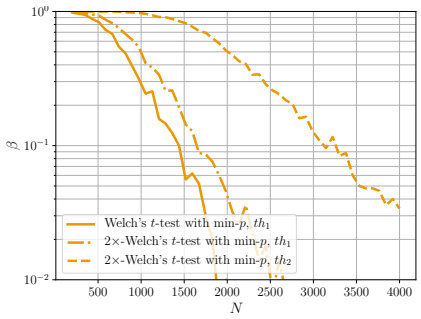


Figure 15: Detection rate of comparison with TVLA

[16] Note that for 2×-Welch's t-test, each test is performed with $N/2$ traces.