A Fair Evaluation Framework for Comparing Side-Channel Distinguishers

Carolyn Whitnall and Elisabeth Oswald

University of Bristol, Department of Computer Science,
Merchant Venturers Building, Woodland Road, BS8 1UB, Bristol, UK
{carolyn.whitnall, elisabeth.oswald}@bris.ac.uk

Abstract. The ability to make meaningful comparisons between side-channel distinguishers is important both to attackers seeking an optimal strategy and to designers wishing to secure a device against the strongest possible threat. The usual experimental approach requires the distinguishing vectors to be estimated: outcomes do not fully represent the inherent *theoretic* capabilities of distinguishers and do not provide a basis for conclusive, like-for-like comparisons. This is particularly problematic in the case of mutual information-based side channel analysis (MIA) which is notoriously sensitive to the choice of estimator. We propose an evaluation framework which captures those theoretic characteristics of attack distinguishers having the strongest bearing on an attacker's general ability to estimate with practical success, thus enabling like-for-like comparisons between different distinguishers in various leakage scenarios. We apply our framework to an evaluation of MIA relative to its rather more well-established correlation-based predecessor and a proposed variant inspired by the Kolmogorov-Smirnov distance. Our analysis makes sense of the rift between the *a priori* reasoning in favour of MIA and the disappointing empirical findings of previous comparative studies, and moreover reveals several unprecedented features of the attack distinguishers in terms of their sensitivity to noise. It also explores—to our knowledge, for the first time—theoretic properties of near-generic power models previously proposed (and *experimentally* verified) for use in attacks targeting injective functions.

1 Introduction

A differential side-channel distinguisher is a statistic used to compare hypothesis-dependent predictions with measured side-channel leakage in order to identify the hypothesis most likely to relate to the true internal state of a cryptographic algorithm. For attackers and designers alike, it is desirable to make meaningful comparisons between different such distinguishers in order to optimize an attack or (respectively) to secure a device against the strongest possible threat (see [32]). This is problematic because of the many and varied factors determining attack outcomes (target function, quality of power model, chosen distinguisher, noise), and the subtle and sometimes unexpected ways in which these factors interplay.

In particular, there is a divide between the theoretical capabilities of a distinguisher and its outcome in practical applications where the distinguishing vector must be estimated. Estimation is differentially problematic across distinguishers, so that comparisons made on the basis of experimental results (whether from simulated or measured traces) are inconclusive: do the observed differences arise from inherent theoretical weaknesses/strengths of the distinguishers or would different estimation procedures yield different relative performance?

Such experimental comparisons therefore only enable statements about the relative performance of different *estimators* and not of the distinguishers (i.e. the *estimands*) themselves. Much of the literature (particularly that pertaining to mutual information, which, as a nonparametric statistic, is notoriously problematic to estimate) has focused on improving or varying estimation procedures in order to attain some sort of 'best case' basis for comparison.

1.1 Our Contribution

This paper explains in greater depth, and applies in greater breadth, the alternative approach presented in [35] and further used in [36]. The objective is to abstract away altogether from the confounding problem of estimation by focussing on the *theoretic* properties of distinguishers as employed in various given leakage scenarios. This addresses a shortcoming of the existing literature, which has tended to overlook the fact that by defining and numerically evaluating possible leakage distributions and the corresponding attacker predictions, the estimands for the distinguishing vectors can be directly computed according to their various definitions[1]. Thus we are able to compare the distinguishers themselves on a like-for-like basis which is independent of choices about estimators.

To this end are presented a selection of theoretic outcome measures which, taken together, are *strongly indicative* of the efficiency with which an attack can be implemented practically. This ensures that the conclusions of our comparisons are *relevant*. However, the differential burden and complexity of estimation associated with different distinguishers does still mean that advantages observed in the theoretic realm will not always directly translate into practice.

We subsequently apply our framework to an example question which is a 'hot topic' in the current literature: Mutual Information Analysis (MIA) was proposed in [11] as an 'optimised' and generic enhancement to correlation-based DPA (CPA), but has disappointed in (most) subsequent comparisons (see [3] for a good overview). By rigorously assessing the *theoretic* capabilities of MIA with respect to a range of leakage scenarios we shed new light on the rift between the *a priori* reasoning and the empirical evidence, demonstrating when and in what sense it *does* represent a superior attack methodology.

Our analysis reveals a surprising sensitivity to noise which resembles a type of stochastic resonance and can even be critical in determining the theoretic success or failure of an MIA distinguisher in certain scenarios. This is by contrast with CPA which is only affected by noise at the *practical* level—that is, the sample size required for precise estimation increases but the underlying distinguishing ability of the theoretic vector remains unchanged.

This noise sensitivity proves especially significant in the case of the 'near-generic' attacks described by the authors of [11], who propose to use the 7 least significant bits (7LSB) as a power model against injective target functions. Although experimentally verified in subsequent investigations such as [21], we provide (to our knowledge) the first detailed theoretical level evaluation. We show that near-generic attacks actually require a certain amount of noise in order to achieve theoretic success—failing comprehensively to recover the key in strong-signal settings—and that even when they do succeed they do not supply the hoped-for advantages displayed by generic attacks against non-injective targets.

Alongside the evaluation of MIA and CPA we present analogous results for a Kolmogorov-Smirnov inspired distinguisher which was proposed in [34] to be conceptually similar to MIA but less problematic to implement. The authors verified the effectiveness of the distinguisher in one typical attack scenario; the work of [36] presents a more in-depth analysis which we further extend here—finding it to be a reasonable substitute for MIA in first-order scenarios, with some evidence that it is more noise robust. Its adaptation to second-order scenarios is less successful and it does not share MIA's potential for higher-order adaptations.

1.2 Outline

The remainder of this paper is structured as follows: First, (Sect. 2) we explain (in greater depth) the background to the motivating problem, by describing DPA attacks in general and our three distinguishers in particular, and by discussing the difficulty of making meaningful comparisons when there are so many factors contributing to outcomes. In Sect. 3 we introduce our framework for comparing attacks on a purely theoretic basis and

[1] Whilst some previous work has taken theoretic distinguishing vectors into consideration (e.g. [25]), this approach has not been applied systematically to an analysis of different types of distinguishers.

reason about the extent to which such comparisons are relevant to practical outcomes. Sect. 4 reports on the comparative analysis, within the proposed framework, of MIA, CPA and KSA. We conclude in Sect. 5.

2 Background

2.1 Differential Power Analysis

We consider a 'standard DPA attack' scenario as defined in [17]. Suppose that the power consumption T of the target cryptographic device depends on some internal value (or state) $f_{k^*}(X)$. The state is a function of some part of the plaintext which is a random variable $X \overset{R}{\in} \mathcal{X}$, as well as some part of the secret key $k^* \in \mathcal{K}$. Consequently, we have that $T = L \circ f_{k^*}(X) + \varepsilon$, where L is some function which describes the data-dependent component and ε comprises the remaining power consumption which can be modeled as independent random noise. The attacker has N power measurements corresponding to encryptions of N known plaintexts $x_i \in \mathcal{X}$, $i = 1, \ldots, N$ and wishes to recover the secret key k^*. The attacker can accurately compute the internal values as they would be under each key hypothesis $\{f_k(x_i)\}_{i=1}^N$, $k \in \mathcal{K}$ and uses whatever information he possesses about the true leakage function L to construct a prediction model $M : f(\mathcal{X}) \longrightarrow \mathcal{M}$.

DPA is based on the intuition that the modeled power traces corresponding to the correct key hypothesis should bear more resemblance to the true power traces than the modeled traces corresponding to incorrect key hypotheses. An attacker is thus concerned with comparing the degree of similarity between the true and modeled traces. A range of comparison tools—'distinguishers'—can be used: Pearson's correlation coefficient [6] is a particularly popular choice and found to perform well in many tested scenarios, particularly when M is a good approximation for L. Mutual Information Analysis (MIA) has been proposed as an enhancement to correlation DPA (CPA) which relies less on M [11], and Kolmogorov-Smirnov Analysis (KSA) has been suggested as an alternative enhancement, conceptually similar to MIA but less sensitive to choices made about estimation procedure [34].

We describe these distinguishers in more detail in Sect. 2.3, but let us first consider what it means for a DPA attack to be successful.

2.2 DPA Outcomes

We concentrate on the notion of *key-recovery success* as formalised by Standaert *et al.* in [32]. The theoretic attack distinguisher is $\mathbf{D} = \{D(k)\}_{k \in \mathcal{K}} = \{D(L \circ f_{k^*}(X) + \varepsilon, M \circ f_k(X))\}_{k \in \mathcal{K}}$, where the plaintext input X takes values in $\mathcal{X}$ according to some known distribution (usually uniform). We say the attack is *theoretically successful* if $D(k^*) > D(k)\ \forall k \neq k^*$. We say it is *o-th order theoretically successful* if $\#\{k \in \mathcal{K} : D(k^*) \leq D(k)\} < o$.

However, in practice $\mathbf{D}$ must be estimated. Suppose we have observations corresponding to the vector of inputs $\mathbf{x} = \{x_i\}_{i=1}^N$, and write $\mathbf{e} = \{e_i\}_{i=1}^N$ to be the observed noise (i.e. drawn from the distribution of ε). Then the size $\#\mathcal{K}$ estimated vector is $\hat{\mathbf{D}}_N = \{\hat{D}_N(k)\}_{k \in \mathcal{K}} = \{\hat{D}_N(L \circ f_{k^*}(\mathbf{x}) + \mathbf{e}, M \circ f_k(\mathbf{x}))\}_{k \in \mathcal{K}}$. We then say the attack is *successful* if $\hat{D}_N(k^*) > \hat{D}_N(k)\ \forall k \neq k^*$ and *o-th order successful* if $\#\{k \in \mathcal{K} : \hat{D}_N(k^*) \leq \hat{D}_N(k)\} < o$.

To avoid over-stating the physical security of a device it is important to take into account the most powerful methods available to an attacker with access to side-channel measurements. Attempts to compare different distinguishers in the search for the 'most effective' have thus received considerable attention in the literature (see [31] for a particularly thorough empirical evaluation). Concurrent efforts such as [20,32] have addressed the meta-challenge of developing a formal model for the notion of physical security. General statements about the relative merits of particular methods are extremely hard to come by as attack outcomes are highly scenario-specific. It is important to understand the multiple contributory factors and the way they interact if we want to begin to make meaningful comparisons between side-channel distinguishers.

Factors Contributing to Theoretic Outcomes The target intermediate function f is known to play an important role in determining DPA outcomes; some operations—most notably those which are designed to be cryptographically secure—are particularly vulnerable [12,24]. This is because small changes in the input produce big changes in the output, so that any wrong key hypothesis leads to predictions which are clearly distinguishable from the true consumption. On the other hand, cryptographically weak functions such as AddRoundKey are far more resilient to DPA, as similar keys produce similar predictions and the true key is thus identified by a much smaller margin. The DES algorithm is a popular context for analysis because of its long-established and widespread use, its general susceptibility to DPA (arising from the nonlinearity of its S-Boxes) and its particular amenability to generic attacks (arising from the non-injectivity of its S-Boxes). In our example application we therefore choose the first DES S-Box as a basis for a comprehensive evaluation over a range of leakage assumptions; however, we also investigate the performance of our distinguishers against the AES S-Box as representing an injective cryptographic function and (to demonstrate the contrast) against bitwise exclusive-or (XOR), i.e. DES and AES AddRoundKey.

The characteristics of the device leakage—the functional form of the data-dependent component and the relative size and shape of the independent noise—will substantially dictate how easily and effectively the side-channel can be exploited. In particular, studies such as [8] have clearly demonstrated the central role of the attack power model in determining distinguisher performance. In the case that an attacker has full control over an identical device, profiling (as, for example, in [29]) can produce a very good approximation of the leakage function. However, we restrict our focus to a weaker adversary, with access only to an unverifiable guess based on what is known or assumed about the underlying technology of the device. Therefore the extent to which the device leakage is 'typical' or predictable will have a significant bearing on the attack outcome.

Some devices, such as those built with CMOS technology, are well-known to consume power approximately proportional to the Hamming weight of/Hamming distance between processed values. This has resulted in the widespread popularity of Hamming weight/distance power models in side-channel research, in particular when little is known about the true form of the leakage and the device cannot be profiled (i.e. the scenario to which this study relates). However, not all technology conforms neatly to predictable behaviour. For example, the authors of [28] show that power consumption in emerging nanoscale technologies is not consistent even between identical devices, so that even an attacker with profiling capabilities has difficulties predicting the leakage of a target device with confidence. Non-standard leakage can also be observed in typical hardware implementations of substitution boxes [**?**].

In our example application in Sect. 4 we assume an uninformed attacker who can use either a 'standard' power model, namely the Hamming weight, or, when appropriate to the distinguisher, a 'generic' (for non-injective target functions) or 'near-generic' (for injective target functions) power model, namely the identity or the 7 least significant bits (7LSB) of the hypothesised processed value (both of which were proposed in [11] as suitable for use in MIA). With this attacker in mind we consider three leakage scenarios: The first, and simplest, is an *optimistic* scenario in which the data-dependent leakage really *is* proportional to the Hamming weight as per the adversary's standard model. The second we term *realistic*: as motivated by [1], we suppose that the true leakage is actually an unevenly weighted sum of the bits.[2] Lastly, as an example of a *challenging* (but still realistic) scenario we suppose that the true leakage is a highly nonlinear function of the intermediate data.[3]

It is natural to expect the presence of noise to have an impact on *practical* outcomes: the weaker the signal-to-noise ratio (SNR, defined as $\frac{\mathrm{var}(L \circ f_{k^*}(X))}{\mathrm{var}(\varepsilon)}$), the more data will be required to estimate the distinguishing vector with sufficient precision to detect the true key (see Chap. 4 of [16]). However, less obvious is the fact that the shape of the underlying theoretic vector can also be sensitive to noise; with the exception of correlation DPA, noise impacts differentially by key hypothesis so that it actually plays a role in determining whether or not the correct key is identified (and by what margin). In order to separately consider the roles of the leakage scenario

[2] Specifically, we allow the least significant bit (LSB) to dominate with a relative weight of 10, since the experiments of [34] identified this as sufficient distortion to enable MIA to outperform CPA.

[3] Specifically, we map the target value to the Hamming weight of the AES S-Box output. There is no significance to this choice other than that it is well-known and specially fitted with the nonlinearity properties useful to produce our hypothetical degraded leakage scenario.

(f, L) and of the noise we will initially consider the behaviour of our distinguishers in a pure-signal setting and then go on to show how Gaussian noise of varying size impacts on distinguisher outcomes.

Theoretic Outcomes vs. Practical Outcomes Supposing, then, that the attacker has chosen a distinguisher which is theoretically capable of determining the correct key in a given setting (i.e. a given combination of (f, L, ε)), distinguishing it from the incorrect hypotheses by a margin of a certain size. The *practical* outcome of the attack will ultimately depend on the attacker's ability to estimate the distinguishing vector sufficiently precisely so as to detect a difference of that size. The theory behind *statistical power analysis* [14] tells us that the amount of data needed to do this depends on the effect size and on the sampling distributions of the estimator under the true and rival hypotheses. Since these sampling distributions depend on the true underlying trace distribution (which is unknown), the overlapping tasks of choosing a 'good' estimator and of computing the sample size required by an estimator (a desirable metric in the evaluation of physical security) are usually extremely difficult, at least under reasonable assumptions. (The sample correlation coefficient is a somewhat exceptional case, as we explain in section 2.3).

In short, there is no such thing as a universal 'ideal' estimator for any given distinguisher, by which to fairly measure its best case capabilities in a given leakage scenario. This rather undermines attempts to compare distinguishers on the basis of practical experiments with simulated or measured traces: perceived advantages/ disadvantages are inconclusive as we do not know if they truly indicate inherent strengths/weaknesses of the distinguishers or merely arise from the choice of estimation procedure. A common theme of the research into MIA, for example, has been the attempt to establish whether by improved estimation techniques it can achieve the much-sought-after advantage over CPA [2,34]. By abstracting away from the estimation problem in order to focus on theoretic distinguisher values we are able to make like-for-like comparisons, but with the drawback that our results will not necessarily translate into the practical realm due to the differential burden of estimation incurred by different statistics. However, we are keen to stress that our theoretic outcome measures are *indicative* of practical performance—as we explain in Sect. 3.

2.3 Background to the Distinguishers

We now introduce the correlation coefficient, mutual information, and the Kolmogorov-Smirnov test statistic, and explain how they can be used to construct DPA distinguishers.

Pearson's Correlation Coefficient-Based Distinguisher Pearson's correlation coefficient measures the total linear dependency between two random variables A and B. It is defined as $\rho(A, B) = \frac{\mathrm{cov}(A,B)}{\sqrt{\mathrm{var}(A)}\sqrt{\mathrm{var}(B)}}$. It takes values from -1 to 1 and is zero whenever A and B are independent. However, the converse is not true; namely, A and B may be (non-linearly) dependent with a (linear) correlation of 0.

It is estimated from samples $\{a_i\}_{i=1}^n$, $\{b_i\}_{i=1}^n$ via the sample correlation coefficient: $r(A, B) = \frac{\sum_{i=1}^n (a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^n (a_i - \bar{a})^2}\sqrt{\sum_{i=1}^n (b_i - \bar{b})^2}}$. This is a consistent estimator for $\rho(A, B)$ and, moreover, is asymptotically unbiased and efficient if A and B have a joint Normal distribution. Under the same assumptions, we can even approximate the sampling distribution which, in the context of DPA, leads to 'nice' results such as the number of trace measurements required for attacks to be successful (see Chap. 6.4 of [16]).

Because we are primarily interested in the magnitude (as opposed to the direction) of the relationship between the true and modeled leakage we base our distinguisher on the absolute value of the correlation, comparing measured traces $T = L + \varepsilon$ with the hypothesis-dependent predictions M_k as follows:

$$D_\rho(k) = |\rho(T, M_k)| = \left| \frac{\mathrm{cov}(T, M_k)}{\sqrt{\mathrm{var}(T)}\sqrt{\mathrm{var}(M_k)}} \right| . \tag{1}$$

If the model M adequately approximates the data-dependent leakage L (up to proportionality) then we expect (1) to be maximised for the correct key hypothesis $k = k^*$.

The impact of noise on the distinguisher is straightforward. In fact, as derived in Chap. 6.3 of [16], $\rho(L + \varepsilon, M_k) = \frac{\rho(L,M_k)}{\sqrt{1+\frac{\sigma_\varepsilon^2}{\mathrm{Var}(L)}}}$, where σ_ε is the noise standard deviation. Thus, the larger the noise, the more diminished are the correlations. But—crucially—the denominator does not depend on the key hypothesis; the theoretic distinguisher vector is thus scaled in such a way that the rankings and other *relative* features (such as the standard score and distinguishability measures defined in 3) are preserved. This does not at all imply that *practical* CPA attacks are immune to noise: As the sample variance of the estimator increases, the number of traces required to reach a sufficient level of precision also increases (see Chap. 4 of [16]).

Multivariate Extensions Pearson's correlation coefficient has no natural multivariate extension, but it has been adapted for use in higher-order DPA attacks against masked implementations by introducing a data pre-processing step in which multivariate trace measurements are mapped to a univariate trace which is then compared with the model predictions in the usual way [19]. A second-order CPA distinguisher takes the form:

$$D_{2\mathrm{O}\rho}(k) = |\rho(P(T_1, T_2), M_k)| = \left|\frac{\mathrm{cov}(P(T_1, T_2), M_k)}{\sqrt{\mathrm{var}(P(T_1, T_2))}\sqrt{\mathrm{var}(M_k)}}\right|, \tag{2}$$

where $P : \mathcal{T}_1 \times \mathcal{T}_2 \longmapsto \mathcal{T}_P$ is the pre-processing function. We choose P to be the normalised product: $P(T_1, T_2) = (T_1 - \mathbb{E}(T_1)) \times (T_2 - \mathbb{E}(T_2))$, as representing the best available in the literature (at least in the case of known Hamming weight leakage—see [26]).

Mutual Information-Based Distinguisher The appeal of mutual information (MI) for use in DPA is that, rather than measuring linear dependencies only, it quantifies the *total* information shared between two random variables. It is measured in bits and is most intuitively expressed in terms of entropies via Shannon's formula: $\mathrm{I}(A; B) = \mathrm{H}(A) - \mathrm{H}(A|B)$.[4]

As a functional of probability distributions, MI is notoriously problematic to estimate [5,13,22,30,33]. All estimators are biased, and further no 'ideal' estimator exists—that is to say, different estimators perform differently depending on the underlying structure of the data. The usual approach is to first estimate the underlying marginal and conditional densities and then to substitute these into Shannon's formula via a 'plug-in' estimator for discrete entropy. There are many different ways to estimate densities and the quality of the resulting estimator for MI is very sensitive to the methods and parameters chosen. If we have a good understanding of the underlying distributions we can fit a parametric model such as a Gaussian mixture (see Veyrat-Charvillon et al. [34]). However, since MIA has been proposed for use in scenarios where our usual assumptions do not hold we are generally more interested in nonparametric methods, which are somewhat sensitive to user approach and known to incur an overhead in terms of estimation costs.

Unfortunately, estimators for MI do not behave so 'nicely' as the sample correlation coefficient; in fact, there are no universal rates of convergence [22], so that whatever estimator we pick, we can always find a distribution for which the error vanishes arbitrarily slowly. In the absence of general results about the sampling distribution of the estimators, we cannot compute (for example) the number of traces needed for an attack to be successful, except under the strongest of assumptions[5].

Its application as an attack distinguisher is as follows:

$$D_{\mathrm{MI}}(k) = \mathrm{I}(T; M_k) = \mathrm{H}(T) - \mathrm{H}(T|M_k) = \mathrm{H}(T) - \underset{m\in\mathcal{M}}{\mathbb{E}}\left[\mathrm{H}(T|M_k = m)\right], \tag{3}$$

[4] The original (but equivalent) definition is $\mathrm{I}(A; B) = \sum_{a\in\mathcal{A}} \sum_{b\in\mathcal{B}} p_{A,B}(a, b) \log_2 \left(\frac{p_{A,B}(a,b)}{p_A(a\, p_B(b)}\right)$, where $p_{A,B}$ is the joint probability density of A and B and p_A, p_B are the marginal densities.

[5] Under strong simplifying assumptions, estimating an MIA parametrically can be shown to be equivalent to conducting a correlation attack [17].

and because the 'unexplained' entropy (the second term) is smallest when the predictions are good, we expect (3) to be maximised for the correct key hypothesis $k = k^*$.

Unlike CPA the impact of noise on the MIA distinguishing vector is complex (see, for example, [15]). In particular, whilst $\mathrm{I}(L+\varepsilon; M_k) \leq \mathrm{I}(L; M_k)$ (L, ε independent), nonetheless $\mathrm{I}(L; M_k) - \mathrm{I}(L+\varepsilon; M_k) \neq \mathrm{I}(L; M_{k'}) - \mathrm{I}(L+\varepsilon; M_{k'})$. Hence, the vector elements are differentially affected so that theoretic outcomes in a pure-signal setting do not directly generalise to theoretic outcomes in the presence of noise.

Multivariate Extensions MI generalises quite naturally to higher-order statistics via several different meaningful extensions. The authors of [3] presented three such notions and explored how each could be adapted to the purposes of DPA. They subsequently demonstrated that the three were essentially (theoretically) equivalent in the case of a perfectly implemented masking scheme (i.e. ensuring that $\mathrm{I}(T_1; M_{k^*}) = \mathrm{I}(T_2; M_{k^*}) = 0$). As this is precisely the scenario we consider for our second-order attacks, we only need pursue one formulation. We pick that which most intuitively relates to the DPA problem—namely, the information shared between the *pair* of trace points taken jointly and the model prediction, as follows:

$$D_{\mathrm{2OMI}}(k) = \mathrm{I}((T_1, T_2); M_k) = \mathrm{H}(T_1, T_2) - \mathrm{H}(T_1, T_2 | M_k). \tag{4}$$

Another opportunity presented by multivariate extensions of MI is the possibility of simultaneously targeting multiple functions in a standard *unprotected* scenario, in order to see if the incorporation of more information can help to further optimise an attack. To test this idea we compute the MI between the pair of trace values and a corresponding pair of predictions:

$$D_{\mathrm{MMI}}(k) = \mathrm{I}((T_1, T_2); (M_1, M_2)_k) = \mathrm{H}(T_1, T_2) - \mathrm{H}(T_1, T_2 | (M_1, M_2)_k). \tag{5}$$

This formulation relates intuitively to the problem we are trying to address. Whilst alternative quantities such as $\mathrm{I}(T_1; T_2; M_{1,k}; M_{2,k})$ are meaningful in the context of information theory generally, it is not clear that such an approach would be meaningful in the specific context of DPA. For completeness, we did test this 4-variate notion, confirming that it is far less effective than the distinguisher in (5); we omit the detailed results as not adding particular insight beyond this summary observation.

Kolmogorov-Smirnov-Based Distinguisher The Kolmogorov-Smirnov (KS) distance between the distributions of random variables A and B is defined as $K(A||B) = \sup_{x \in \mathcal{A} \cup \mathcal{B}} |F_A(x) - F_B(x)|$ where F_A, F_B are the cumulative distribution functions (CDFs) of A and B, i.e. $F_A(x) = \mathbb{P}(A \leq x)$. In a two-sample KS test designed to test the null hypothesis that A and B share the same distribution, the empirical CDFs are estimated from samples $\{a_i\}_{i=1}^n$, $\{b_i\}_{i=1}^n$, e.g. $\hat{F}_A(x) = \frac{1}{n} \sum_{i=1}^n I_{\{a_i \leq x\}}$ ($I_{\{a_i \leq x\}}$ is the indicator function, taking the value 1 if $a_i \leq x$ and 0 otherwise). The fact that the KS test statistic does not require explicit density estimation is what makes it appealing as an alternative to MI.

Just as MIA can be understood to operate by comparing the global traces T with the hypothesis-dependent conditional traces $T|M_k$—via the expected change in entropy—a KS-inspired distinguisher measures the maximum distance between the global and the conditional trace distributions, as averaged over the prediction space:

$$D_{\mathrm{KS}}(k) = \mathbb{E}[K(T||T|M_k)] = \mathop{\mathbb{E}}_{m \in \mathcal{M}} \left[\sup_y |F_T(y) - F_{T|M_k=m}(y)| \right]. \tag{6}$$

In case of the correct key hypothesis we expect the test statistic to return a large difference.

Multivariate Extensions Multivariate extensions of the KS test are somewhat more difficult to achieve, as we first need to formulate an appropriate notion of a multivariate CDF. In the one-dimensional case there are only two ways of ordering the data, namely $\mathbb{P}(A \geq x)$ and $\mathbb{P}(A \leq x)$. As we have that $\mathbb{P}(A \geq x) = 1 - \mathbb{P}(A \leq x)$ the choice turns out to be arbitrary.

In higher dimensions the choice of ordering is no longer inconsequential: there is no direct way to map (e.g.) between $\mathbb{P}(A_1 \leq x, A_2 \leq y)$ and $\mathbb{P}(A_1 \geq x, A_2 \leq y)$. In fact for d different random variables, there are 2^d possible orderings we need to consider. Peacock (in [23]) proposes to define the KS distance as the maximum distributional difference taken over *all* the orderings, so that (for example) in the bivariate case:

$$K(A_1, A_2||B_1, B_2) = \max \left\{ \sup_{(x,y)\in(\mathcal{A}_1\cup\mathcal{B}_1)\times(\mathcal{A}_2\cup\mathcal{B}_2)} \left| F^{(i)}_{A_1,A_2}(x, y) - F^{(i)}_{B_1,B_2}(x, y)\right| \right\}_{i=1}^{4},$$

where $\mathcal{S} = (\mathcal{A}_1 \cup \mathcal{B}_1) \times (\mathcal{A}_2 \cup \mathcal{B}_2)$ and $F^{(1)}$ to $F^{(4)}$ are the CDFs based on all four possible orderings. He shows that a bivariate KS test statistic according to this approach is close enough to being distribution-free to be useful in practice. Inevitably, this impacts not just the data complexity of estimation but also the computational costs; in spite of various optimisations proposed in [9], extensions to dimensions greater than 2 quickly become infeasible.

For second-order attacks against a masked implementation we adapt the KSA distinguisher to compare the global joint CDF of the traces with the joint CDFs as partitioned by the model predictions under each key hypothesis (this is conceptually comparable to our second-order MIA distinguisher):

$$D_{\text{2OKS}}(k) = \mathbb{E}[K(T_1, T_2||T_1, T_2|M_k)] = \mathop{\mathbb{E}}_{m\in\mathcal{M}} \left[\max \left\{ \sup_{y_1,y_2} \left| F^{(i)}_{T_1,T_2}(y_1, y_2) - F^{(i)}_{T_1,T_2|M_k=m}(y_1, y_2)\right| \right\}_{i=1}^{4} \right]. \quad (7)$$

Similarly, following the construction of our two-target MIA distinguisher we consider a KSA analogue whereby the global joint CDFs are compared with the joint CDFs as conditioned on the bivariate predictions:

$$\begin{aligned} D_{\text{MKS}}(k) &= \mathbb{E}[K(T_1, T_2||T_1, T_2|(M_1, M_2)_k)] \\ &= \mathop{\mathbb{E}}_{(m_1,m_2)\in\mathcal{M}_1\times\mathcal{M}_2} \left[\max \left\{ \sup_{y_1,y_2} \left| F^{(i)}_{T_1,T_2}(y_1, y_2) - F^{(i)}_{T_1,T_2|(M_1,M_2)_k=(m_1,m_2)}(y_1, y_2)\right| \right\}_{i=1}^{4} \right]. \end{aligned} \quad (8)$$

3 Comparison Framework

As mentioned above, a useful measure of physical security would be the number of traces needed for an attack to be successful. We can compute this for a given estimator using the techniques of *statistical power analysis* [14], provided the sampling distribution can be approximated—but this is not achievable in general (see Sect. 2.3), besides which we are seeking to avoid estimator-specific comparisons. Our solution, as first introduced in [35], is to choose measures based on those characteristics of the theoretic vectors which have the greatest bearing on the trace efficiency of a practical attack. The first needs little explanation:

1. *Correct key ranking*: The position of the correct key when ranked by distinguisher value. If the correct key is ranked joint first the *ranking order* is the number of keys sharing position 1, so that an attack with a ranking order of o is o^{th}-order theoretically successful as defined in Sect. 2.2. The relationship with practical efficiency is obvious: attacks which are not first-order successful will not be able to uniquely extract the correct key from *any* number of trace measurements (except by random chance).

$$\textit{Rank}(\mathbf{D}) = 1 + \sum_{k\in\mathcal{K}} I_{\{D(k)>D(k^*)\}};$$

$$\textit{RankOrder}(\mathbf{D}) = \sum_{k\in\mathcal{K}} I_{\{D(k)=D(k^*)\}}.$$

The theory behind statistical power analysis tells us that, when estimating population quantities, the sample size required to detect a statistically significant difference increases as the actual magnitude of the true difference decreases. Therefore, of crucial practical relevance to *practical* attack outcomes are the *theoretical* margins by which the true key is isolated from the remaining keys. Such is the motivation for the next three measures:

2. *Relative distinguishing margin*: The distance between the correct key distinguisher value and the value for the highest ranked alternative, normalised by the standard deviation of the distinguishing vector so that scale-free comparisons can be made between different distinguishers in different leakage scenarios. (Note that it is zero for attacks with success orders greater than 1, and negative for failed attacks, where it gives further indication of the extent of the failure).

$$\mathit{RelMarg}(\mathbf{D}) = \frac{D(k^*) - \max\{D(k)|k \neq k^*\}}{\sqrt{\mathit{Var}\{D(k)|k \in \mathcal{K}\}}}.$$

3. *Absolute distinguishing margin*: The relative margin allows us to summarise the *shape* of a distinguishing vector and how this responds to noise or scenario degradation. However, it disguises changes in the *actual magnitude* of the margin and the fact that this is more sensitive for some methodologies than for others. We need some way to take into account raw margin size as well as size relative to the vector as a whole, which is still scale-independent so that we can make like-for-like comparisons between distinguishers. We therefore report the ratio between the nearest-rival margin and that of the corresponding 'optimal' vector: the univariate equivalent in an optimistic (i.e. known Hamming weight power model) noise-free setting. This will allow us to comment meaningfully on the impact of model degradation and noise on the real size of the margins to be estimated.

$$\mathit{AbsMarg}(\mathbf{D}) = \frac{D(k^*) - \max\{D(k)|k \neq k^*\}}{D(L \circ f_{k^*}(X), L \circ f_{k^*}(X)) - \max\{D(L \circ f_{k^*}(X), L \circ f_{k}(X))|k \neq k^*\}}.$$

4. *Standard score*: This is the same as the "DPA signal-to-noise ratio" described by [12]: the number of standard deviations above (or below) the mean, for the correct key distinguisher value. It provides a more general measure of the sensitivity of an attack in isolating the correct key. A theoretically 'unsuccessful' attack may still be able to return a small candidate subset containing the correct key if the standard score is high.

$$\mathit{StdScore}(\mathbf{D}) = \frac{D(k^*) - \mathbb{E}\{D(k)|k \in \mathcal{K}\}}{\sqrt{\mathit{Var}\{D(k)|k \in \mathcal{K}\}}}.$$

By computing the above measures for uniformly drawn plaintexts $X \overset{unif.}{\leftarrow} \mathcal{X}$, we are able to compare theoretic behaviour of attacks when provided with full information. We propose to explore the sensitivity of attacks to incomplete information by inspecting theoretic attack vectors as restricted on reduced subsets of the plaintext space: $\mathbf{D}|_{\mathcal{X}'}$ where $\mathcal{X}' \subseteq \mathcal{X}$. These vectors depend not only on the size but also on the composition of the input set $\mathcal{X}'$; we cannot perform the computations exhaustively over the entire space of possible subsets (it is too large), but by repeated random draws of increasing size we can estimate the support size needed for theoretic success. We argue that this provides insight into the relative data complexity of distinguishers and their particular limitations in small samples. We thus add the following measures (defined for theoretically successful distinguishers only):

5. *Average critical support*: On average, the required support size of the input distribution for the attack to achieve o^{th}-order success (where o is the ranking order).

$$\mathit{AveSupp}(\mathbf{D}) = \hat{\mathbb{E}}[\min\{\#\mathcal{X}'|\mathcal{X}' \subseteq \mathcal{X} \wedge \mathit{Rank}(\mathbf{D}|_{\mathcal{X}'}) = 1\}].$$

6. *Critical support for* $100 \times p\%$ *success rate*: The support size for which the rate of success (of the appropriate order) is at least $100 \times p$ per cent.

$$\mathit{PctSupp}(\mathbf{D}, p) = \min\{\#\mathcal{X}'|\mathcal{X}' \subseteq \mathcal{X} \wedge \hat{\mathbb{E}}[I_{\{\mathit{Rank}(\mathbf{D}|_{\mathcal{X}'})=1\}}] = p\}.$$

Our criteria are best viewed in conjunction with one another rather than in isolation, and trade-offs between them will interplay differently with practical considerations. For instance, a methodology which achieves only o^{th}-order success (where $o > 1$) might be preferable to one achieving 1^{st}-order success if the distinguisher vector can be estimated more precisely and/or efficiently. Likewise, nearest-rival distinguishability may be more important than average critical support in the presence of high noise.

Computing the Theoretic Vectors For each possible input $x \in \mathcal{X}$ to the cryptographic function we obtain a vector evaluating the Gaussian density centred at the corresponding data-dependent leakage value $L \circ f_{k^*}(x)$ and having variance $\mathrm{Var}(\varepsilon)$. The average of these vectors, weighted by the input probabilities $\mathbb{P}(X = x)$, then gives the probability density of the power consumption evaluated over the full range of possible leakage values. Conditional densities, corresponding to each possible prediction value $m \in \mathcal{M}$ under each key hypothesis $k \in \mathcal{K}$, are constructed similarly. From these probability densities we are able to directly compute (via numerical integration) the moments, entropies and cumulative probabilities comprising the formulae for our distinguishers (as detailed in Sect. 2.3).

4 Example Application

In this section we apply our proposed framework to an evaluation of MIA relative to its (less optimal?) precursor, CPA, and its (more user-friendly?) relative, KSA.

4.1 Evaluation Scenarios

Recall that we consider a non-profiling adversary who is limited to 'standard', 'generic' and 'near-generic' power models (i.e., the Hamming weight, identity and 7LSB functions, respectively). We evaluate the theoretic capabilities of such an adversary—using CPA, MIA and KSA distinguishers—in optimistic, realistic and challenging leakage scenarios, as defined in Sect. 2.2.

To this end we examine the theoretical vectors underlying DPA attacks against the DES and AES block ciphers. Within each scenario we will first consider the behaviour of our distinguishers in a noise-free setting—to demonstrate the role of data-dependent leakage in determining attack outcomes—and then go on to show how Gaussian noise of varying size impacts on the distinguisher outcomes.

4.2 Univariate Attacks Targeting the First DES S-Box

The Noise-Free Setting The first two blocks of Table 1 report the outcomes of standard and generic univariate attacks on an unprotected DES S-Box with noise-free data-dependent leakage. In the optimistic scenario, the MIA distinguishers exhibit substantially larger relative margins than standard CPA, confirming that in some sense MIA *does* meet the *a priori* expectation of enhanced data exploitation. However, it also requires a larger support to be successful, and it is this initial 'information overhead', combined with the relative efficiency of estimating the correlation coefficient, which accounts for the consistently reported CPA advantage in practical attacks with a good power model. Unsurprisingly, when the standard Hamming weight power model is a good fit to the true leakage, generic MIA offers no advantage, exhibiting a substantially reduced margin in absolute terms and requiring a larger input support to succeed.

As the true leakage diverges from the standard power model, the advantage to MIA increases. In the challenging scenario, CPA actually fails whilst MIA continues to identify the correct key. Moreover, the generic capabilities of the latter become apparent as the distinguishing margins and the critical support size are remarkably robust to the deterioration of the leakage. Hence it appears that the ability of generic attacks to recover the key is in some sense independent of the true leakage: whilst it is always preferable to use a power model when a good one is available, a generic model will work just as well however typical or unusual an unknown leakage function really is.

Generic KSA performs very similarly to generic MIA in each scenario, with slightly diminished relative margins. Standard Hamming weight KSA performs similarly to its MIA counterpart in the optimistic scenario but is less robust to model degradation.

The Impact of Noise on Distinguishing Margins Figure 1 shows the impact of noise on distinguishing margins. As we know already (from Sect. 2.3), the CPA vector is merely scaled by a constant as the SNR varies, so that the relative distinguishing margin is unchanged. By contrast, the relative margins for MIA *are* affected by noise, and in such a way that the relationships are not monotonic. In each leakage scenario there seems to be an optimal SNR at which the margin reaches a maximum, subsequently diminishing to that of the noise-free setting. Such a phenomenon is a type of *stochastic resonance* [4], which can (in principle) occur in any nonlinear measurement system. The impact on KSA margins is less marked.

In the optimistic scenario, standard MIA exhibits the largest relative margins across the tested noise range (in particular maintaining an advantage over generic MIA). Generic KSA gains an advantage over its standard power model counterpart in the presence of sufficient noise, but the margins of each reduce to below those of CPA when the SNR is less than around 0.5. In the realistic scenario the impact of noise is more marked, and with greater implications for the relative effectiveness of the distinguishers. For one, it can now be seen that the advantages exhibited by the generic attacks are actually far more substantial in low-signal settings, so that they may well prove more practically efficient than their standard counterparts. Note also that convergence to the noise-free setting occurs (for all distinguishers) at a larger SNR threshold, hence the different x-axes. In the challenging scenario the generic attacks remain clearly favourable throughout the tested range; in fact the standard MIA and KSA attacks are actually rendered *unsuccessful* by high levels of noise, only achieving key recovery once the signal begins to dominate in the leakage.

The lower part of the figure shows *absolute* margins as the SNR varies. These are most robust for CPA, in such cases that the attack *is* theoretically successful (i.e. the optimistic and realistic scenarios). Since the actual size of the margins to be estimated has a bearing on the amount of data needed for estimation (in addition to the size relative to the variation in the vector), this is likely only to enhance its proven advantage in *practical* attacks in the presence of noise. It is interesting to note that KSA absolute margins are more robust to noise than those of MIA, so that the former method may actually prove the preferable of the two in (noisy) practical settings. This is particularly relevant, for example, in the challenging scenario where the generic MIA and KSA attacks are the only two which remain successful across the tested range.

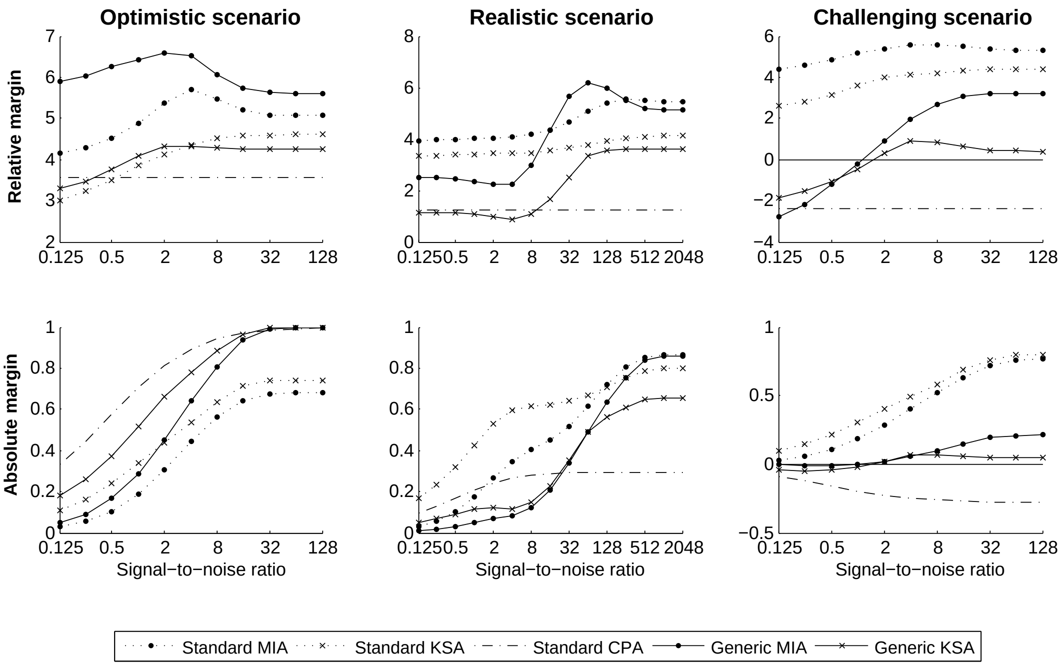

Fig. 1: *Theoretic relative and absolute distinguishing margins as SNR varies, for standard and generic univariate attacks against the first DES S-Box.*

The Impact of Noise on Critical Support Size Within each scenario, we tested the strongest MIA and KSA variants (standard in the optimistic scenario, generic in the realistic and challenging scenarios) to see whether or not noise had any detrimental effect on the support size required for key recovery. We found that it did not—i.e. the outcome measures relating to support size remained constant across the tested SNR range. (For CPA we do not need to test this because of the noise-invariance of the shape of the distinguishing vector). Thus the advantages of MIA and KSA in terms of distinguishing margin size and (in the generic case) scenario and noise robustness are not undermined by any increased support size costs as noise varies.

4.3 Multi-Target Attacks Against Unprotected DES

The Noise-Free Setting To put our multi-target attacks in context we first consider standard Hamming weight attacks against DES AddRoundKey. The outcomes in the noise-free setting (block 3 of Table 1) nicely illustrate the reduced effectiveness of DPA against linear target functions. Even in the optimistic scenario the distinguishers are no longer able to uniquely identify the correct key, instead ranking it equally with its bitwise complement $\bar{k}^*$. The standard scores are reduced relative to those reported for the S-Box attacks with the standard power model; MIA and KSA again seem to have an advantage over CPA but are no longer robust to severe model degradation. Note that, since AddRoundKey is injective, there is no opportunity for the generic strategy.

Block 4 relates to multivariate extensions of MIA and KSA by which we attempt to enhance outcomes by jointly targeting AddRoundKey and the first S-Box. Although we are exploiting a larger amount of information, this increase applies across the range of key hypotheses so that distinguishing margins are not automatically increased. In fact, even in the optimistic scenario the true key is less strongly distinguished (in relative and even absolute terms) than in the attacks against the S-Box alone, and with a substantial increase in the support size required.

The attacks moreover appear less robust to model degradation, and, as with the AddRoundKey attacks, they actually fail in the challenging scenario.

The Impact of Noise on Distinguishing Margins Figure 2 confirms that, in each of our three leakage scenarios, the advantage of the univariate attacks over the multivariate persists across the tested noise range, thus reinforcing our conclusion that more information does not necessarily imply greater attack effectiveness.

This discovery appears counter-intuitive alongside what is known about other classes of multivariate attack. Template attacks, in particular, are known to be *enhanced* by the incorporation of multiple data points (see [7,27]). Careful investigation into the subtle conceptual differences between the two approaches is clearly needed to explain this apparent incongruity.

4.4 Second-Order Attacks Against the First DES S-Box With Masking

The Noise-Free Setting We next consider an implementation in which the outputs of the first DES S-Box are XOR-ed with a uniformly distributed random mask of the same length as the S-Box outputs (i.e. 4 bits).

In the optimistic scenario the relative distinguishing margins for all three second-order attacks do not appear greatly reduced from those of their first-order counterparts; the penalty in absolute terms is more evident. Second-order MIA is the most robust to the masking with an absolute margin of 57% compared with 45% for second-order CPA and just 14% for second-order KSA. The critical support sizes for MIA and CPA are comparable and modest compared with the substantially inflated requirements for KSA (note that the full support set comprises all the possible input-mask pairs—$64 \times 16 = 1024$ in total).

Although MIA is rivaled by CPA in the optimistic scenario, and that the latter continues to succeed even in the challenging scenario (as does second-order KSA), it is decidedly the most robust to degradation of

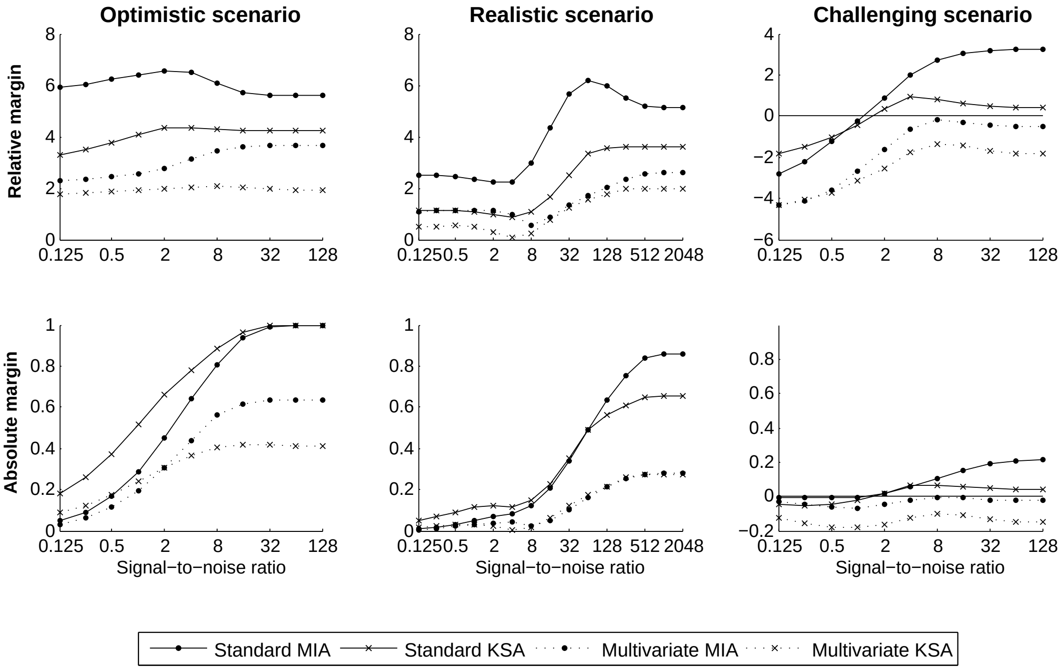

Fig. 2: *Theoretic relative and absolute distinguishing margins as SNR varies, for multivariate attacks against AddRoundKey and the first S-Box in an unprotected implementation of DES.*

the model. On that basis it appears that (theoretically) MIA could be the most well suited of the tested distinguishers to exploiting masked leakage in unusual leakage scenarios. Its advantage over CPA—which relies on pre-processing—has been anticipated in the literature (in particular, [10]) but it is interesting to note that second-order KSA no longer achieves close performance in spite of its conceptual similarity and consistently similar behaviour in first-order scenarios.

The Impact of Noise on Distinguishing Margins Figure 3 indicates that the impact of noise on second-order attacks can be quite profound. In the optimisticscenario the relative margin of second-order MIA is enhanced by noise, whilst in the realistic scenario noise works to the substantial detriment of MIA until the SNR is at least 8. In the challenging scenario noise appears to enhance the relative margins of both MIA and KSA.

In absolute terms the MIA margins are substantially reduced by noise, requiring a strong signal before they begin to approach their pure-signal capabilities. CPA and KSA appear more robust, so that in spite of their inferiority in noise-free settings they may yet be preferable in certain noisy settings.

4.5 Univariate Attacks Targeting the AES S-Box

The Noise-Free Setting We next consider attacks against the AES algorithm. Block 6 of Table 1 reports outcomes of standard attacks against the AES S-Box with noise-free data-dependent leakage. As was the case for DES, the MIA attacks with the Hamming weight power model have substantially larger relative margins than CPA, with KSA falling only slightly behind. However, all three standard attacks are less robust to severe model degradation and fail in the challenging scenario.

Block 7 summarises the theoretic outcomes of 'near-generic' attacks in the absence of noise; it is immediately apparent that these are *not* comparable to the generic variants as used against DES. In fact, MIA with a 7LSB power model fails dramatically, with the correct key appearing last in the ranked vector (by a substantial

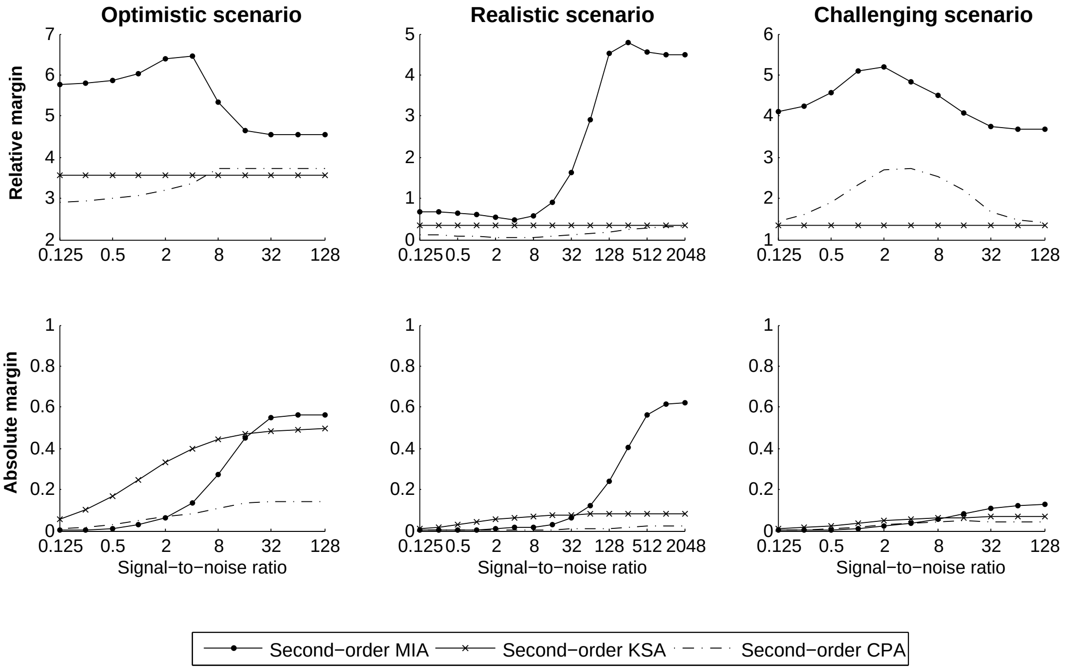

Fig. 3: *Theoretic relative and absolute distinguishing margins as SNR varies, for second-order attacks against the first DES S-Box in a masked implementation.*

margin). The equivalent KSA attack succeeds, but with greatly reduced distinguishing power and requiring almost half of the total input support space.

This is surprising—and even concerning—given the reported effectiveness of the attack in practical scenarios (for example in the experiments of [21]). For clues to this apparent incongruity we must look beyond the noise-free setting.

The Impact of Noise on Distinguishing Margins Examining the first panel of Figure 4 it becomes clear that it is *only* in the strong-signal setting (SNR > 10, approx.) that the attack fails to distinguish the correct key. As the data-dependent signal weakens, the attack *becomes* theoretically distinguishing, though with consistently smaller relative and absolute margins than CPA.[6]

The relative margins of the distinguishers using the standard Hamming weight power model are less affected by noise than in the attacks against DES. In absolute terms (see top right panel of Figure 4) KSA once more appears more robust to noise than MIA, with CPA the most robust of all.

[6] A closer inspection of Shannon's formula in the noise-free setting ($\mathrm{I}(L; M_k) = \mathrm{H}(L) - \mathrm{H}(L|M_k)$) reveals the reason for the eventual failure: *except for the correct key hypothesis*, certain model predictions induce conditional distributions which are supported on a single point and therefore contribute zero entropy to the overall (expected) conditional entropy component $\mathrm{H}(L|M_k)$. All other conditional distributions, including *all* those induced under the correct key hypothesis, are supported on two values and thus contribute entropy of one bit. So the expected conditional entropy is 1 for the correct key and less than 1 everywhere else; since the global entropy $\mathrm{H}(L)$ does not change, $\mathrm{I}(L; M_k)$ is minimal for the correct key.

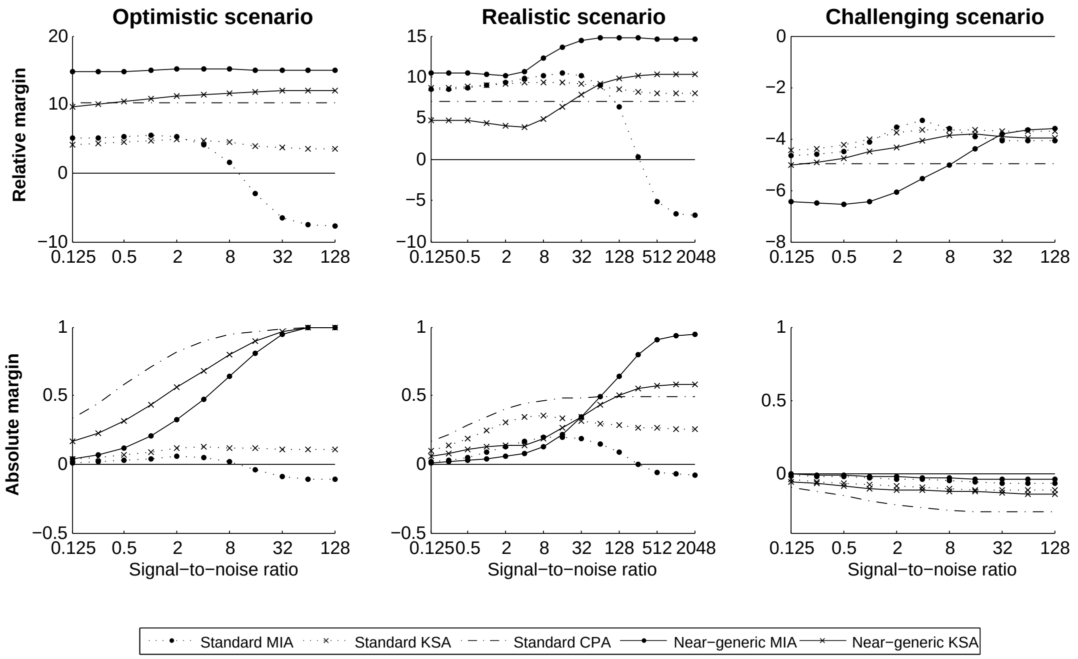

Fig. 4: *Theoretic relative and absolute distinguishing margins as SNR varies, for standard and generic univariate attacks against the AES S-Box.*

4.6 Multi-Target Attacks Against Unprotected AES

The Noise-Free Setting Block 8 of Table 1 shows theoretic outcomes against AES AddRoundKey in the absence of noise. In the optimistic and realistic scenarios we observe the same reduction in effectiveness relative to S-Box attacks that we reported for DES; in the challenging scenario all three attacks fail.

However, block 9—reporting outcomes for the multivariate attacks against AddRoundKey and the S-Box jointly—reveals something interesting. Whilst in the optimistic and realistic scenarios these exhibit smaller margins and greater support requirements than univariate S-Box attacks, in the challenging scenario we actually observe successful multivariate attacks where the separate univariate attacks failed.

The Impact of Noise on Distinguishing Margins Figure 5 confirms that the diminished distinguishing margins exhibited by multivariate MIA and KSA persist across all tested noise settings in the optimistic and realistic scenarios. In the challenging scenario the margins of (unanticipated) success demonstrate good robustness to noise. Thus we arrive at further evidence that attacks against multiple targets *can* sometimes be useful in special circumstances where the true leakage is sufficiently non-standard. But—given the failure of the corresponding attack against DES—this clearly depends also on the target functions chosen and it is hard to make any general concrete statements.

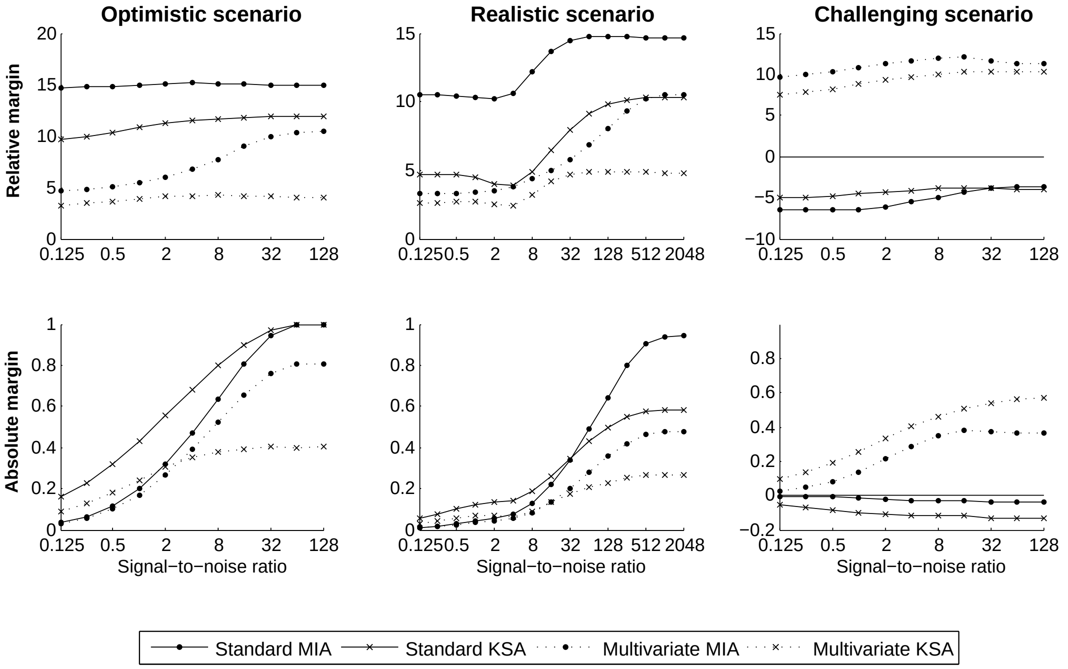

Fig. 5: *Theoretic relative and absolute distinguishing margins as SNR varies, for multivariate attacks against AddRoundKey and the S-Box in an unprotected implementation of AES.*

Table 1: *Theoretic outcomes in optimistic, realistic and challenging scenarios with noise-free data-dependent leakage.*

	Optimistic			**Realistic**			**Challenging**		
	CPA	**MIA**	**KSA**	**CPA**	**MIA**	**KSA**	**CPA**	**MIA**	**KSA**
1. Standard attacks against DES S-Box									
Correct key ranking (order)	1	1	1	1	1	1	12	1	1
Standard score	5.14	6.59	5.95	3.21	6.38	5.49	0.74	5.23	2.66
Relative margin	3.56	5.61	4.24	1.22	5.12	3.61	-2.38	3.22	0.40
Absolute margin	1.00	1.00	1.00	0.30	0.86	0.66	-0.28	0.21	0.04
Average critical support	6	8	8	17	10	12	–	26	39
Critical support for 90% SR	8	11	11	32	14	20	–	37	61
Critical support for 100% SR	16	19	19	49	21	34	–	46	64
2. Generic attacks against DES S-Box									
Correct key ranking (order)	1	1	1	8	1	1	64	1	1
Standard score	5.39	6.35	6.20	1.45	6.66	5.77	-1.29	6.48	5.94
Relative margin	3.61	5.08	4.60	-0.81	5.45	4.12	-3.95	5.30	4.41
Absolute margin	0.85	0.68	0.74	-0.14	0.86	0.80	-0.55	0.77	0.80
Average critical support	9	16	16	–	15	15	–	15	15
Critical support for 90% SR	14	19	19	–	17	17	–	18	18
Critical support for 100% SR	27	24	24	–	21	21	–	25	25
3. Standard attacks against DES AddRoundKey									
Correct key ranking (order)	1 (2)	1 (2)	1 (2)	1 (2)	1 (4)	1 (2)	27	51	9
Standard score	2.62	4.48	3.76	2.24	3.44	4.14	0.37	-0.80	0.93
Relative margin	0.00	0.00	0.00	0.00	0.00	0.00	-1.85	-3.33	-1.41
Absolute margin	–	–	–	–	–	–	–	–	–
Average critical support	6	9	9	26	10	14	–	–	–
Critical support for 90% SR	8	12	12	47	15	24	–	–	–
Critical support for 100% SR	14	21	21	58	23	51	–	–	–
4. Multivariate attacks against DES									
Correct key ranking (order)	–	1	1	–	1	1	–	4	8
Standard score	–	6.04	4.70	–	5.08	4.51	–	1.60	1.04
Relative margin	–	3.66	1.93	–	2.62	1.91	–	-0.53	-2.12
Absolute margin	–	0.64	0.41	–	0.28	0.27	–	-0.02	-0.17
Average critical support	–	14	14	–	28	23	–	–	–
Critical support for 90% SR	–	16	16	–	39	33	–	–	–
Critical support for 100% SR	–	25	25	–	53	47	–	–	–
5. Second-order attacks against DES S-Box									
Correct key ranking (order)	1	1	1	1	1	1	1	1	1
Standard score	5.14	6.16	5.39	2.46	5.90	2.31	3.57	5.23	3.19
Relative margin	3.56	4.55	3.72	0.35	4.48	0.30	1.34	3.67	1.42
Absolute margin	0.45	0.57	0.14	0.07	0.62	0.02	0.08	0.13	0.04
Average critical support	28	29	115	75	44	195	207	140	292
Critical support for 90% SR	50	46	219	189	72	449	503	247	574
Critical support for 100% SR	137	74	385	894	117	953	913	434	909
6. Standard attacks against AES S-Box									
Correct key ranking (order)	1	1	1	1	1	1	186	193	214
Standard score	12.24	15.60	13.91	9.61	15.49	12.69	-0.77	-0.75	-1.03
Relative margin	10.14	14.88	11.95	7.00	14.61	10.30	-4.98	-3.62	-3.98
Absolute margin	1.00	1.00	1.00	0.49	0.94	0.58	-0.26	-0.04	-0.13
Average critical support	5	9	9	22	12	18	–	–	–
Critical support for 90% SR	6	11	11	39	16	30	–	–	–
Critical support for 100% SR	9	15	15	71	23	47	–	–	–
7. Near-generic attacks against AES S-Box									
Correct key ranking (order)	1	256	1	1	256	1	153	212	164
Standard score	11.23	-5.75	6.12	3.86	-3.98	10.23	-0.44	-1.08	-0.42
Relative margin	8.88	-7.74	3.50	0.70	-6.79	7.96	-3.87	-4.06	-3.72
Absolute margin	0.57	-0.11	0.11	0.03	-0.08	0.26	-0.19	-0.06	-0.11
Average critical support	21	–	125	153	–	76	–	–	–
Critical support for 90% SR	35	–	155	234	–	89	–	–	–
Critical support for 100% SR	73	–	186	255	–	110	–	–	–
8. Standard attacks against AES AddRoundKey									
Correct key ranking (order)	1 (2)	1 (2)	1 (2)	1 (2)	1 (4)	1 (2)	193	185	187
Standard score	3.24	7.06	5.18	2.65	5.68	5.50	-0.91	-0.51	-0.74
Relative margin	0.00	0.00	0.00	0.00	0.00	0.00	-3.34	-2.93	-3.83
Absolute margin	–	–	–	–	–	–	–	–	–
Average critical support	7	11	11	59	14	23	–	–	–
Critical support for 90% SR	9	14	14	117	20	41	–	–	–
Critical support for 100% SR	15	25	25	193	40	96	–	–	–
9. Multivariate attacks against AES									
Correct key ranking (order)	–	1	1	–	1	1	–	1	1
Standard score	–	13.81	9.69	–	13.44	9.59	–	13.24	12.65
Relative margin	–	10.44	4.05	–	10.49	4.87	–	11.31	10.38
Absolute margin	–	0.81	0.40	–	0.48	0.27	–	0.36	0.57
Average critical support	–	19	19	–	45	46	–	57	35
Critical support for 90% SR	–	22	22	–	62	68	–	80	47
Critical support for 100% SR	–	28	28	–	86	97	–	105	67

5 Conclusion

We have discussed the many interacting factors contributing to the outcomes of a differential side-channel attack, and explained how these make it difficult to draw meaningful, like-for-like comparisons between different methodologies and/or different leakage scenarios. We proposed a solution to this problem whereby comparisons are made based on the theoretic performance of distinguishers in well-defined hypothetical settings, abstracting away from the confounding problem of practical estimation.

Our methodology entails summarising the key features of a distinguishing vector which contribute to an attacker's ability to estimate it with sufficient precision to achieve practical success. This is something which has not hitherto been explored in the side-channel literature: the difference between estimating attack outcomes from simulated traces and numerically approximating the theoretical quantities directly from the density functions has largely been overlooked.

We subsequently showed how our framework could be applied to the current 'hot topic' question of whether and in what sense MIA distinguishers ever have an advantage over CPA distinguishers, and extended this to include a comparison with a conceptually similar but implementationally 'simpler' distinguisher based on the Kolmogorov-Smirnov distance.

We found that MIA has theoretic advantages (in some senses) even in scenarios which are particularly favourable to CPA (i.e. when the attacker has a good power model), so confirming that the underperformance frequently observed in practical experiments can be largely attributed to estimation overheads. It gains in superiority as the true leakage diverges from the attacker's power model, especially when the 'generic' (power model-free) approach can be used, as when targeting non-injective functions (such as the first DES S-Box). It can therefore be seen as a practically useful alternative to CPA in unusual leakage scenarios.

However, the 'near-generic' approach using the 7LSB power model does not, as hoped, supply an equivalent functionality against injective targets (such as the AES S-Box)—rather it produces some very unexpected results and actually fails quite catastrophically in strong-signal settings. Whether or not the generic capabilities of MIA *can* be exploited against injective targets remains an open question.

The noise dependency of theoretic outcomes is itself a very new result: whilst it has always been expected that the presence of noise affects an attack at the *practical* stage—i.e. the precision with which the distinguishing vector can be estimated—it has not, to our knowledge, been hitherto observed that the underlying ability of a distinguisher to recover the key can itself vary, and to a substantial degree. CPA distinguishers inherently do *not* possess this property, which accounts for the fact that it has not been previously investigated. KSA distinguishers, whilst consistently inferior to MIA in noise-free settings, do exhibit a similar adaptability to non-standard leakage and moreover appear to be more robust to increasing noise so that they may become practically useful alternatives to CPA and MIA when the side-channel leakage is both unusual *and* noisy.

6 Acknowledgements

The authors would like to thank Luke Mather for his contribution relating to the use of the KS test statistic (see [18,36]), and Dave Cliff for pointing them towards the concept of stochastic resonance. The first author of this paper has been funded via an Engineering and Physical Sciences Research Council studentship. The second author has been supported by an Engineering and Physical Sciences Research Council Leadership Fellowship I005226.

References

1. Akkar, M., Bevan, R., Dischamp, P., Moyart, D.: Power Analysis, What is Now Possible... In: T. Okamoto (ed.) Advances in Cryptology, Proceedings of ASIACRYPT 2000, LNCS, pp. 489–502 (2000)

2. Aumonier, S.: Generalized Correlation Power Analysis. Proceedings of the Ecrypt Workshop Tools For Cryptanalysis (2007)
3. Batina, L., Gierlichs, B., Prouff, E., Rivain, M., Standaert, F.X., Veyrat-Charvillon, N.: Mutual Information Analysis: A Comprehensive Study. Journal of Cryptology **24**, 269–291 (2011)
4. Benzi, R., Parisi, G., Sutera, A., Vulpiani, A.: Stochastic Resonance in Climatic Change. Tellus **34**(1), 10–16 (1982)
5. Bonachela, J., Hinrichsen, H., Munoz, M.: Entropy Estimates of Small Data Sets. Journal of Physics A – Mathematical and Theoretical **41**(20) (2008)
6. Brier, E., Clavier, C., Olivier, F.: Correlation Power Analysis with a Leakage Model. In: M. Joye, J.J. Quisquater (eds.) Proceedings of CHES 2004, *LNCS*, vol. 3156, pp. 135–152. Springer Berlin / Heidelberg (2004)
7. Chari, S., Rao, J., Rohatgi, P.: Template Attacks. In: B. Kaliski, Ç. Koç, C. Paar (eds.) Proceedings of CHES 2002, *LNCS*, vol. 2523, pp. 51–62. Springer Berlin / Heidelberg (2003)
8. Doget, J., Prouff, E., Rivain, M., Standaert, F.X.: Univariate Side Channel Attacks and Leakage Modeling. In: Proceedings of COSADE 2011, pp. 1–15 (2011)
9. Fasano, G., Franceschini, A.: A Multidimensional Version of the Kolmogorov-Smirnov Test. Monthly Notices of the Royal Astronomical Society **225**, 155–170 (1987)
10. Gierlichs, B., Batina, L., Preneel, B., Verbauwhede, I.: Revisiting Higher-Order DPA Attacks: Multivariate Mutual Information Analysis. In: J. Pieprzyk (ed.) Topics in Cryptology – CT-RSA 2010, *LNCS*, vol. 5985, pp. 221–234. Springer-Verlag, San Francisco, CA, USA (2010)
11. Gierlichs, B., Batina, L., Tuyls, P., Preneel, B.: Mutual Information Analysis: A Generic Side-Channel Distinguisher. In: E. Oswald, P. Rohatgi (eds.) Proceedings of CHES 2008, *LNCS*, vol. 5154, pp. 426–442. Springer-Verlag Berlin (2008)
12. Guilley, S., Hoogvorst, P., Pacalet, R.: Differential Power Analysis Model and Some Results. In: J.J. Quisquater, P. Paradinas, Y. Deswarte, A. El Kalam (eds.) Smart Card Research and Advanced Applications VI, *IFIP*, vol. 153, pp. 127–142. Springer Boston (2004)
13. Hutter, M.: Distribution of Mutual Information. In: T.G. Dietterich, S. Becker, Z. Ghahramani (eds.) Advances in Neural Information Processing Systems, vol. 14, pp. 399–406. MIT Press, Cambridge, MA (2002)
14. Kraemer, H.C., Thiemann, S.: How Many Subjects?: Statistical Power Analysis in Research, 1 edn. Sage Publications, Inc (1987)
15. Madiman, M.: On the Entropy of Sums. In: 2008 IEEE Information Theory Workshop (2008)
16. Mangard, S., Oswald, E., Popp, T.: Power Analysis Attacks: Revealing the Secrets of Smart Cards. Springer (2007)
17. Mangard, S., Oswald, E., Standaert, F.X.: One for All – All for One: Unifying Standard DPA Attacks. IET Information Security **5**(2), 100–110 (2011)
18. Mather, L.: The Multivariate Kolmogorov-Smirnov Test In Differential Power Analysis Attacks. Master's thesis, University of Bristol, Department of Computer Science (2010)
19. Messerges, T.S.: Using Second-Order Power Analysis to Attack DPA Resistant Software. In: Ç. Koç, C. Paar (eds.) Proceedings of CHES 2000, *LNCS*, vol. 1965, pp. 27–78. Springer Berlin / Heidelberg, London, UK (2000)
20. Micali, S., Reyzin, L.: Physically observable cryptography. In: M. Naor (ed.) Theory of Cryptography, *LNCS*, vol. 2951, pp. 278–296. Springer Berlin / Heidelberg (2004)
21. Moradi, A., Mousavi, N., Paar, C., Salmasizadeh, M.: A Comparative Study of Mutual Information Analysis Under a Gaussian Assumption. In: H. Youm, M. Yung (eds.) Information Security Applications, *LNCS*, vol. 5932, pp. 193–205. Springer Berlin / Heidelberg (2009)
22. Paninski, L.: Estimation of Entropy and Mutual Information. Neural Computation **15**(6), 1191–1253 (2003)
23. Peacock, J.: Two-Dimensional Goodness-of-Fit Testing in Astronomy. Monthly Notices of the Royal Astronomical Society **202**, 615–627 (1983)
24. Prouff, E.: DPA Attacks and S-Boxes. In: H. Gilbert, H. Handschuh (eds.) Fast Software Encryption, *LNCS*, vol. 3557, pp. 424–441. Springer Berlin / Heidelberg (2005)
25. Prouff, E., Rivain, M.: Theoretical and Practical Aspects of Mutual Information Based Side Channel Analysis. In: M. Abdalla, D. Pointcheval, P.A. Fouque, D. Vergnaud (eds.) ACNS, *LNCS*, vol. 5536, pp. 499–518. Springer Berlin / Heidelberg (2009)
26. Prouff, E., Rivain, M., Bevan, R.: Statistical Analysis of Second Order Differential Power Analysis. IEEE Transactions on Computers **58**(6), 799–811 (2009). DOI 10.1109/TC.2009.15
27. Rechberger, C., Oswald, E.: Practical Template Attacks. In: WISA, *LNCS*, vol. 3325, pp. 440–456 (2004)
28. Renauld, M., Standaert, F.X., Veyrat-Charvillon, N., Kamel, D., Flandre, D.: A Formal Study of Power Variability Issues and Side-Channel Attacks for Nanoscale Devices. In: K.G. Paterson (ed.) Proceedings of EUROCRYPT 2011, *LNCS*, vol. 6632, pp. 109–128. Springer (2011)
29. Schindler, W., Lemke, K., Paar, C.: A Stochastic Model for Differential Side Channel Cryptanalysis. In: J. Rao, B. Sunar (eds.) Proceedings of CHES 2005, *LNCS*, vol. 3659, pp. 30–46. Springer Berlin / Heidelberg (2005)
30. Shiga, M., Yokota, Y.: An Optimal Entropy Estimator for Discrete Random Variables. In: Proceedings of the International Joint Conference on Neural Networks, IEEE International Joint Conference on Neural Networks (IJCNN), pp. 1280–1285. IEEE, New York (2005)

31. Standaert, F.X., Gierlichs, B., Verbauwhede, I.: Partition vs. Comparison Side-Channel Distinguishers: An Empirical Evaluation of Statistical Tests for Univariate Side-Channel Attacks against Two Unprotected CMOS Devices. In: P. Lee, J. Cheon (eds.) ICISC 2008, *LNCS*, vol. 5461, pp. 253–267. Springer Berlin / Heidelberg (2009)
32. Standaert, F.X., Malkin, T.G., Yung, M.: A Unified Framework for the Analysis of Side-Channel Key Recovery Attacks. In: A. Joux (ed.) Advances in Cryptology, Proceedings of EUROCRYPT 2009, *LNCS*, vol. 5479, pp. 443–461. Springer-Verlag, Berlin, Heidelberg (2009)
33. Treves, A., Panzeri, S.: The Upward Bias in Measures on Information Derived From Limited Data Samples. Neural Computation **7**(2), 399–407 (1995)
34. Veyrat-Charvillon, N., Standaert, F.X.: Mutual Information Analysis: How, When and Why? In: C. Clavier, K. Gaj (eds.) Proceedings of CHES 2009, *LNCS*, vol. 5747, pp. 429–443. Springer Berlin / Heidelberg (2009)
35. Whitnall, C., Oswald, E.: A Comprehensive Evaluation of Mutual Information Analysis Using a Fair Evaluation Framework. In: Proceedings of CRYPTO 2011. Springer (2011)
36. Whitnall, C., Oswald, E., Mather, L.: An Exploration of the Kolmogorov-Smirnov Test as Competitor to Mutual Information Analysis. Cryptology ePrint Archive, Report 2011/380 (2011). `http://eprint.iacr.org/`