Does My Device Leak Information? An *a priori* Statistical Power Analysis of Leakage Detection Tests

Luke Mather[1], Elisabeth Oswald[1], Joe Bandenburg[2] and Marcin Wójcik[1]

[1] University of Bristol, Department of Computer Science,
Merchant Venturers Building, Woodland Road, BS8 1UB, Bristol, UK
[1]{Luke.Mather,Elisabeth.Oswald,Marcin.Wojcik}@bris.ac.uk,
[2]joe@bandenburg.com

Abstract. The development of a leakage detection testing methodology for the side-channel resistance of cryptographic devices is an issue that has received recent focus from standardisation bodies such as NIST. Statistical techniques such as hypothesis and significance testing appear to be ideally suited for this purpose. In this work we evaluate the candidacy of three such detection tests: a t-test proposed by Cryptography Research Inc., and two mutual information-based tests, one in which data is treated as continuous and one as discrete. Our evaluation investigates three particular areas: statistical power, the effectiveness of multiplicity corrections, and computational complexity. To facilitate a fair comparison we conduct a novel *a priori* statistical power analysis of the three tests in the context of side-channel analysis, finding surprisingly that the continuous mutual information and t-tests exhibit similar levels of power. We also show how the inherently parallel nature of the continuous mutual information test can be leveraged to reduce a large computational cost to insignificant levels. To complement the *a priori* statistical power analysis we include two real-world case studies of the tests applied to software and hardware implementations of the AES.

1 Introduction

The evaluation of the resilience of cryptographic devices against side-channel adversaries is an issue of increasing importance. The potential of side-channel analysis (SCA) as an attack vector is driving the need for standards organisations and governing bodies to establish an acceptance-testing methodology capable of robustly assessing the vulnerability of devices; the National Institute of Standards and Technology (NIST) held a workshop in 2011 driving the requirements [4] and recent papers have been published on this topic by industry [13,16].

Current evaluation methodologies such as Common Criteria [2], used by bodies such as ANSSI [1] and BSI [3], consist of executing a battery of known side-channel attacks on a device and considering whether the attack succeeds and, if so, the quantity of resources expended by an adversary to break the device. This methodology is likely to prove unsustainable in the long-term: the number and type of Simple Power Analysis (SPA), and particularly Differential Power Analysis (DPA) attacks is steadily increasing year-on-year, lengthening the testing process and forcing evaluation bodies to keep up-to-date with an increasingly large, technically complex and diverse number of researched strategies.

A desirable complement or alternative to an attack-focused evaluation strategy is to take a 'black-box' approach; rather than attempting to assess security by trying to find the data or computational complexity of an optimal adversary against a specific device, we can attempt to quantify whether *any* side-channel information is contained in power consumption data about underlying secrets without having to precisely characterise and exploit leakage distributions. We describe this as a *detection* strategy; the question any detection test answers is whether *any* side-channel information is present, and *not* to precisely quantify the exact amount or how much of it is exploitable. Detection-based strategies can be used to support 'pass or fail' type decisions about the security of a device [13], or can be used to identify time points that warrant further investigation.

In practice we *estimate* information leakage, and so any reasonable detection strategy should ideally incorporate a degree of statistical rigour. In this paper we provide a comprehensive evaluation of three leakage detection hypothesis tests in the context of power analysis attacks: a t-test proposed by [13], and two tests for detecting the presence of zero mutual information (MI)—one in which power traces are treated as continuous data (hereafter the CMI test) [10], and one as discrete (hereafter the DMI test) [9].

Our contribution Previous work in the context of side-channel analysis has assessed detection tests through practical experimentation only [13]. This approach creates flawed comparisons of tests for reasons similar to those encountered in the practical analysis of distinguishers in DPA [28]; the effects of sample size and estimation error on detection test performance cannot be quantified in a practical experiment and consequently it becomes difficult to draw fair comparisons that apply in a general context. To ensure a fair comparison in this work we perform an *a priori* statistical power analysis[1] of the three detection tests using a variety of practically relevant side-channel analysis scenarios. The analysis allows us to study the effects that sample size, leakage functions, noise and other hypothesis testing criteria have on the performance of the detection tests in a fair manner. In addition to statistical power, we also investigate the computational complexity of the tests and the effectiveness of multiplicity corrections.

Related work An alternative to the black-box strategy is the 'white-box' leakage evaluation methodology proposed by Standaert et al. [26]. Their methodology re-

[1] The overlap in terminology of the *statistical* power analysis of hypothesis tests with the entirely different differential or simple power analysis technique is unfortunate. To establish a reasonable separation of terminology we will use 'DPA' or 'SPA' to address the latter technique, and 'statistical power' when referencing the former topic.

quires an estimation of the conditional entropy of a device's leakage distribution using an estimated leakage model. This allows for a tighter bound on the amount of information available to an adversary, but requires additional computational expense and the ability to profile a device, and bounding estimation error in the results is non-trivial. The black-box detection approach outlined in this work does not require any device profiling, trading-off the ability to estimate the *exploitable* information leakage contained within the device for efficiency gains and the ability to increase robustness through statistical hypothesis testing. A detection strategy may be used as a complement to the approach of Standaert et al. by identifying a subset of time points that are known to leak information and can be further explored in a white-box analysis.

There is no previous *a priori* power analysis study of these three tests in the context of SCA. A generic analysis of the CMI test and additional non-parametric hypothesis tests was conducted in [10], but does not consider the influence of variables such as noise and leakage function in the context of side-channel analysis, and cannot be used in comparison with the DMI or t-tests.

Organisation In Section 4 of this work we present the results of the first *a priori* statistical power analysis of the three detection tests in the context of side-channel analysis. To support the *a priori* analysis we also provide a case study illustrating an example application of the tests to real-world traces acquired from a software and a hardware implementation of the AES in Section 5. Section 6 discusses the computational complexity of the three tests.

2 Introduction to selected hypothesis tests

2.1 Side-channel analysis

We will consider a 'standard' SCA scenario whereby the power consumption T of a device is dependent on the value of some internal function $f_k(x)$ of plaintexts and secret keys evaluated by the device. Using the random variable $X \in \mathcal{X}$ to represent a plaintext and the random variable $K \in \mathcal{K}$ to represent a sub-key, the power consumption T of the device can be modelled using $T = L \circ f_k(x) + \varepsilon$, where L is a function that describes the data-dependent component of the power consumption and ε represents the remaining component of the power consumption modelled as additive random noise.

2.2 Candidate tests

There are many hypothesis tests that may be used to detect information leakage: one can test for differences between particular moments (such as the mean) of leakage distributions, or one can test for any general differences between leakage distributions. In this work we consider three tests, one from the former category and two from the latter. In the former category, the Welch t-test [27], used to assess the difference between the means of two distributions, has been proposed by Cryptography Research Inc. [13]. One can also analyse higher moments using tests such as the F-test [20]. Information leakage solely occurring in a particular higher moment is rare—to our knowledge, one example of this is in [20]—and so a natural progression is to use a generic non-parametric test instead. Chatzikokolakis et al. and Chothia et al. present hypothesis tests capable of detecting the presence of discrete and continuous mutual information [9,10].

Whilst alternative non-parametric tests are available, mutual information-based methods provide an intuitive measure and are frequently used in other contexts [23,26]. There is a generic *a priori* power analysis comparing the CMI test and additional non-parametric hypothesis tests in [10], finding that the CMI test compared favourably. The analysis does not discuss any of the side-channel specific variables described in Section 2.1 and cannot be used in comparison with the t-test, but does suggests that an MI-based test is a natural choice for a generic test candidate. As such, we focus on the t-test and the two MI-based methods, and note that our evaluation strategy can be easily applied to other detection tests in the future.

The null hypothesis for any hypothesis testing procedure used in a detection context is that there is no information leakage: using the t-test, any statistically significant difference of means is evidence for an information leak, and using MI-based tests, any significant non-zero mutual information is evidence.

The generic strategy followed by each test is to systematically evaluate each individual time point in a set of traces in turn. This is a 'univariate' approach, and in many cases is likely to be sufficient; vulnerabilities arising from sub-optimal security measures are likely to manifest themselves as leakage detectable within a single time point. To detect leakage exploitable by n-th order attacks would necessitate the joint comparison of n time points. This results in a considerable increase on the the amount of computation required—the brute force strategy would be to analyse the joint distribution of every possible n-tuple of points—and additionally can substantially increase the complexity of the test statistics, with multivariate mutual information in particular becoming costly. Whilst an efficient multivariate strategy would be desirable, it is beyond the scope of this initial work.

2.3 Difference-of-means and the t-test

Exploiting the difference-of-means $\overline{T_1} - \overline{T_2}$ between two sets of power traces T_1 and T_2 partitioned on a single bit of a targeted intermediate state was proposed by Kocher et al. and is the canonical example of a generic DPA attack [17]. The same difference-of-means can also be used to detect information leakage, and was proposed as a candidate detection test in [13].

Welch's t-test is a hypothesis test that (in the two-tailed case) tests the null hypothesis that the population means of two variables are equal, where the variables have possibly unequal variances, yielding a p-value that may or may not provide sufficient evidence to reject this hypothesis. The test statistic t is:

$$t = \frac{\overline{T_1} - \overline{T_2}}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}, \tag{1}$$

where $\overline{T_i}$, s_i^2 and N_i are the sample means, sample variances and sample size of the i-th set T_i. Using this test statistic and the Welch-Satterthwaite equation[2] to compute the degrees of freedom ν, a p-value can be computed to determine whether there is sufficient evidence to reject the null hypothesis at a particular significance level $1 - \alpha$. Using the quantile function for the t distribution at a significance level α and with ν degrees of freedom, a confidence interval for the difference-of-means can also be computed.

Leveraging the t-test requires a partitioning of the traces based on the value of a particular bit of an intermediate state with the targeted algorithm, and therefore to comprehensively evaluate a device every single bit of every single intermediate state must be tested. To assess the i-th bit of a particular state for leakage (e.g. the output of SubBytes in a particular round), an evaluator must compute the intermediate values for the chosen state, using a set of chosen messages. Having recorded the encryption or decryption of the chosen messages, the resulting traces can be partitioned into two sets T_1 and T_2, depending on the value of the i-th bit of the intermediate state. The test statistic t and corresponding p-values or confidence intervals can then be used to determine whether a difference between the means exists.

The t-test by design can only detect differences between subkeys that are contained within the mean of the leakage samples, and assumes that the populations being compared are normally distributed. In practice univariate leakage from unprotected devices is typically close enough to Gaussian for this condition to not be too restrictive [7,8,17].

2.4 Mutual information

Given two random variables X and Y, the MI $\mathrm{I}(X;Y)$ computes the *average* information gained about X if we observe Y (and vice-versa). The application of MI to detecting information leaks from a cryptographic device is straightforward: any dependence between subkeys and the power consumed by the device, giving $\mathrm{I}(K;T) > 0$, may be evidence for an exploitable information leak[3].

The rationale for using MI to detect information leaks is that it compares distributions in a general way, incorporating all linear and non-linear dependencies between sub-keys and power values. Unfortunately, the estimation of MI is well-known to be a difficult problem. There are no unbiased estimators, and

[2] Using Welch-Satterthwaite, the degrees of freedom ν for a t-distribution can be calculated as $\nu = \frac{(s_1^2/N_1 + s_2^2/N_2)^2}{(s_1^2/N_1)^2/(N_1-1) + (s_2^2/N_2)^2/(N_2-1)}$.

[3] Under the assumption of the ‘equal images under different sub-keys’ property [24] we can safely compute $\mathrm{I}(X;T)$, if simpler.

it has been proven that there is no estimator that does not perform differently depending on the underlying structure of the data [22].

Recent results on the behaviour of zero MI can help to alleviate this problem. Chatzikokolakis et al. find the sampling distribution of MI between two discrete random variables when it is zero, where the distribution of one of the variables is known and the other unknown, and use this to construct a confidence interval test [9]. A second result from Chothia and Guha establishes a rate of convergence, under reasonable assumptions, for the sampled estimate for zero MI between one discrete random variable with a known distribution and one *continuous* random variable with an unknown distribution [10]. This result is then used to construct a non-parametric hypothesis test to assess whether sampled data provides evidence of an information leak within a system.

Discrete mutual information As side-channel measurements are typically sampled using digital equipment, it may be viable to treat the sampled data as discrete. The most common way to make continuous data discrete is to split the continuous domain into a finite number of bins. Using the standard formula for marginal and conditional entropy, the discrete MI estimate can be computed as

$$\hat{\mathrm{I}}(K;T) = \sum_{k \in \mathcal{K}} \sum_{t \in \mathcal{T}} \hat{p}(k,t) \log_2 \left(\frac{\hat{p}(k,t)}{p(k)\hat{p}(t)} \right). \tag{2}$$

The test of Chatzikokolakis et al. is biased by $(I-1)(J-1)/2n$, where I and J are the sizes of the distribution domains of two random variables in question, and n is the number of samples acquired. In our context, $I = |\mathcal{K}|$, the number of possible sub-keys, and $J = |\mathcal{T}|$, the number of possible power values as a result of discretisation. Consequently, the point estimate e for MI is the estimated value minus this bias: $e = \hat{\mathrm{I}}(K;T) - (I-1)(J-1)/2n$. We can use this to compute $100(1-\alpha)\%$ confidence intervals for zero and non-zero MI (full details can be found in [9]).

As a result of the bias of the test, to be sure of good results it is necessary to ensure that the number of traces sampled is larger than the product of the number of sub-keys and the number of possible power values. The applicability of this discrete test is then dictated by the ability of an evaluator to sample enough traces to meet this condition.

Continuous mutual information The test of Chothia and Guha requires two assumptions about the data to guarantee a convergence result for zero MI [10]. The first is that the power values are continuous, real-valued random variables with finite support. This may or may not hold theoretically, depending on the distribution of the leakages, but in practice will be true; the sampling resolution used dictates the range of the recorded power consumption. The second is that for $u = \{0,1\}$, the probability $p(u,t)$ must have a continuous bounded second derivative in t. This can be fulfilled with the leakage analysis of a single bit of a key only. However, Chothia and Guha also demonstrate experimentally that the test works well in cases of multiple inputs, often outperforming other two-sample tests [10].

Under the assumption of a continuous leakage distribution, we are estimating a hybrid version of the MI:

$$\hat{\mathrm{I}}(K;T) = \sum_{k \in \mathcal{K}} \int_T \hat{p}(k,t) \log_2 \left(\frac{\hat{p}(k,t)}{p(k)\hat{p}(t)} \right) dt. \tag{3}$$

To compute this estimate we are required to estimate a conditional probability density function $\hat{\Pr}\{t|k\}$ using kernel density estimation. The assumptions underlying the test's convergence result dictate the use of a function such as the Epanechnikov kernel[4] as the chosen kernel function, and a bandwidth function such as Silverman's [25] general purpose bandwidth[5].

Using this estimated density function, we can compute an estimate of the MI, $\hat{\mathrm{I}}(K;T)$. The following step of the hypothesis test is a permutation stage requiring s permutations of the sampled data T': for each sampled power value, we randomly assign a new sub-key to the value without replacement. The power values contained in each permuted set should now have no relation with the sub-keys, and so the MI of the s sets can now be computed $\hat{\mathrm{I}}_1(K;T'_1), \ldots, \hat{\mathrm{I}}_s(K;T'_s)$, providing a baseline for zero MI.

An *estimated* p-value can be computed by computing the percentage of the MI estimates $\hat{\mathrm{I}}_1, \ldots, \hat{\mathrm{I}}_s$ that have a value greater than the observed point estimate $\hat{\mathrm{I}}(K;T)$. The suggested number of shuffled estimates to achieve useful baseline results is given to be 100 by Chothia and Guha, but to increase the power of the test and the precision of the estimated p-values a few thousand shuffles may be required.

3 Evaluation methodology

3.1 Comparing detection tests

The most important notion in hypothesis testing is of the quantification and classification of the error involved. The type I error rate α is defined as the probability of incorrectly rejecting a true null hypothesis, usually termed the significance criterion. Tests are also associated with a type II error rate β: the probability of failing to reject a false null hypothesis. The exact valuation assigned to these error rates is an important factor to balance; typically decreasing one error rate will result in an increase in the other, and the only way to reduce both in tandem is to increase the sample size available to the test. The statistical power of a test is defined as the probability of correctly rejecting a false null hypothesis, $\pi = 1 - \beta$. This is the key factor for our detection tests: higher statistical power indicates increased robustness and lessens reliance on large sample sizes.

[4] Epanechikov's kernel function is defined as $K(u) = 3/4(1-u^2)_{\chi\{|u|\leq 1\}}$.

[5] $h = 1.06 s_T N^{-1/5}$, where s_T is the sample standard deviation of T and N is the number of sampled traces.

A common motivation for performing an *a priori* statistical power analysis[6] is to compute or estimate the minimum sample size required to detect an *effect* of a given size, or to determine the minimum effect size a test is likely to detect when supplied with a particular sample size. The determination of sample sizes required to achieve acceptable power has two-fold uses: firstly, data acquisition from a cryptographic device is an expensive and time-consuming operation, and so tests that are less data-hungry are likely to be preferable, and secondly, knowledge of the sample sizes required to detect a particular effect can serve as a guideline for evaluators to determine the number of trace acquisitions sufficient for detecting an information leak.

3.2 Multiple testing

When considering the results of large numbers of simultaneously-computed hypothesis tests, we must take into account that the probability a single test falsely rejects the null hypothesis will increase in proportion with the number of tests computed. A single test computed at significance level $\alpha = 0.05$ has a 5% chance of incorrectly rejecting the null hypothesis; when conducting a large number of simultaneous tests the probability of a false positive increases. The intuitive solution is to control the overall false rejection rate by selecting a smaller significance level for each test. There are two main classes of procedure: controlling the *familywise error rate* (FWER) and controlling the *false discovery rate* (FDR).

Familywise error rate The FWER is defined as the probability of falsely rejecting one or more true null hypotheses (one or more type I errors) across a family of hypothesis tests. The FWER can be controlled, allowing us to bound the number of false null hypothesis rejections we are willing to make—in our device evaluation context this would allow the evaluator to control the probability a device is falsely rejected. FWER controlling procedures are conservative, and typically trade-off FWER for increasing type II error.

False discovery rate Proposed by Benjamini and Hochberg in 2005, the FDR is defined as the *expected* proportion of false positives (false discoveries) within the hypothesis tests that are found to be significant (all discoveries). Procedures that control the FDR are typically less stringent than FWER-based methods, and have a strong candidacy for situations where test power is important. The Benjamini-Hochberg (BH) procedure is a 'step-up' method that strongly controls the FDR at a rate α [6]. Given m simultaneous hypothesis tests, the BH procedure sorts the p-values and selects the largest k such that $p_k \leq \frac{k}{m}\alpha$, where all tests with p-values less than or equal to p_k can be rejected. Many additional FWER and FDR controlling methods exist, e.g. [14,15], but are beyond the scope of this paper.

A trade-off with multiplicity corrections that control the FWER is that generally decreasing the FWER results in an increase in type II error. As a consequence the FDR approach may be more suitable if an evaluator is particularly

[6] For further discussion of statistical power analysis, see [11].

concerned with ensuring that the type II error rate is kept low—that the statistical power remains high. It may also serve a useful purpose by identifying a small candidate set of time points that are *likely* to contain information leakage—the evaluator can then perform further analysis on the set of points, for example by inspecting the effect sizes reported for each of the points, re-sampling additional data and performing new hypothesis tests, or even by trying to attack the points using an appropriate method. We demonstrate an example application of the BH procedure in Section 5.

3.3 Why perform an *a priori* power analysis?

Having established the importance of statistical power to our detection tests, the motivation for performing an *a priori* power analysis for our three candidate tests is that it is not possible to make generally true inferences based on practical experiments alone; given that it is only possible to establish the vulnerability of a time point by successfully attacking it, it becomes impossible to establish whether a reported rejection of the null hypothesis is a false positive—in other words, the type II error rate β cannot be estimated—and hence any *a posteriori* (or post-hoc) power analysis is likely to be misleading.

To be able to perform an *a priori* statistical power analysis, we need to be able to produce or simulate data, ideally with characteristics as close as possible to those observed in practice, for which we are sure of the presence of information leakage. The most straightforward way to do this is to simulate trace data under the 'standard' DPA model commonly used throughout the existing body of literature, detailed in Section 2.1.

4 *A priori* power analysis

As all of the variables in the standard SCA model outlined in Section 2.1 have an effect on detection test performance, to perform a useful *a priori* power analysis we defined a variety of leakage scenarios that have relevance to practice, and then estimated the power π of each of the detection tests under many combinations of the different parameters in the SCA model for each scenario. For each leakage scenario, power was estimated under varying sample sizes, noise levels and using two different significance criteria: $\alpha = 0.05$ and $\alpha = 0.00001$. The former provides a general indication of test power with a common level of significance, and the intention with the latter level of significance is to gain an understanding of how much statistical power is degraded by the typical tightening of the significance criteria enforced by multiple testing corrections.

Leakage model We defined five different practically-relevant leakage models L under which to simulate trace data:

1. Hamming weight—a standard model under which the device leaks the Hamming weight of the intermediate state;

2. Weighted sum—the device leaks an unevenly weighted sum of the bits of the intermediate state, where the least significant bit (LSB) dominates with a relative weight of 10, as motivated by [5];
3. Toggle count—the power consumption of hardware implementations has been shown to depend on the number of *transitions* that occur in the S-Box. The model used here is computed from back-annotated netlists as in [19], and creates non-linear leakage distributions;
4. Zero value—for this model we set the power consumption for every non-zero intermediate value to be 1, and for the value zero we set the power consumption to be 0; this will typically produce small amounts of information leakage and should stress the data efficiency of the tests;
5. Variance—the mean of the power consumption does not leak, and the variance of the power consumption follows the distribution given in Maghrebi et al. [18]. The t-test will not be able to detect any leakage, but the model can be used to evaluate the relative performances of the MI tests.

A statistical power analysis would ideally be performed for each candidate target function; given the limited space available we have focused on the AES. For this comparison we targeted, without loss of generality, the first byte of the key. For each leakage model, we simulated traces under a wide range of signal-to-noise ratios (SNRs), ranging from 2^{-14} to 2^{12}, enabling us to assess the maximum amount of noise a test can overcome when provided with a particular sample size.

Estimation process The estimated power for the test is computed as the fraction of times the test correctly[7] rejects the null hypothesis for $1,000$ tests run. For the CMI and t-tests we used the significance criterion α to determine rejection or acceptance, and for the DMI test we checked whether the corrected estimate for the MI was inside the $100(1-\alpha)\%$ confidence interval for zero MI.

In the following section we present the results of our *a priori* statistical power analysis on the five leakage models in terms of the number of samples required to achieve 80% power for each combination of model, SNR and sample size. We performed $1,000$ permutations of the simulated traces for each CMI test, and used the Epanechnikov kernel with Silverman's bandwidth for the kernel density estimation. To enable a fair comparison between the bit and byte level tests, we chose to represent the results for the t-test corresponding to the most leaky bit of the state. Graphs illustrating the number of samples required by each test to achieve 80% power for each leakage model and SNR are shown in Figure 1.

Hamming weight We can see that the t-test is the most powerful test in general, as we would expect given the unbiased estimator for the mean values and the Gaussian noise assumption holding true in the model. The CMI test

[7] Each of these scenarios contain information leakage; even for the extremely low SNRs, given sufficiently large data an attacker will eventually be able to exploit the leakage, and as a consequence candidate detection tests should, for some level of sample size, be able to consistently detect information leakage.

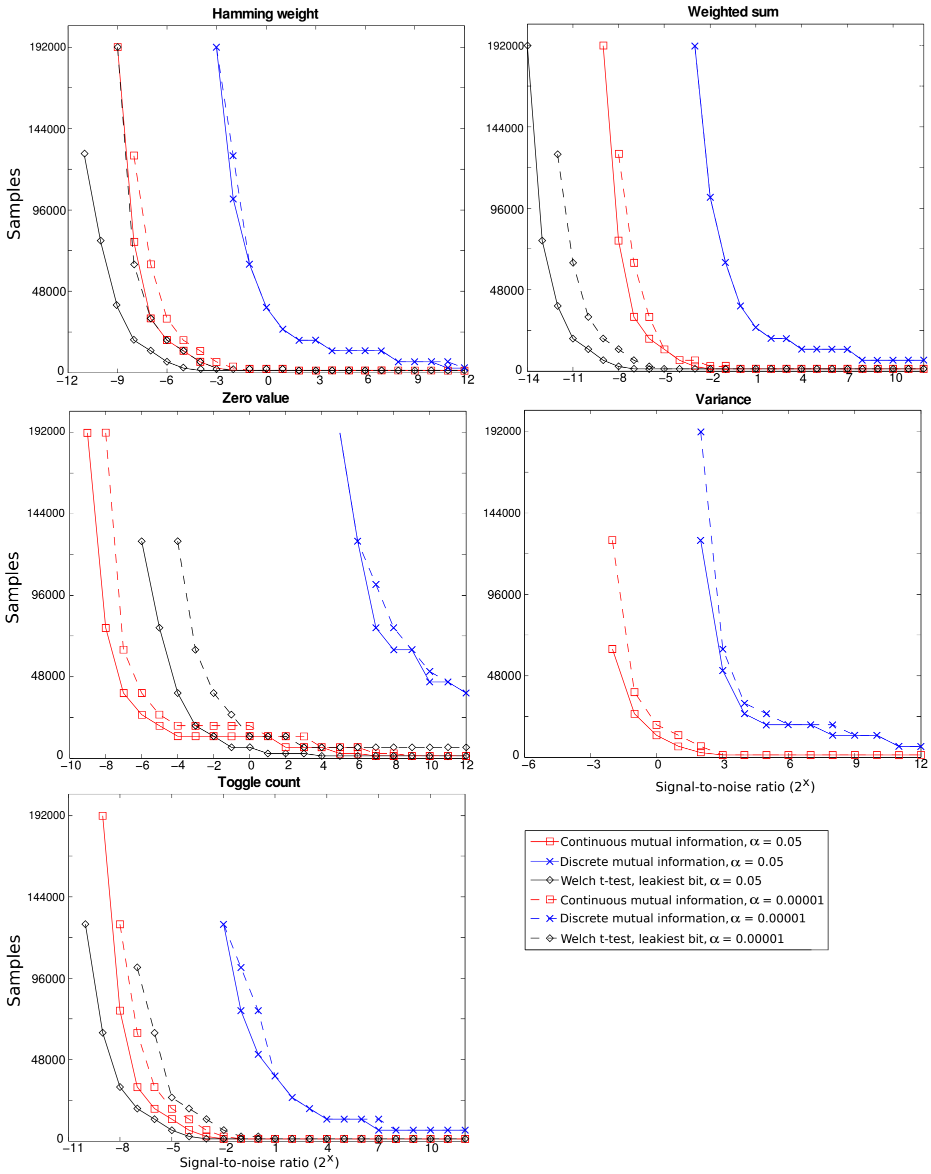

Fig. 1: *Number of samples required for the* t*-test, CMI and DMI tests to achieve estimated* 80% *power for a variety of leakage models and SNRs.*

requires slightly more samples to achieve the requisite power in the presence of high noise, and both tests seem to perform equivalently for mid-range and low levels of noise.

The DMI test appears to be significantly less powerful; this is unsurprising given a loss of information from the treatment of continuous data as discrete is to be expected, and we also see that the test struggles to cope with high levels of noise—the lowest SNR for which we could detect an information leak with up to 192,000 samples was 2^{-3}. A closer inspection indicates that this is caused by the bias correction required; the size of the input space for the AES often necessitates a large sample size to minimise the size of the correction to within manageable bounds.

The stricter significance criterion $\alpha = 0.00001$ seems to have a small but noticeable effect on the test power for the CMI and t-tests. Under the DMI test we see little change in behaviour; the dominant factor influencing power is the bias correction rather than the precise width of the confidence intervals.

Weighted sum The relative dominance of the LSB in the leakage provides an additional advantage for the t-test and we found as expected that the test achieved its highest power when evaluating this bit. This results in a relative increase in overall power compared to the CMI test than we observed in the Hamming weight scenario and also allows for detection of leakage at lower SNRs. The CMI test seems to exhibit performance consistent with that under the Hamming weight model, and similarly for the DMI test. The effects of the stricter significance criterion are also similar, with noticeable reductions in power observed for each of the tests under the smaller α values save for the DMI test, where again the bias correction is the predominant factor.

Toggle count An analysis of the underlying true distance of means for the Toggle count model indicated that the largest information leakage was contained within the second-least significant bit, which was also twice the leakage in the next most leaky bit. As with the Weighted sum model, the relative dominance of this bit supplies the t-test with an advantage over the CMI test but in this instance the advantage is by a smaller margin. We can also see that the CMI test appears to be significantly more robust to the stricter significance criterion, outperforming the more sensitive t-test in all of the high noise settings. Here we also see the DMI test exhibiting an increased sensitivity to the significance criterion.

Zero value The size of the information leak present in a noise-free setting for the Zero value model is small relative to those in the other models: the true MI in a noise-free setting is 0.0369 and the true distance-of-means 0.0078. As such it is interesting to note the stronger performance of the CMI test in high noise settings relative to that of the t-test observed in these results—the additional information on the non-linear dependencies contained in the estimated MI values increases the power of the CMI test whereas the quantity of noise has a stronger effect on the difference-in-means estimated by the t-test. The low power estimates for the DMI test are consistent with the small size of the information leak in the model coupled with the loss of information in the conversion process of continuous to discrete data.

VARIANCE By design the mean of the power consumption for all sub-key values is equivalent in the VARIANCE model, and so the t-test cannot be applied. As a test for the applicability of the CMI and DMI to situations in which only higher-order moments leak, the CMI test appears to be robust, so that small sample sizes suffice to achieve the requisite power at medium and low noise levels. The true information leakage contained within the variances is strongly affected by the amount of noise in the samples, which explains why both tests soon begin to struggle as the SNR drops below 2^0.

Conclusion The t-test was generally shown by the *a priori* power analysis to be the most powerful. This is not unexpected: the sample mean is a consistent, unbiased estimator for the population mean and converges quickly to the true value. The performance of the CMI test was close to that of the t-test in all scenarios, indicating that it remains a robust, if slightly inferior alternative in the majority of settings. The DMI test was expected to be less powerful due to the loss of information by the conversion of continuous data to discrete, and this was observed in our analysis; the results indicate that the test is a viable choice only when supplied with large amounts of trace data and only when the SNR is high.

Of note was the superior performance of the CMI test when detecting the small leaks produced by our ZERO VALUE model, particularly in high-noise settings. This suggests that the CMI test may be a better, or safer, choice when applied to devices with these sorts of characteristics. The results obtained under the VARIANCE model indicate that the CMI test is sufficiently robust to handle 'tough' leakage scenarios in which the leakage is solely contained in higher moments of the power consumption distribution.

5 Case studies

The *a priori* statistical power analysis is the primary method for comparison of the detection tests. To complement the analysis, and to further explore the effectiveness of multiplicity corrections, in the following section we demonstrate the application of the three detection tests to the evaluation of two cryptographic devices implementing the AES. The first device we analyse is an ARM7 microcontroller implementing the AES in software, with no countermeasures applied. This device would be expected to exhibit significant information leakage in Hamming-weight form, and hence is a good opportunity to analyse the efficacy of multiple testing correction procedures. The second device analysed is a Sasebo-R evaluation board manufactured using a 90nm process implementing AES in hardware with a Positive-Prime Reed-Muller (PPRM) based SubBytes operation using single-stage AND-XOR logic [21]. This second case study is intended to investigate the performance of the detection tests under increasingly complex leakage distributions as well as acting as a further test for the multiplicity corrections.

5.1 ARM7 microcontroller

Our data set contained $32,000$ traces from the device and we chose to evaluate the first key byte for information leakage. For the t-test we analysed the output of the first SubBytes operation. Figure 2 illustrates the estimated MI values and t-test statistics produced by the detection tests ran at a significance level $\alpha = 0.05$ for each of the $200,000$ time points in our traces. For the CMI test we performed $1,000$ permutations of the traces at each time point, and as we found that all 8 of the bits in the intermediate state produced similar information leakage we elected to display the results for the LSB.

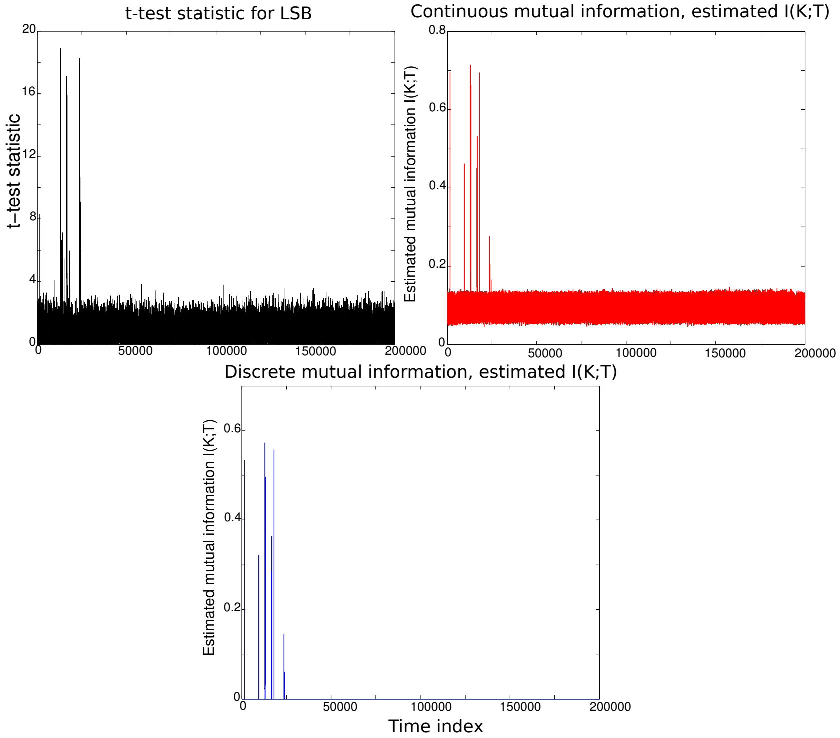

Fig. 2: *Estimated CMI and DMI values and* t*-test statistics produced using 32,000 traces during an evaluation of an ARM7 microcontroller implementing a software version of the AES.*

At the initial significance level $\alpha = 0.05$, the CMI test identified $9,360$ time points consistent with information leakage, the discrete test 178, and the t-test $9,713$. These occur across the full range of the traces, and account for around 4.8% of the total in the CMI and t-test cases. Using our prior knowledge of the device we could ascertain that many of these points are likely to be false positives.

To gain an *indication* of how many of these time points actually contain exploitable leakage, we conducted a battery of attacks on the output of the SubBytes operation on all of the time points using the same set of traces including Brier et al.'s correlation (CPA) [7], Gierlichs et al.'s mutual information analysis (MIA) [12], both using a Hamming weight power model, and Kocher et al.'s difference of means [17]. Whilst we have argued that practical results should not be used to perform a *post hoc* power analysis, the results of the DPA attacks can be used to quantify under-performances of the three tests—time points that *can* be successfully attacked that are missed by detection tests are indicative of low statistical power given the available sample size. In this regard the only notable false acceptances of time points occurred under the DMI test, with the CMI and t-tests able to spot the vast majority of the vulnerable time points. These results appear to be consistent with those observed under the Hamming-weight scenario in the statistical *a priori* power analysis.

False discovery rate Applying any correction to the results produced by the DMI test is redundant as the 'raw' results are already highly unlikely to contain falsely rejected null hypotheses. The FDR controlling procedures are likely to be the most successful of the multiple testing corrections for our purposes, so we applied the Benjamini-Hochberg correction to the results produced by the CMI and t-tests, controlling the FDR at the levels 0.05 and 0.5. Using prior knowledge of the device and the results of the DPA attacks we would not expect to observe any information leaked about the first key byte after time 25,000.

The effect of increasing the value of the FDR on the type I error can be observed by the larger number of false positives produced when the FDR is 0.5. The t-test appears to react more effectively to the corrective procedure, eliminating larger numbers of the false positives previously observed at the time points greater than $25,000$. An inspection of the p-values reported by the CMI test indicates that the number of permutations performed is the proximate cause for the under-performance: the $1,000$ executed do not appear to produce enough precision in the estimated p-values to allow the step-up procedure to differentiate between neighbouring tests. The procedures do not appear to result in a significant rise in type II error—the increase is lessened with the looser FDR of 0.5, but appears to be slight in both cases. As always, increasing the sample size available would reduce the size of any increase in type II error.

5.2 Hardware AES with PPRM SubBytes implementation

The dataset contained $79,360$ traces from the device at 5 giga-samples per second and we again chose to evaluate the first key byte for information leakage; for the t-test we analysed the output of the first SubBytes operation. Figure 4 illustrates the estimated MI values and t-test statistics produced by the detection tests run at a significance level $\alpha = 0.05$ for each of the $50,000$ time points in our traces. The first and last $10,000$ points are not displayed as they do not correspond to any part of the full AES operation. For the CMI test we increased the number of permutations to $10,000$ per time point in an attempt to gain additional precision

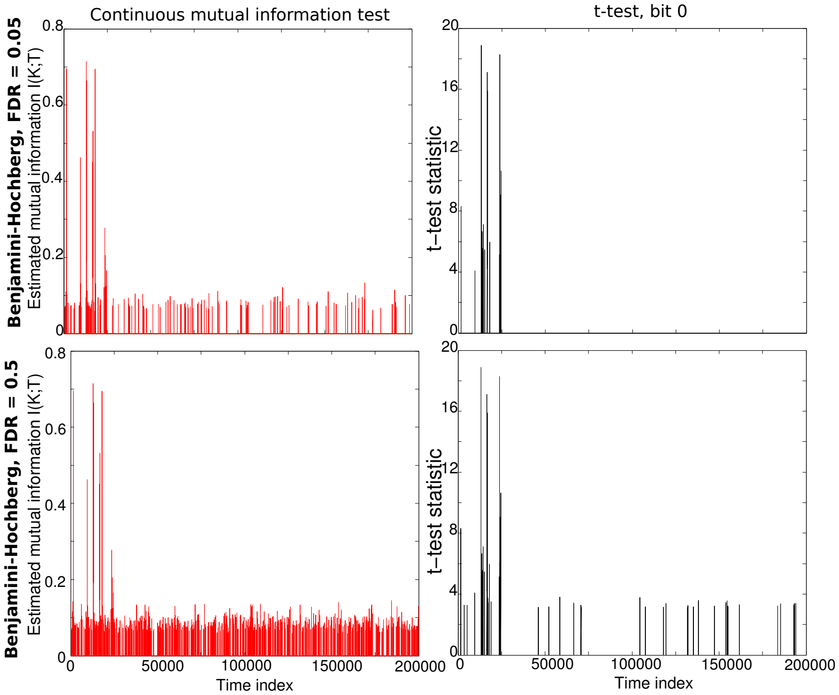

Fig. 3: *Plots of the time points consistent with information leakage after applying the Benjamini-Hochberg FDR controlling procedure to the results produced by the* t*-test and CMI test.*

on the estimated p-values. Information leakage was found to occur to a varying degree across all 8 bits of the intermediate state when using the t-test—as such, we have elected to superimpose the results for all of the state bits on a single graph. The DMI test was not able to identify any information leakage.

A visual inspection of the results produced by both the CMI test and t-tests indicate that there are 10 groups of points within the power traces that contain significant amounts of information leakage. As would be expected the shape and scale of the leakages differ: the t-test is only assessing the SubBytes operation *and* the leakage of individual bits. We were able to confirm the vulnerability of the device by successfully executing a reduced Bayesian template attack on the intermediate values of the SubBytes operation at the time points the detection tests indicated would be vulnerable. The hardware device exhibits less, but still significant leaking behaviour when compared to the ARM7 microcontroller implementation, as evidenced by the lower mutual information estimates and the smaller t-test statistic scores.

The performance of the CMI and t-tests appears to be similar. The extra definition in the CMI graph is likely due to the t-test assessing leakage from the output of the SubBytes operation only. The DMI test could not identify any

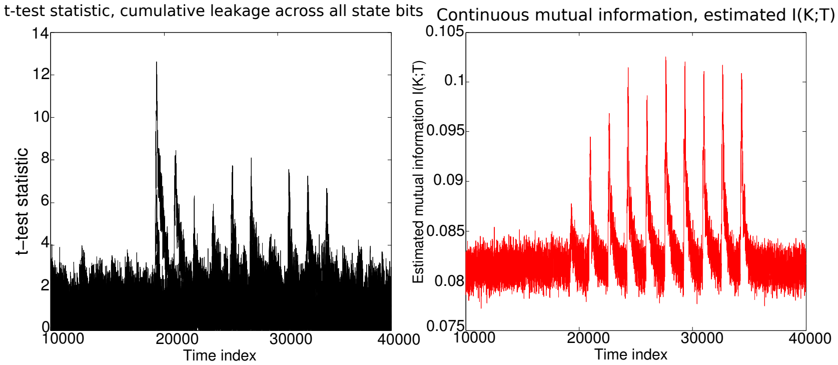

Fig. 4: *Estimated $\hat{I}(K;T)$ values produced by the CMI test and* t*-test statistics produced using* 79, 360 *traces taken from an evaluation of a hardware AES device with the SubBytes operation using Positive-Prime-Reed-Muller (PPRM) logic.*

information leakage, indicating that many more samples would be required to begin to match the power of the CMI and t-tests.

False discovery rate The Benjamini-Hochberg correction was applied to the results produced by the CMI and t-tests, this time controlling the FDR at the levels 0.05 and 0.005. The previous FDR of 0.5 used in the analysis of the ARM7 device yielded too many clear false rejections of the null hypothesis, possibly due to the smaller number of time points, and as a consequence two stricter criteria were used. Figure 5 shows the results of applying the two criteria to the results produced by the CMI and t-test. The effectiveness of the multiplicity corrections is lessened in the hardware device evaluation. The t-test again reacts better to the stricter corrective procedure, eliminating larger numbers of likely false positives. Despite the increase of permutations per time point from 1, 000 to 10, 000 for the CMI test, the effectiveness of the multiplicity correction is again dampened by the lack of precision available in the estimated p-values. It is likely that a different, more complex approach may be required to effectively mitigate the multiplicity problem under the CMI test.

6 Computational complexity

If we consider commercial and logistical pressures on the evaluation process then we must also include the computational complexity of the detection tests as a factor in our evaluation. In this regard, the CMI test is particularly expensive. Under reasonable parameters of a data set of 80, 000 traces each consisting of 50, 000 sampled time points, and where the test computes 1, 000 permuted estimates of the MI at each time point, a full run of the detection test on a single key byte necessitates the evaluation of 50 million continuous MI values. If we factor in the cost of finding conditional probability density functions, then we

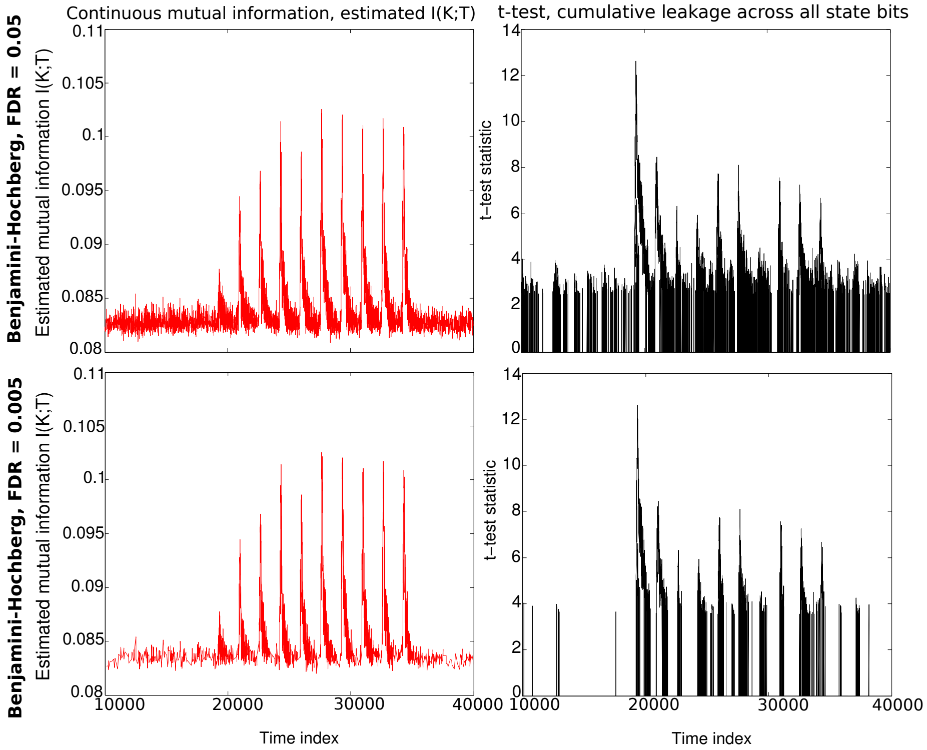

Fig. 5: *Plots of the time points consistent with information leakage after applying the Benjamini-Hochberg FDR controlling procedure at levels* 0.05 *and* 0.005 *to the results produced by the* t*-test and CMI test for the hardware AES implementation.*

may expect to perform in total 2.05×10^{15} ($\approx 2^{51}$) evaluations of the kernel function used in the density estimation, at a total cost of roughly 1.64×10^{16} floating point operations.

This presents a significant obstacle; we estimated that our naive single-CPU implementation would take around a month to analyse a device. However the problem is ‘embarrassingly parallel’ and we implemented the test in parallel form using OpenCL: using two AMD Radeon 7970 GPUs we were able to execute a test with the above parameters in approximately 14 hours; a throughput of 300 GFLOPS. The addition of inexpensive GPUs decreases the running time linearly, ensuring that the CMI test, even with large data set parameters, is feasible to run. By comparison the DMI and t-tests are efficient; a key byte can be fully assessed for leakage in under 30 minutes.

7 Conclusion

Taking the perspective of a 'black-box' evaluation, in which the evaluator may have little knowledge about the leakage characteristics of the device, it would be desirable to select a leakage detection test that is the most generally applicable and that has the best all-round performance. In the majority of our *a priori* analysis this was, by a small margin, the t-test. However we must also take into account the inherent limitations in the t-test's inability to measure leakage in any moment other than the mean. If an evaluator wished to gain the most coverage over *all* possible leakage scenarios, then, given the significant under-performance of the discrete version in the *a priori* analysis, the CMI test is the *only* viable candidate.

The complexity of the tests is an additional factor to consider. The t-test must be re-run for every bit and every intermediate operation within the algorithm implemented on the device, whereas the CMI and DMI tests need only to be run once per bit or byte of key analysed. At first glance the computational cost of the CMI test appears to be prohibitive, but we have demonstrated that using relatively inexpensive GPUs and the inherently parallel nature of the problem, the running time can easily and cheaply be reduced to insignificant levels.

In the absence of any general result that can translate MI, entropy or a difference of means into the trace requirements for an adversary, the interpretation of the results of any standardised detection test becomes heavily reliant on the tools provided by statistics. The large body of work on multiplicity corrections is a rich resource to draw upon, and further research in this area may yield useful results. In addition, a multivariate detection procedure capable of detecting any higher-order information leakage warrants research effort.

Acknowledgements The authors would like to acknowledge the help of the anonymous reviewers whose comments helped to improve the paper. This work has been supported in part by EPSRC via grant EP/I005226/1.

References

1. Agence nationale de la sécurité des systèmes d'information (ANSSI). `http://www.ssi.gouv.fr/en/products/certified-products`. Accessed 25 Feb 2013.
2. Common Criteria v3.1 Release 4. `http://www.commoncriteriaportal.org/cc/`. Accessed 25 Feb 2013.
3. Federal Office for Information Security (BSI)—Common Criteria for examination and evaluation of IT security. `https://www.bsi.bund.de/ContentBSI/EN/Topics/CommonCriteria/commoncriteria.html`. Accessed 25 Feb 2013.
4. National Institute of Standards and Technology: Non-Invasive Attack Testing Workshop. `http://csrc.nist.gov/news_events/non-invasive-attack-testing-workshop`, 2011. Accessed 25 Feb 2013.
5. Mehdi-Laurent Akkar, Régis Bevan, Paul Dischamp, and Didier Moyart. Power Analysis, What Is Now Possible... In *ASIACRYPT 2000*, LNCS, vol. 1976, pp. 489–502, Springer, Heidelberg (2000).

6. Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300 (1995).
7. Eric Brier, Christophe Clavier, and Francis Olivier. Correlation Power Analysis with a Leakage Model. In *CHES 2004*, LNCS, vol. 3156, pp. 16–29, Springer, Heidelberg (2004).
8. Suresh Chari, Josyula R. Rao, and Pankaj Rohatgi. Template Attacks. In *CHES 2002*, LNCS, vol. 2523, pp. 13–28, Springer, Heidelberg (2002).
9. Konstantinos Chatzikokolakis, Tom Chothia, and Apratim Guha. Statistical Measurement of Information Leakage. In *TACAS 2010*, pp. 390–404, LNCS, vol. 6015, Springer (2010).
10. Tom Chothia and Apratim Guha. A Statistical Test for Information Leaks Using Continuous Mutual Information. In *CSF*, pp. 177–190, IEEE Computer Society (2011).
11. Paul D. Ellis. The Essential Guide to Effect Sizes: An Introduction to Statistical Power, Meta-Analysis and the Interpretation of Research Results. Cambridge University Press, United Kingdom (2010).
12. Benedikt Gierlichs, Lejla Batina, Pim Tuyls, and Bart Preneel. Mutual Information Analysis. In *CHES 2008*, LNCS, vol. 5154, pp. 426–442, Springer, Heidelberg (2008).
13. Gilbert Goodwill, Benjamin Jun, Josh Jaffe, and Pankaj Rohatgi. A Testing Methodology for Side-Channel Resistance Validation. In *NIST Non-invasive attack testing workshop* (2011).
14. Y. Hochberg and A. C. Tamhane. *Multiple Comparison Procedures.* John Wiley & Sons, Inc., New York, NY, USA (1987).
15. Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 2(6):65–70 (1979).
16. Josh Jaffe, Pankaj Rohatgi, and Marc Witteman. Efficient Side-Channel Testing For Public Key Algorithms: RSA Case Study. In *NIST Non-invasive attack testing workshop* (2011).
17. Paul C. Kocher, Joshua Jaffe, and Benjamin Jun. Differential Power Analysis. In *CRYPTO 1999*, LNCS, vol. 1666, pp. 388–397, Springer, Heidelberg (1999).
18. Houssem Maghrebi, Jean-Luc Danger, Florent Flament, and Sylvain Guilley. Evaluation of Countermeasures Implementation Based on Boolean Masking to Thwart First and Second Order Side-Channel Attacks. In *Signals, Circuits and Systems (SCS)* (2009).
19. Stefan Mangard, Norbert Pramstaller, and Elisabeth Oswald. Successfully Attacking Masked AES Hardware Implementations. In *CHES 2005*, LNCS, vol. 3659, pp. 157–171. Springer, Heidelberg (2005).
20. Amir Moradi, Oliver Mischke, and Thomas Eisenbarth. Correlation-Enhanced Power Analysis Collision Attack. In *CHES 2010*, LNCS, vol. 6225, pp. 125–139, Springer, Heidelberg (2010).
21. Sumio Morioka and Akashi Satoh. An Optimized S-Box Circuit Architecture for Low Power AES Design. In *CHES 2002*, LNCS, vol. 2523, pp. 172–186, Springer, Heidelberg (2002).
22. Liam Paninski. Estimation of Entropy and Mutual Information. *Neural Computation*, 15(6):1191–1253, (2003).
23. Oscar Reparaz, Benedikt Gierlichs, and Ingrid Verbauwhede. Selecting Time Samples for Multivariate DPA Attacks. In *CHES 2012*, LNCS, vol. 7428, pp. 155–174, Springer, Heidelberg (2012).

24. Werner Schindler, Kerstin Lemke, and Christof Paar. A Stochastic Model for Differential Side Channel Cryptanalysis. In *CHES 2005*, LNCS, vol. 3659, pp. 30–46, Springer, Heidelberg (2005).
25. B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London (1986).
26. François-Xavier Standaert, Tal Malkin, and Moti Yung. A Unified Framework for the Analysis of Side-Channel Key Recovery Attacks. In *EUROCRYPT 2009*, LNCS, vol. 5479, pp. 443–461, Springer, Heidelberg (2009).
27. B. L. Welch. The generalization of "Student's" problem when several different population variances are involved. *Biometrika*, 34(1–2):28–35 (1947).
28. Carolyn Whitnall and Elisabeth Oswald. A fair evaluation framework for comparing side-channel distinguishers. *J. Cryptographic Engineering*, 1(2):145–160, (2011).