How (Not) to Use Welch's T-Test in Side-Channel Security Evaluations

François-Xavier Standaert(✉)

ICTEAM/ELEN/Crypto Group, Université catholique de Louvain, Ottignies-Louvain-la-Neuve, Belgium
fstandae@uclouvain.be

Abstract. The Test Vector Leakage Assessment (TVLA) methodology is a qualitative tool relying on Welch's T-test to assess the security of cryptographic implementations against side-channel attacks. Despite known limitations (e.g., risks of false negatives and positives), it is sometimes considered as a pass-fail test to determine whether such implementations are "safe" or not (without clear definition of what is "safe"). In this note, we clarify the limited quantitative meaning of this test when used as a standalone tool. For this purpose, we first show that the straightforward application of this approach to assess the security of a masked implementation is not sufficient. More precisely, we show that even in a simple (more precisely, univariate) case study that seems best suited for the TVLA methodology, detection (or lack thereof) with Welch's T-test can be totally disconnected from the actual security level of an implementation. For this purpose, we put forward the case of a realistic masking scheme that looks very safe from the TVLA point-of-view and is nevertheless easy to break. We then discuss this result in more general terms and argue that this limitation is shared by all "moment-based" security evaluations. We conclude the note positively, by describing how to use moment-based analyses as a useful ingredient of side-channel security evaluations, to determine a "security order".

1 Introduction

Leakage detection tests have recently emerged as a convenient solution to perform preliminary (black box) evaluations of resistance against side-channel analysis. Cryptography Research (CRI)'s non-specific (fixed vs. random) T-test is a popular example of this trend [8,13]. It works by comparing the leakages of a cryptographic (e.g., block cipher) implementation with fixed plaintexts (and key) to the leakages of the same implementation with random plaintexts (and fixed key)[1], thanks to Welch's T-test [31]. Besides its conceptual simplicity, the main advantage of such a test, that was carefully discussed in [11,19,27], is its

[1] The Test Vector Leakage Assessment methodology in [8,13] includes other options such as non-specific semi-fixed vs. random tests and specific tests – we focus on the non-specific fixed vs. random test that is the most popular in the literature.

B. Bilgin and J.-B. Fischer (Eds.): CARDIS 2018, LNCS 11389, pp. 65–79, 2019.
https://doi.org/10.1007/978-3-030-15462-2_5

low sampling complexity. That is, by comparing only two (fixed vs. random) classes of leakages, one reduces the detection problem to a simpler estimation task. And since these tests are generally applied independently to many leakage samples (e.g., corresponding to a full block cipher execution), they generally take advantage of the larger signal (i.e., the larger difference of means between the fixed and random classes) that occur for some samples with high probability.

Limitations and Improvements. The counterpart to this lower sampling complexity is a risk of false negatives and positives. Regarding false negatives, it may for example happen that for some informative samples, the mean values of the fixed and random classes are identical (resp., very similar), which makes detection impossible (resp., measurement-intensive). Yet, by applying the TVLA methodology to large enough traces (possibly with a few different fixed classes), the risk that significant leakages remain unnoticed for a complete (e.g., block cipher) implementation is usually expected to remain negligible. Regarding false positives, they rather relate to the fact that a (non-specific) T-test spots informative samples independent of their exploitability with standard Differential Power Analysis (DPA) attacks [18]. For example, the latter attacks typically target an enumerable part of the key that is manipulated in the first block cipher rounds, while the real and random classes differ in all the cipher rounds. More specific (and informative) detections can however be obtained by computing more specific metrics (i.e., targeting specific computations of the implementation), at the cost of a more expensive estimation. So in summary, the state-of-the-art typically views the TVLA methodology as a tradeoff between the sampling complexity and the informativeness of the leakage detection. Note that as discussed in [11], the sampling complexity of non-specific T-tests can be further reduced by considering two fixed classes (rather than a fixed and a random one).

A Tempting Shortcoming. In view of these advantages and limitations, it is sometimes considered that the TVLA methodology is "*a pass-fail test which determines whether the crypto implementation is safe or not*" [26]. But this naturally raises the question of what is precisely meant by "safe". For example, it is tempting (and as will be shown, incorrect) to expect that a device successively passing a non-specific T-test with Q traces is secure against side-channel attacks with up to Q traces. Clearly, this cannot hold in general. Indeed, and even assuming that the aforementioned false positives and negatives do not occur, another limitation of the original TVLA methodology is that it is inherently univariate. This implies that whenever multivariate attacks are more powerful than univariate ones, a leaking device can pass a non-specific T-test despite being weak in the general sense (i.e., breakable with less traces than used by the TVLA methodology). Concrete examples of this situation include the exploitation of static leakages [20,24], and serial implementations of masking schemes for which the number of exploitable leakage samples grows quadratically in the number of shares, which implies that univariate attacks become less and less relevant to evaluate their security level as this number of shares increases [3]. Note that the work of Schneider and Moradi in [27] mitigates this limitation by integrating the possibility to estimate mixed statistical moments in their leakage detection. Yet,

even in that case the resulting evaluation remains insufficient since corresponding to the exploitation of one tuple of leaking samples, while the optimal attack should take advantage of all the informative tuples in the leakage traces [14].

Note also that this kind of limitation was already mentioned from the introduction of the TVLA methodology. In particular, [16] (Section 5) clearly points out that blinded RSA implementations suffering from SPA leakages (which are one more example of highly multivariate attacks) may pass the T-test despite being vulnerable to other attacks, and therefore require additional analyses.

The Case of Parallel Masked (e.g., Threshold) Implementations. In practice, non-specific T-tests have been the method of choice for the security evaluation of higher-order threshold implementations manipulating their shares in parallel, such as discussed in [4,6,7]. Based on this state-of-the-art, our goal in this note is to further clarify what is learned (and what is missed) by the standalone application of the TVLA methodology in this case. Admittedly, our results do not contradict the published literature. (Precisely: the previous papers did not claim that the application of this methodology was correlated with a quantitative security level). We only recall that performing univariate T-tests is only an ingredient of a sound side-channel security evaluation that has to be combined with other ones, and that the gap between the standalone application of this methodology and a sound security evaluation increases with the security levels. More precisely, in the case of masking the TVLA methodology is good to detect a "security order" (i.e., the lowest key-dependent statistical moment of the leakage distribution). But in general a high security order is not sufficient to guarantee a high security level (e.g., number of traces for key recovery): one also needs to ensure a sufficient noise. So in order to claim quantitative results for masked/threshold implementations, the TVLA methodology has to be combined with a noise analysis and/or information theoretic evaluation.

In order to make our discussion concrete, we next consider side-channel attacks exploiting a single leakage sample corresponding to the parallel manipulation of several shares in a masked/threshold implementation. Based on this example, we compare the number of samples needed to detect fixed and random (or fixed) classes with a non-specific T-test and the DPA security of the implementation. None of our conclusions are new from the theoretical point-of-view. We only use this example to make explicit that even ignoring the issue of highly multivariate attacks, the standalone application of the TVLA methodology can be highly misleading regarding the actual security level of an implementation (i.e., the number of traces needed for key recovery). In this respect, the main concern of this note is not the use of the TVLA methodology for research purposes, but its potential misuse in the security evaluation of real products.

Cautionary Remarks. Despite the goal of this note is to prevent the misuse of the TVLA methodology when evaluating real products, we are not claiming that it is currently misused by any evaluation laboratory. We wrote it as a complement to several informal discussions that we had over the last months with researchers and engineers unconvinced that applying the TVLA methodology is not sufficient to state quantitative conclusions on the physical security of a cryptographic

implementation, which is now clarified by the next example. Conceptually, this example in fact falls under the general (and known) observation that the TVLA methodology is unable to detect SPA leakages (e.g., mentioned in [16]). So it should be viewed as a reminder that such SPA leakages can happen even in the case of univariate attacks against parallel masking schemes. In this respect, the note is also of (mostly) prospective nature, since the limitation it points out relates to (very) high order masking schemes, while the TVLA methodology has mostly been used for low order masked implementations so far. Besides, and as will be clear in Sect. 3, our results do not contradict the value of the TVLA methodology, as an ingredient to detect the security order of a masked implementation, or as a useful first step before more advanced analyses.

2 Case Study: How Not to Use the T-Test

2.1 Setup and Metrics

Our following discussions will be based on the parallel implementation of a simple masking scheme such as described in [2]. More precisely, we will consider the simplest example where all the shares are in $\mathsf{GF}(2)$ (generalizations to larger fields follow naturally). In this setting, we have a sensitive variable x that is split into m shares such that $x = x_1 \oplus x_2 \oplus \ldots \oplus x_m$, with $\oplus$ the bitwise XOR. The first $m-1$ shares are picked up uniformly at random: $(x_1, x_2, \ldots, x_{m-1}) \xleftarrow{R} \{0,1\}$, and the last one is computed as $x_m = x \oplus x_1 \oplus x_2 \oplus \ldots \oplus x_{m-1}$.

Denoting the vector of shares $(x_1, x_2, \ldots, x_m)$ as $\bar{x}$, we will consider an adversary who observes a single leakage sample corresponding to the parallel manipulation of these shares. A simple model for this setting is to assume this sample to be a linear combination of the shares, namely:

$$\mathsf{L}_1(\bar{x}) = \left(\sum_{i=1}^{m} \alpha_i \cdot x_i \right) + N,$$

where $+, \cdot$ are the addition and multiplication in $\mathbb{R}$, the α_i's are coefficients in $\mathbb{R}$ and N is a noise random variable that we will assume Gaussian distributed with variance σ_n^2. The case with all α_i's equal to one corresponds to the popular Hamming weight leakage function. A slightly more sophisticated model would additionally consider quadratic terms, leading to:

$$\mathsf{L}_2(\bar{x}) = \left(\sum_{i=1}^{m} \alpha_i \cdot x_i \right) + \left(\sum_{i,j=1}^{m} \beta_{i,j} \cdot (x_i \wedge x_j) \right) + N,$$

with $\wedge$ the bitwise AND. The algebraic degree of this function can be extended similarly up to $d \leq m$, capturing increasingly complex leakages.

A standard (worst-case) metric to capture the informativeness of these leakages is the mutual information [29] that can be computed as follows:

$$\mathrm{MI}(X; \mathsf{L}_d(\bar{X})) = \mathrm{H}[X] + \sum_{x \in \mathcal{X}} \Pr[x] \cdot \sum_{l \in \mathcal{L}} \mathsf{f}(l|x) \cdot \log_2 \Pr[x|l].$$

In this equation, $\mathsf{f}(l|x)$ is the conditional Probability Density Function (PDF) of the leakages $\mathsf{L}(\bar{X})$ given the secret X, which (assuming Gaussian noise) can be written as the following Gaussian mixture model:

$$\mathsf{f}(l|x) = \sum_{\bar{x} \in \mathcal{X}^{d-1}} \mathcal{N}\left(l|(x,\bar{x}), \sigma_n^2\right),$$

and the conditional probability $\Pr[x|l]$ is computed thanks to Bayes' theorem as:

$$\Pr[x|l] = \frac{\mathsf{f}(l|x)}{\sum_{x^* \in \mathcal{X}} \mathsf{f}(l|x^*)}.$$

We recall that this mutual information metric is correlated with the measurement complexity of a worst-case template attack, as demonstrated in [10], which we next use as a relevant (quantitative) metric to capture side-channel security.

In our simple (single-bit secret) case, the TVLA methodology works by collecting Q_0 (resp. Q_1) traces corresponding to the secret value $X = 0$ (resp. $X = 1$) and stores them in vectors $\bar{L}_0$ (resp. $\bar{L}_1$). In order to capture higher-order security, and following what was done in [4,6,7,27], we then process these vectors by removing their mean (so that we next estimate central moments) and raise them to a power o, that we will denote as the attack order. This leads to vectors $\bar{L}'_0$ (resp. $\bar{L}'_1$) of which the samples equal (e.g., for $\bar{L}'_0$):

$$\bar{L}'_0(i) = \left(\bar{L}_0(i) - \hat{\mathsf{E}}(\bar{L}_0)\right)^o,$$

with $\hat{\mathsf{E}}$ the sample mean operator and for $1 \leq i \leq Q_0$. Based on these leakage vectors, the TVLA methodology computes Welch's T statistic as follows:

$$\Delta = \frac{\hat{\mathsf{E}}(\bar{L}'_0) - \hat{\mathsf{E}}(\bar{L}'_1)}{\sqrt{\frac{\hat{\mathsf{var}}(\bar{L}'_0)}{Q_0} + \frac{\hat{\mathsf{var}}(\bar{L}'_1)}{Q_1}}},$$

with $\hat{\mathsf{var}}$ the sample variance operator. The side-channel literature usually assumes this T statistic to be significant when a threshold of 5 is passed.[2]

2.2 Experimental Results

Based on the setup in the previous section, we started by performing an information theoretic evaluation of our parallel implementation of a Boolean encoding, which is reported in Fig. 1. In order to allow an easier interpretation of the results, we use the Signal-to-Noise Ratio (SNR) as X axis, defined as the variance of the noise-free traces (e.g., m/4 for a Hamming weight model) divided by the variance of the noise. It better reflects the fact that the impact of the noise depends on the scaling of the signal. The figure carries the usual intuitions:

[2] In general, this threshold has to be set in function of the number of samples in the traces, to reflect the probability that a high Δ is observed by chance [9].

Boolean masking provides limited security for low noise levels; the slope of the IT curve reveals the security order of the implementation (i.e., relates to the smallest key-dependent moment of the leakage distribution) for high noise levels; and a leakage function mixing the shares in a non-linear manner (e.g., a quadratic one for the dotted curve) reduces the security order according to its algebraic degree.[3] For our discussions, it is mostly the first observation that matters.

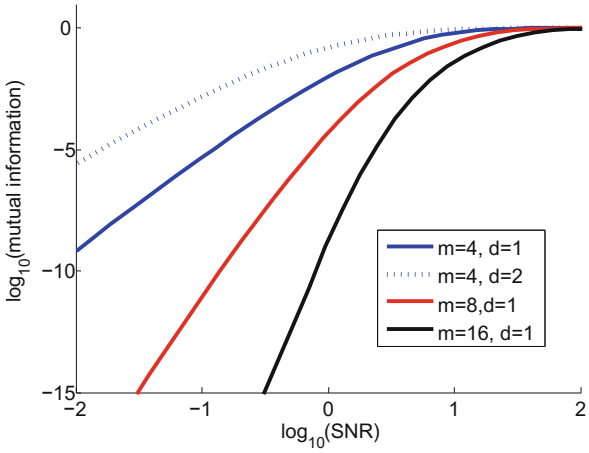

Fig. 1. Information theoretic evaluation of the (parallel) Boolean encoding. (Color figure online)

Note that in the case of the degree 1 leakage function with all α_i's equal to 1, it is easy to see that the high information observed for low noise levels corresponds to a powerful and concrete attack. Namely, without noise the adversary just has to check whether the leakage sample he obtains is odd or even.

As a complement to this information theoretic evaluation, we launched the TVLA methodology. For this purpose, we started with the case of an $m = 4$-share masking, leaking according to a linear leakage function (i.e., $d = 1$) and for a very low noise level ($\sigma_n^2 = 10^{-2}$). It corresponds to the rightmost point of the plain blue curve of Fig. 1 and therefore to an insecure implementation. Since the security order in this 4-share case study is expected to be four, we carried out Welch's T-test with traces raised to powers $o = 3$ and $o = 4$ and reported the results of ten independent experiments in Fig. 2. As expected, the third-order test does not succeed while the fourth-order one does. However, it already requires a couple of hundreds traces to detect with confidence, which seems a lot compared to the (large) information leaked by this sample.

[3] A higher-degree leakage function manipulating shares in parallel is in fact the natural mathematical model to capture the independence issues discussed in [2], which can be caused in practice by glitches, transition-based leakages or couplings.

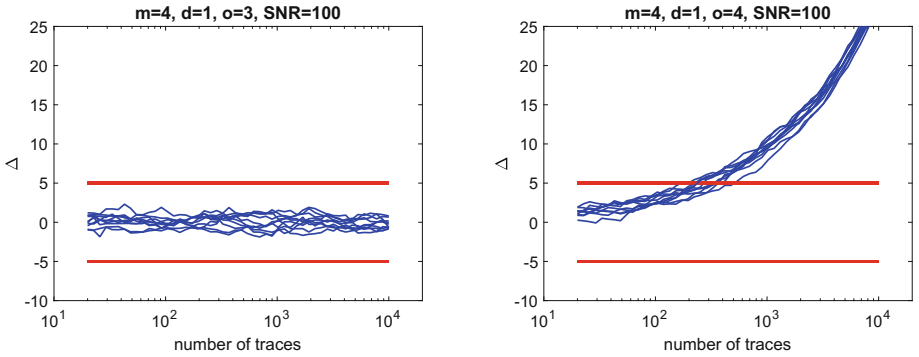

Fig. 2. Results of the TVLA methodology for 4-share (parallel) masking.

In order to confirm this first impression, we then launched the TVLA methodology for the cases of of an $m = 8$-share and $m = 12$-share masking (same leakage function, same noise level). As expected again, the lowest successful detection orders were respectively 8 and 12. But as reported in Fig. 3, the complexity of the detection task increases significantly (in fact, exponentially) with the number of shares, which clearly contradicts the information theoretic analysis of the Boolean encoding for low noise levels. Hence, this case study highlights an issue with the (tempting shortcoming of the) TVLA methodology, since the number of traces needed to detect with it can be made arbitrarily larger than the one needed to recover the secret (by increasing the number of shares m).

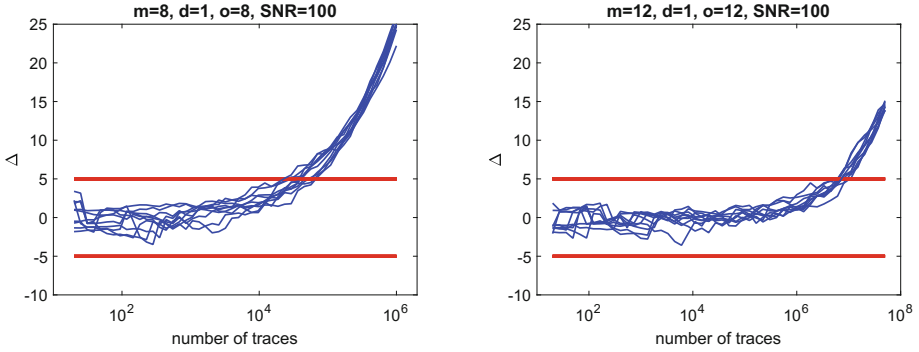

Fig. 3. Results of the TVLA methodology for 8- and 12-share parallel masking.

2.3 Interpretation

What Went Wrong? In short, the main issue of TVLA methodology as applied in the previous subsection is that *it assumes an adversarial strategy*, which relies on estimating the statistical moments of the leakage distribution. In theory this

is a risky approach since security arguments generally aim at being independent of the adversarial strategy. Our example shows that even in practice, estimating statistical moments is in fact not the best strategy to attack a masked implementation with low noise levels (which naturally follows from the hypotheses in masking proofs [10]). Furthermore, the gap between this strategy and the optimal one increases with the security order. Note that our previous examples focus on parallel implementations (which are a more natural target for the application of Welch's T-test since mitigating the dimensionality issue discussed in introduction), but the same observation holds for serial implementations.[4]

An Analogy. A similar situation was observed in [15,17] when comparing the Gaussian mixture and Gaussian adversaries: the latter one does in fact exactly the same "mistake" as the TVLA methodology since "summarizing" a mixture into a statistical moment, namely the (co)variance. So for low noise levels, the Gaussian adversary will generally overstate the security level of a protected implementation, by interpreting mask (or supply voltage) variations as a single Gaussian with larger (co)variance. As in our previous example, this amounts to implicitly assume the existence of a large enough noise without testing it.

Impact for Threshold Implementations. These results illustrate that testing a masked/threshold implementation with the TVLA methodology *only* is not sufficient to gain accurate insights on its security level, especially as the security order increases. However, our observations do not contradict the results in [4,6,7] where the authors only claimed a security order (which is exactly what the TVLA methodology is good for – see next). Reading these papers, it is also clear that their authors are well aware that noise is needed for their countermeasure to provide security. So concretely, the only limitation of these works is that they are not quantitative. In this respect, our results come with the important cautionary remark that a quantitative approach is increasingly needed when masking security orders increase, since the gap between the number of traces needed to detect fixed and random (or fixed) classes with the TVLA methodology and the actual (worst-case) security level of an implementation also increases in this context. In order to avoid this caveat, the TVLA methodology has to be combined with an analysis of the noise (and ideally, an information theoretic evaluation of the leakages), which then enables a quantified implementation security assessment. As mentioned in introduction, we again insist that the main concern in this note is not the use of the TVLA methodology for research purposes (where claiming a security order and assuming noise to be a security parameter is acceptable), but its potential misuse in the security evaluation of real products for which the noise is fixed (i.e., not a security parameter) and the most relevant metric is the number of traces needed to perform a successful key recovery.

We note also that we would obtain similar conclusions with more complex (i.e., not only linear) leakages since noise is in general a necessary condition for the security of the masking countermeasure. Yet, trivial examples (e.g., checking

[4] In a trivial manner: an adversary getting d noise-free leakages corresponding to the d shares of a secret x will not estimate moments but simply XOR them together.

whether the leakage is odd or even in the parallel case and XORing leakage samples in the serial cas) would not work anymore in this case.

Impact for Other Security Evaluation Tools. Quite naturally, the TVLA methodology is not the only side-channel distinguisher focusing on the estimation of statistical moments. In fact, the higher-order DPAs described in [25,30] or higher-order variations of the Correlation Power Analysis (CPA) described in [21] suffer from the same drawback. Namely, they are only indicative of the actual security level of an implementation *if* the best adversarial strategy is to estimate statistical moments of the leakage distribution. Yet, not sufficient does not mean not necessary. In the next section, we will show that moment-based evaluations remain a useful ingredient for sound side-channel security evaluations.

3 Clarification: How to Use the T-Test

3.1 Separation of Duties

First recall that the only thing our previous experiments showed is that launching a T-test cannot be sufficient for the side-channel security evaluation of a masked/threshold implementation (even in univariate case studies that seem the most suitable context for such tests). In fact, this observation again derives from masking security proofs (e.g., in [10]) where it is explicitly mentioned that such a countermeasure provides security under two hypotheses: sufficient noise and independence. So recast positively from this more theoretical viewpoint, the take home message of this note becomes that the TVLA methodology is useful to determine the security order of an implementation, and that the noise level (which also depends on the number of exploitable leakage samples [14]) has to be tested independently. Interestingly, looking back at the information theoretic plot of Fig. 1 allows putting these observations together, since it shows that when the noise is sufficiently large, the slope of the IT curves reflects the security order, suggesting that the best adversarial strategy is indeed to estimate higher-order statistical moments in this case (e.g., as discussed in [10,21]).

3.2 Beyond the TVLA Methodology

Given that we restrict the goal of the TVLA methodology to the detection of the security order of a masked/threshold implementation, the remaining question is to know whether it is an efficient solution for this purpose. In this respect, one can notice that the main drawback of the processing described in Sect. 2.1 is that it directly raises the leakage samples to a certain power o. This implies that as the noise increases, the number of samples needed to detect will increase exponentially with the number of shares (because the noise is amplified), just as expected from secure masking. But this also implies that this approach is inherently limited if one wants to claim very high security levels. So as for other security evaluation tasks (e.g., key enumeration vs. rank estimation [23]), one

can wonder whether an evaluator can benefit from some shortcut to determine the security order, thanks to additional knowledge he may have access to?

A natural option for this purpose is to take advantage of mask knowledge (if available). That is, say the evaluator has access to the shares' vector $\bar{x}$ for each of his leakage samples. Then, he will be able to identify repeated samples for each of the 2^{m-1} possible sharings of the sensitive variable x. Further say that the number of samples per sharing is N_a for simplicity, then the evaluator can pre-process his leakage samples by averaging them (for each sharing). As a result of this pre-processing, the vectors $\bar{L}_0$ and $\bar{L}_1$ of Sect. 2.1 now have $\tilde{Q}_0 = \tilde{Q}_1 = 2^{m-1}$ values (rather than $N_a \cdot 2^{m-1}$ ones without this pre-processing). But the noise of these pre-processed samples has been reduced *before* raising them to the power o, which mitigates the "noise amplification" of the masking scheme. Concretely, it then remains to determine the averaging parameter N_a which naturally depends on the SNR. Typically, one can choose it so that $\mathrm{SNR} \cdot N_a = 10$ (which means that the pre-processed measurements have $\mathrm{SNR} = 10$).

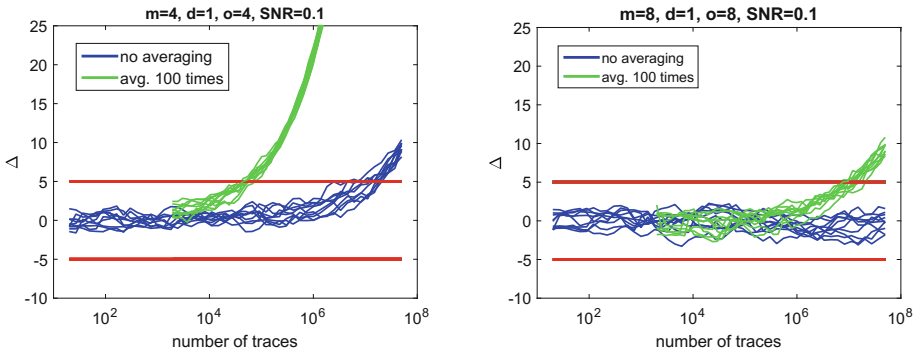

Fig. 4. Comparison between the TVLA and TVLA^2 methodologies.

For illustration, the results of such a "TVLA + averaging" methodology (next denoted as TVLA^2) for a smaller SNR of 0.1, with $m = 4$ and $m = 8$ shares, are represented in Fig. 4. Note that the value of the X axis corresponds to $Q_0 + Q_1$ for the standard TVLA methodology, and to $N_a \cdot 2^m$ for the TVLA^2 one. In other words, it represents the total number of leakage samples used to detect in both cases (which explains why the TVLA^2 curves are shifted by a factor N_a). Several interesting observations can be highlighted. First, the TVLA methodology starts detecting with confidence after 10^7 leakage samples for the $m = 4$ case. This value is nicely related to the MI value of Fig. 1 for the same case ($m = 4$, $\mathrm{SNR} = 10^{-1}$), which is worth $\approx 10^{-6}$ and implies that the number of samples to perform a key recovery should be larger than 10^6 [10]. Similarly, we see that the TVLA methodology does not detect anything for the $m = 8$ case, which is expected since the the MI is then below 10^{-10} for a $\mathrm{SNR} = 10^{-1}$. Second, the average pre-processing of the TVLA^2 methodology significantly improves the

complexity of the detection task. This is due to the previously mentioned noise reduction before amplification. In order to make this gain more explicit, Fig. 5 additionally compares the results of the TVLA^2 methodology for SNRs of 10^{-1} and 10^{-2}. It confirms that the reduction of the SNR by a factor 10 causes an increase of the number of traces needed to detect by a similar factor 10 (and not a factor 10^m as would be observed with the TVLA methodology).

Note that when applying the TVLA^2 methodology, the number of traces needed to detect is even less correlated with the security level of the target implementation than with the TVLA methodology (since concrete adversaries do not know mask values and are not able to perform an average pre-processing). Yet, in view of the limited quantitative meaning of the TVLA methodology in general, and if the TVLA^2 methodology is only used to detect a security order, this drawback is not very critical (when mask knowledge is accessible!).

Eventually, and more negatively, we see from Fig. 4 that the complexity of the TVLA^2 detection still (inevitably) increases exponentially in the number of shares m (since the left and right plots of the have the same SNR). This is in fact exactly the cause of our negative examples in Sect. 2.2. So the average pre-processing is only useful to mitigate the exponential increase of the noise.

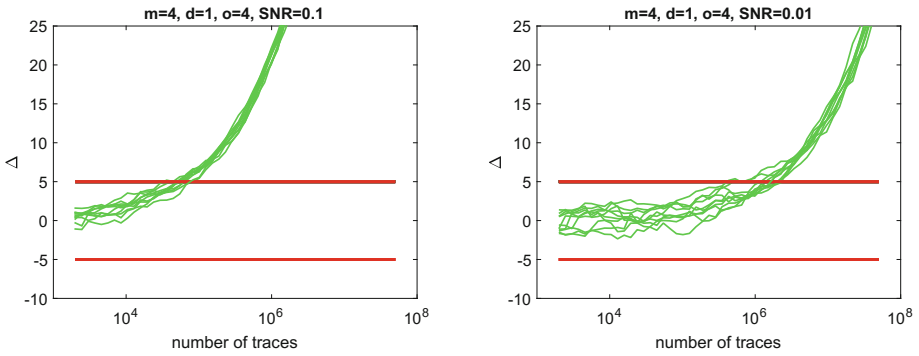

Fig. 5. Results of the TVLA^2 methodology for different noise levels.

Quite naturally, the improvement in this last section can be combined similarly with other statistical tools such as the previously mentioned higher-order DPAs (in [25,30]) or higher-order variations of the CPA (in [21]). In those cases as well, the trick is to take advantage of the masks knowledge in order to pre-process the traces by averaging before estimating higher-order statistical moments. And of course, there as well, the effectiveness of the distinguisher will then only reflect the security order, and be uncorrelated with the attack complexity.

4 Conclusions

Evaluating the security of a leaking device is a challenging problem (see [28] for a recent survey). For the masking countermeasure, it implies to test whether the hypotheses required to deliver its security promises are fulfilled.

The first hypothesis is that the leakage of the shares are independent of each other. Concretely this can be tested by computing a security order, which is the lowest statistical moment of the leakage PDF that depends on the target secret. The TVLA methodology is good for this purpose. Yet, as the security order increases, the exponential amplification of the noise provided by masking renders the sampling complexity of such an approach unreachable. In case the evaluator can access the masks during a profiling phase, it is possible to mitigate this noise amplification, by averaging the leakage traces before computing the security order (i.e., before raising the samples to some power).

Independent of the security order, the second hypothesis is that the leakages are sufficiently noisy. In this respect, the main observation of this note is that launching the TVLA methodology does not allow to guarantee a sufficient noise (since it in fact only tests the security order). This implies that claiming concrete security levels for masked/threshold implementations requires an additional step such as a noise analysis or an information theoretic evaluation with worst-case profiling – an approach that is not yet systematically followed. While it is not a big issue for research works, where claiming a security order is sufficient to indicate that the countermeasure has a potential for noise amplification, it may be a serious limitation for the concrete security evaluations of real products, of which the goal eventually is to determine the number of measurements needed for key recovery (which is a function of the security order and noise level).

In general, our results provide a nice illustration of the separation given in [2]. Namely, "bounded moment security" is a strictly weaker notion than "noisy leakage security", and can only imply it under the necessary condition that the leakages are noisy. More concretely, they also recall that as cryptographic implementations become more and more protected, the gap between (cost-efficient) "conformance/validation-style" testing and (more expensive) "evaluation-style" testing is likely to increase. In this respect, combining conformance/validation-style testing for checking simple properties that implementations have to fulfill "locally" (e.g., a security order and a noise level in the case of masking, or their combination via an information theoretic metric) with more formal approaches to analyze security "globally", such as proposed in [1], seems promising.

As a closing note, we mention that the detection of a security order discussed in this paper is based on univariate statistics. While one may (intuitively) expect that reductions of the security order via glitches, transitions or coupling (as mentioned in Footnote 2) happen mostly at this univariate level, and that increasing the number of dimensions exploited by the adversary will be more prejudicial to the noise level of the implementations, this is certainly something that requires further practical investigations (e.g., by analyzing security order reductions via mixed statistical moments for serial masked implementations – a task for which the tools of Schneider and Moradi in [27] are a good starting point). In this

respect, it is worth observing that most tools used to extend the T-test to multiple samples rely on an independence assumption. Investigating the impact of this assumption is yet another interesting open problem.

Acknowledgments. The author is grateful to Carolyn Whitnall for useful feedback. The author is is an associate researcher of the Belgian Fund for Scientific Research (FNRS-F.R.S.). This work was funded in parts by the ERC project 724725 (acronym SWORD) and by the H2020 project REASSURE.

References

1. Barthe, G., Belaïd, S., Dupressoir, F., Fouque, P.-A., Grégoire, B., Strub, P.-Y.: Verified proofs of higher-order masking. In: Oswald and Fischlin [22], pp. 457–485
2. Barthe, G., Dupressoir, F., Faust, S., Grégoire, B., Standaert, F.-X., Strub, P.-Y.: Parallel implementations of masking schemes and the bounded moment leakage model. In: Coron, J.-S., Nielsen, J.B. (eds.) EUROCRYPT 2017, Part I. LNCS, vol. 10210, pp. 535–566. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-56620-7_19
3. Battistello, A., Coron, J.-S., Prouff, E., Zeitoun, R.: Horizontal side-channel attacks and countermeasures on the ISW masking scheme. In: Gierlichs and Poschmann [12], pp. 23–39
4. Bilgin, B., Gierlichs, B., Nikova, S., Nikov, V., Rijmen, V.: Higher-order threshold implementations. In: Sarkar, P., Iwata, T. (eds.) ASIACRYPT 2014, Part II. LNCS, vol. 8874, pp. 326–343. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-662-45608-8_18
5. Carlet, C., Hasan, M.A., Saraswat, V. (eds.): SPACE 2016. LNCS, vol. 10076. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49445-6
6. De Cnudde, T., Bilgin, B., Reparaz, O., Nikov, V., Nikova, S.: Higher-order threshold implementation of the AES S-Box. In: Homma, N., Medwed, M. (eds.) CARDIS 2015. LNCS, vol. 9514, pp. 259–272. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-31271-2_16
7. De Cnudde, T., Reparaz, O., Bilgin, B., Nikova, S., Nikov, V., Rijmen, V.: Masking AES with d+1 shares in hardware. In: Gierlichs and Poschmann [12], pp. 194–212
8. Cooper, J., De Mulder, E., Goodwill, G., Jaffe, J., Kenworthy, G., Rohatgi, P.: Test vector leakage assessment (TVLA) methodology in practice (extended abstract). In: ICMC 2013 (2013). http://icmc-2013.org/wp/wp-content/uploads/2013/09/goodwillkenworthtestvector.pdf
9. Ding, A.A., Zhang, L., Durvaux, F., Standaert, F.-X., Fei, Y.: Towards sound and optimal leakage detection procedure. In: Eisenbarth, T., Teglia, Y. (eds.) CARDIS 2017. LNCS, vol. 10728, pp. 105–122. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75208-2_7
10. Duc, A., Faust, S., Standaert, F.-X.: Making masking security proofs concrete - or how to evaluate the security of any leaking device. In: Oswald and Fischlin [22], pp. 401–429
11. Durvaux, F., Standaert, F.-X.: From improved leakage detection to the detection of points of interests in leakage traces. In: Fischlin, M., Coron, J.-S. (eds.) EUROCRYPT 2016, Part I. LNCS, vol. 9665, pp. 240–262. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-662-49890-3_10

12. Gierlichs, B., Poschmann, A.Y. (eds.): CHES 2016. LNCS, vol. 9813. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-662-53140-2
13. Goodwill, G., Jun, B., Jaffe, J., Rohatgi, P.: A testing methodology for side channel resistance validation. In: NIST Non-invasive Attack Testing Workshop (2011). http://csrc.nist.gov/news_events/non-invasive-attack-testing-workshop/papers/08_Goodwill.pdf
14. Grosso, V., Standaert, F.-X.: Masking proofs are tight and how to exploit it in security evaluations. In: Nielsen, J.B., Rijmen, V. (eds.) EUROCRYPT 2018, Part II. LNCS, vol. 10821, pp. 385–412. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-78375-8_13
15. Grosso, V., Standaert, F.-X., Prouff, E.: Low entropy masking schemes, revisited. In: Francillon, A., Rohatgi, P. (eds.) CARDIS 2013. LNCS, vol. 8419, pp. 33–43. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-08302-5_3
16. Jaffe, J., Rohatgi, P., Witteman, M.: Efficient side-channel testing for public key algorithms: RSA case study. In: NIST Non-invasive Attack Testing Workshop (2011). http://csrc.nist.gov/news_events/non-invasive-attack-testing-workshop/papers/09_Jaffe.pdf
17. Kamel, D., et al.: Towards securing low-power digital circuits with ultra-low-voltage Vdd randomizers. In: Carlet et al. [5], pp. 233–248
18. Mangard, S., Oswald, E., Standaert, F.-X.: One for all - all for one: unifying standard differential power analysis attacks. IET Inf. Secur. **5**(2), 100–110 (2011)
19. Mather, L., Oswald, E., Bandenburg, J., Wójcik, M.: Does my device leak information? An *a priori* statistical power analysis of leakage detection tests. In: Sako, K., Sarkar, P. (eds.) ASIACRYPT 2013, Part I. LNCS, vol. 8269, pp. 486–505. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-42033-7_25
20. Moradi, A.: Side-channel leakage through static power - should we care about in practice? In: Batina, L., Robshaw, M. (eds.) CHES 2014. LNCS, vol. 8731, pp. 562–579. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-662-44709-3_31
21. Moradi, A., Standaert, F.-X.: Moments-correlating DPA. In: Proceedings of the 2016 ACM Workshop on Theory of Implementation Security, TIS 2016, pp. 5–15. ACM, New York (2016)
22. Oswald, E., Fischlin, M. (eds.): EUROCRYPT 2015, Part I. LNCS, vol. 9056. Springer, Heidelberg (2015). https://doi.org/10.1007/978-3-662-46800-5
23. Poussier, R., Standaert, F.-X., Grosso, V.: Simple key enumeration (and rank estimation) using histograms: an integrated approach. In: Gierlichs and Poschmann [12], pp. 61–81
24. Del Pozo, S.M., Standaert, F.-X., Kamel, D., Moradi, A.: Side-channel attacks from static power: when should we care? In: Nebel, W., Atienza, D. (eds.) Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition, DATE 2015, Grenoble, France, 9–13 March 2015, pp. 145–150. ACM (2015)
25. Prouff, E., Rivain, M., Bevan, R.: Statistical analysis of second order differential power analysis. IEEE Trans. Comput. **58**(6), 799–811 (2009)
26. Roy, D.B., Bhasin, S., Patranabis, S., Mukhopadhyay, D., Guilley, S.: What lies ahead: extending TVLA testing methodology towards success rate. Cryptology ePrint Archive, Report 2016/1152 (2016). http://eprint.iacr.org/2016/1152
27. Schneider, T., Moradi, A.: Leakage assessment methodology - extended version. J. Cryptogr. Eng. **6**(2), 85–99 (2016)
28. Standaert, F.-X.: Towards fair and efficient evaluations of leaking cryptographic devices - overview of the ERC project CRASH, part I (invited talk). In: Carlet et al. [5], pp. 353–362

29. Standaert, F.-X., Malkin, T.G., Yung, M.: A unified framework for the analysis of side-channel key recovery attacks. In: Joux, A. (ed.) EUROCRYPT 2009. LNCS, vol. 5479, pp. 443–461. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-01001-9_26
30. Waddle, J., Wagner, D.: Towards efficient second-order power analysis. In: Joye, M., Quisquater, J.-J. (eds.) CHES 2004. LNCS, vol. 3156, pp. 1–15. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-28632-5_1
31. Welch, B.L.: The generalization of student's problem when several different population variances are involved. Biometrika **34**, 28–35 (1947)