Choosing Distinguishers for Differential Power Analysis Attacks

Elisabeth Oswald, Luke Mather, and Carolyn Whitnall

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

Abstract. Differential power analysis attacks are among the 'classical' non-invasive types of attacks against physical devices. Attacks belonging to that class are well studied in the literature, however a seemingly simple yet very important question has proven to be exceptionally difficult to answer: given a cryptographic device, how do I best choose a distinguisher to actually perform a differential power analysis attack? This question needs to be unpicked before an attempt to an answer can be made: what is known about the power consumption characteristics of the device (everything—i.e. power profiles are available, not much—i.e. one can realistically assume a certain standard power model such as Hamming weight can be used, or nothing). Does the device allow control over its inputs to the cryptographic routine that is targeted? Are there any countermeasures built in, and if so which? In this article we aim to illuminate one particular aspect of such considerations. Namely, is there any best distinguisher, and consequently, can the choice of distinguisher and the modelling of the power consumption be made independently? Our approach in answering these questions is to draw from our own recent results and research into evaluation strategies for distinguishers, and linking them to other recent works. The conclusion that we can draw is that there is no generally best distinguisher, but for well defined scenarios there are best choices for a distinguisher in conjunction with a power model.

1 Introduction

Differential power analysis (DPA) attacks[1] use key-dependent hypotheses to extract information about secret data used in the cryptographic algorithm executed on a physical device. In this process, which can be broken up into five steps [15, ch. 6], the attacker selects a power model by which to map key-dependent intermediate values to hypothetical power consumption values, which are then somehow 'compared' to the real power consumption from the device using a statistic. This statistic is commonly referred to as a 'distinguisher' in the side channel literature and is the focus of this article.

An obvious question, that is somewhat simple and yet difficult to fully answer, is what is the best distinguisher? Is there any generally best method; is it possibly dependent on the devices' power consumption characteristics, or maybe even on the attackers' model thereof? The research community has been hunting for the best distinguisher ever since the introduction of power analysis attacks: shortly after the original DPA article, Messerges published so-called multi-bit attacks ([17] and [19]) which essentially involved using all predictable bits in a distance-of-means style attack. This was seen as an extension to the classical power analysis attack ([11]) that made only use of one predictable bit, also in a distance-of-means type distinguisher. Based on the generalisation of Messerges, more papers have claimed to improve attacks, e.g. [4] and [13] by using distinguisher results from various bits. Concurrently, researchers started exploring covariance/correlation type distinguishers ([6], [20]) which were then coined correlation power analysis (CPA) in [5].

The obvious advantage of using correlation with a suitable power model was clear from those papers, yet the jury was still out there whether correlation would always outperform other methods, or if certain conditions (e.g. quality of the available power model) would favour other distinguishers. Finally, template attacks [7] were recognised as the most powerful flavour but required profiling, a disadvantage that was seemingly overcome

[1] Whilst we use the term DPA in this article, our analysis and argument is not dependent on the type of side channel used.

by using mutual information (MI) as a distinguisher in [9]. However, MI based attacks failed to live up to the expectations as shown in subsequent, at first mainly empirical, studies such as [2].

In this article we focus our investigation on a fair comparison of the aforementioned distinguishers. For this purpose we analyse the existing literature and bring together those results that clarify the 'best choice' in certain scenarios. We make use of our own recent work, in particular, this article draws heavily from [32]. Our article is structured as follows. In Sect 2 we set the choice of the distinguisher in context with the overall DPA scenario and define give a precise definition of the distinguishers under consideration. In Sect. 3 we review the key points of research results that focus on studying distinguisher properties. We conclude in Sect 4.

2 Differential Power Analysis Attacks

We consider a 'standard DPA attack' scenario as defined in [16]. In short, we assume that the power consumption L of the target cryptographic device depends on some internal state $f_{k^*}(X)$. The state is a function of some part of the plaintext, which we denote by the 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 $L = L \circ f_{k^*}(X) + \varepsilon$, 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. In our scenario 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 present in the device 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: a distance-of-means inspired distinguisher was used in [11] as well as in follow-up work [19]. Pearson's correlation coefficient [5] is a particularly popular choice. Mutual Information Analysis (MIA) has been proposed as an enhancement to correlation DPA (CPA) which relies less on M [9], 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 [30].

We describe these distinguishers in more detail in Sect. 2.2, but let us first consider what it means for a DPA attack to be successful.

2.1 Success of DPA attacks

We concentrate on the notion of *key-recovery success* as formalised by Standaert *et al.* in [27]. 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' (i.e. achieve o-th order success with the least number of N of measurements) have thus received considerable attention in the literature (see [26] for thorough empirical

evaluation). As the available literature shows, general statements about the relative merits of particular methods are extremely hard to come by as attack outcomes are highly scenario-specific. Hence we now briefly review some of the contributing factors and the way they interact.

Factors Contributing to Success of DPA Attacks

Target intermediate function: 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 [10,21]. 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. In this article we hence concentrate on the DES and AES substitution boxes as well as bitwise exclusive-or to compare distinguishers.

Device leakage vs attacker's power model: 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 [7]) 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 about the leakage function 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. In particular we study adversaries who either use a Hamming weight model or the identity function (representing the absence of any power model). In line with previous work we choose three different leakage scenarios in this article. In the *optimistic* scenario, we assume that the data-dependent leakage really *is* proportional to the Hamming weight of the target intermediate function. In the *realistic* scenario, as motivated by [1], we assume that the true leakage is actually an unevenly weighted sum of the bits.[2] In the *challenging* (but still realistic) scenario we assume that the true leakage is a highly nonlinear function of the intermediate data.[3]

Noise: 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 [15]). In this paper we show results for different SNRs to represent high quality to low quality measurements, and include the possibility that signal processing might have been used to achieve an improved SNR. 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 (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.

[2] Specifically, we allow the least significant bit (LSB) to dominate with a relative weight of 10, since the experiments of [30] 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. Such non-standard leakage has recently been observed in the context of emerging nanoscale technologies [23], but also previously in typical hardware implementations of substitution boxes [14].

Theoretic Outcomes vs. Practical Outcomes At the beginning of Sect. 2.1 we briefly introduced theoretic success and (practical) success, which are related to theoretic and practical distinguishing vectors. As the distinction and relationship between these two vectors is crucial for the comparison of distinguishers, we return now to these concepts.

Suppose 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 (this is called effect 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* [12] 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 N required by an estimator are usually extremely difficult, at least under reasonable assumptions. The sample correlation coefficient is a somewhat exceptional case, as we explain in section 2.2.

Consequently, in general 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.

By abstracting away from the estimation problem in order to focus on theoretic distinguisher values we are however 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, it is possible to define theoretic outcome measures that are *highly indicative* of practical performance—as we explain in Sect. 2.3.

2.2 Background to Distinguishers

We now briefly describe the correlation coefficient, mutual information, and the Kolmogorov-Smirnov test statistic, and explain how they can be used to construct DPA distinguishers. We omit the often used distance-of-means test as a single-bit correlation based attack is equivalent.

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 [15]).

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 $L = L + \varepsilon$ with the hypothesis-dependent predictions M_k as follows:

$$\begin{aligned} D_\rho(k) &= |\rho(L, M_k)| \\ &= \left| \frac{\mathrm{cov}(L, M_k)}{\sqrt{\mathrm{var}(L)}\sqrt{\mathrm{var}(M_k)}} \right|. \end{aligned} \tag{1}$$

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. (as derived in Chap. 6.3 of [15]); in short, the larger the noise, the more diminished are the correlations. Most importantly, the theoretic distinguisher vector is scaled in such a way that the rankings and other *relative* features (such as the standard score and distinguishability measures defined in 2.3) 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 [15]).

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 ([18], Chap. 10 of [15] gives a good overview of available options, [22] details the best possible choice for pre-processing in the context of an optimistic Hamming weight leakage scenario).

Mutual Information-Based Distinguisher Mutual information is measured in bits and is most intuitively expressed in terms of entropies via Shannon's formula: $\mathrm{I}(A;B) = \mathrm{H}(A) - \mathrm{H}(A|B)$.

As a function of probability distributions, MI is notoriously problematic to estimate as we have explained in [31]. All estimators are biased, and furthermore 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.

Unfortunately, estimators for MI do not behave so 'nicely' as the sample correlation coefficient. 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[4].

Its application as an attack distinguisher is as follows:

$$\begin{aligned} D_{\mathrm{MI}}(k) &= \mathrm{I}(L; M_k) \\ &= \mathrm{H}(L) - \mathrm{H}(L|M_k) \\ &= \mathrm{H}(L) - \mathop{\mathbb{E}}_{m \in \mathcal{M}} \left[\mathrm{H}(L|M_k = m)\right], \end{aligned} \tag{2}$$

and because the 'unexplained' entropy (the second term) is smallest when the predictions are good, we expect (2) to be maximised for the correct key hypothesis $k = k^*$.

Unlike CPA the impact of noise on the MIA distinguishing vector is complex. 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.

MI generalises quite naturally to higher-order statistics via several different meaningful extensions. The authors of [2] presented three such notions and explored how each could be adapted to the purposes of DPA.

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

[4] Under strong simplifying assumptions, estimating an MIA parametrically can be shown to be equivalent to conducting a correlation attack [16].

$a_i \leq x$ 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 L with the hypothesis-dependent conditional traces $L|M_k$—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:

$$\begin{aligned} D_{\mathrm{KS}}(k) &= \mathbb{E}[K(L||L|M_k)] \\ &= \mathop{\mathbb{E}}_{m \in \mathcal{M}} \left[\sup_y |F_L(y) - F_{L|M_k=m}(y)| \right]. \end{aligned} \tag{3}$$

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

Whilst the sampling distribution of the KS test statistic is known, the distribution over the average of such statistics is not. Multivariate extensions of the KS test are somewhat more difficult to achieve, as one first needs to formulate an appropriate notion of a multivariate CDF; a discussion of this problem can be found in [33].

2.3 A Comparison Framework for Distinguishers

As mentioned before, 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* [12], provided the sampling distribution can be approximated—but this is not achievable in general (see Sect. 2.2).

Our solution, as first introduced in [31] and further extended in [32], 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 precise formulae are provided in [32]; descriptions and rationale are provided below.

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. 2.1. 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).

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 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).
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.

4. *Standard score*: This is the same as the "DPA signal-to-noise ratio" described by [10]: 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.

By computing the above measures for uniformly drawn plaintexts $X \overset{unif.}{\leftarrow} \mathcal{X}$, we are able to compare the 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).
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.

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. 2.2).

3 Results about 'Best' Distinguishers

We now analyse and summarise results that have been achieved for the distinguishers that we defined in the previous section. As we have pointed out, comparisons based on practical vectors suffer from the drawback that they rely on the quality of the used estimator. Hence conclusive results are not always possible. Comparing theoretic vectors overcomes this problem, but suffers from the drawback that one can not always deduce practical efficiency from theoretic distinguishability. Consequently if we want to gain a rounded view we must take into account results looking at both theoretic and practical vectors.

In the subsequent analysis we first look at results from comparing theoretic vectors as they will represent the 'best' case in terms of DPA outcomes. We then look at how they relate to practical outcomes and focus on trace efficiency. We provide a summary of key observations made when studying theoretic vectors. Thereafter we look at results based on comparing practical vectors only. We compare and contrast these results with each other and results from theoretic vectors.

3.1 Results concerning Theoretic Distinguishing Vectors

Our own recent work has been within the framework we have outlined in Sect. 2.3. In particular, our published work [31], [32], and [33] includes a comprehensive overview of correlation, MI, and KS based DPA attacks based on theoretic distinguishing vectors (we have provided results relating to practical vectors as well in [31] and [33] which confirm the meaningfulness of studying theoretic vectors).

Whilst we do not want to repeat our previous results and analysis here, we want to illustrate the meaning of the previously definitions on one concrete and very simple example in the following section. Thereafter we merely summarise key observations and results from our previous work on theoretic vectors.

Univariate Attacks Targeting the First DES S-Box We study the performance of several distinguishers in three leakage scenarios that we identified before as practically meaningful (i.e. the optimistic, realistic, and challenging scenarios, see Sect.2.1).

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. 2.2), 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* [3], 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.

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 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.

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

Practical Outcomes (Simulations) We also test MIA and KS distinguishers against simulated traces with different levels of Gaussian noise. For our MIA estimations we employ the heuristic rule favoured by the literature, and estimate PDFs via histograms with the number of bins equal to the cardinality of the power model image (i.e. 5 for the HW power model, 16 for the identity power model). Therefore, though these are not 'definitive' results (as no universally 'best' estimator exists) they do represent an established methodology and, as such, a meaningful basis for comparison with KSA. As it is well known that correlation-based DPA is the most efficient distinguisher in this optimistic scenario we omit it from the following Figures.

The first panel of Fig. 2 shows the mean number of traces needed to recover the key; the second panel shows the 90^{th}-percentile, i.e. the number needed to achieve a 90% success rate. KSA(HW) performs almost identically

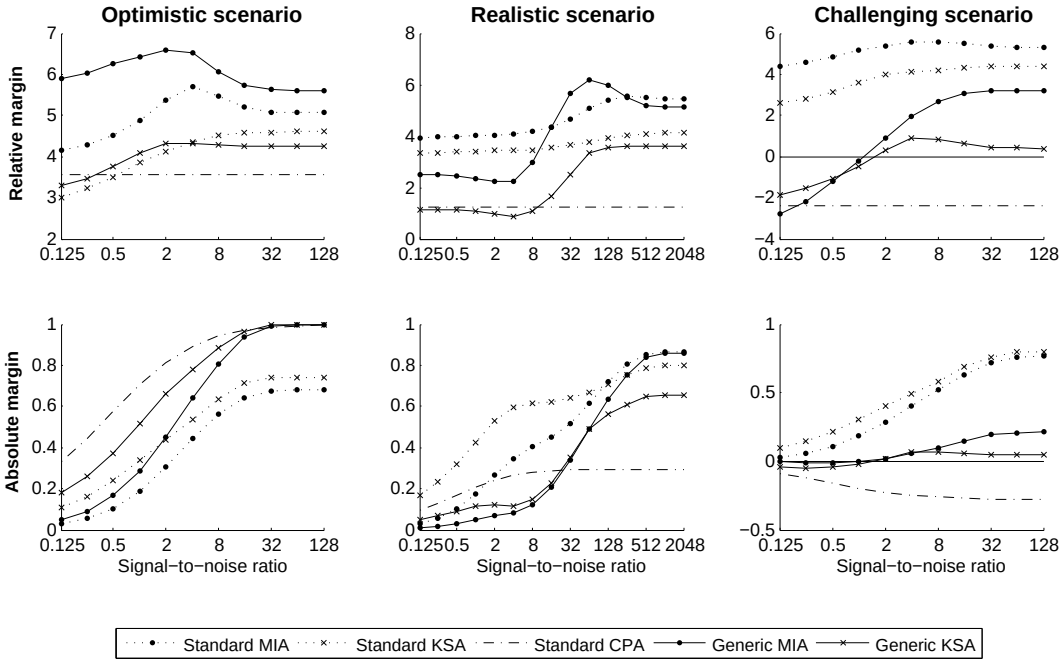

Fig. 1. Theoretic relative and absolute distinguishing margins as SNR varies, for standard and generic univariate attacks against the first DES S-Box.

to MIA(HW) (as could be expected from the theoretic vectors), with some evidence of a small advantage in weak-signal settings (again in keeping with the theoretic vectors). The ID attacks are more data intensive in both cases, but KSA(ID) exhibits consistently better performance than MIA(ID), probably due to the heavy estimation overhead incurred by the large number of bins required by the latter.

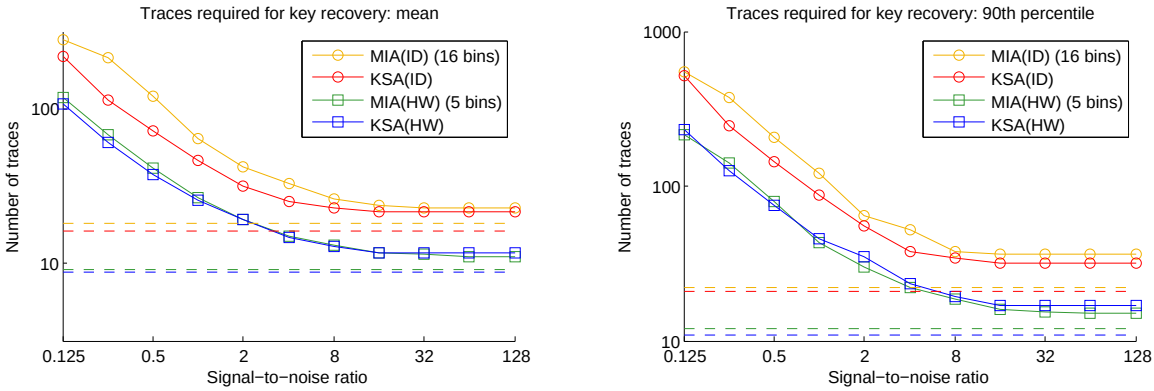

Fig. 2. Mean and 90^{th} percentile of the trace requirement for key recovery, in repeated experiments against simulated HW leakage of the first DES S-Box, as SNR varies.

Realistic Scenario—Hamming Distance to unknown reference state The scenario in which a device exhibits a Hamming distance leakage to an unknown but constant reference state turned out to favour MIA as a distinguisher over correlation based DPA in practice (see [31]). When including KSA into the comparison the picture is similar to that of the weighted bits leakage that we studied in [32]: KSA with the identity function

as power model performs better than MIA (with the identity function as power model) in low signal scenarios, and is outperformed by MIA in high signal scenarios. Depending on the reference state, correlation based DPA might or might not at all be able to recover the key.

Summary We summarise our key observations and conclusions in the following list:

1. We found that MIA has theoretic advantages 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.
2. MIA 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.
3. As a special case to the aforementioned unusual leakage scenarios we have found that MIA is able to distinguish keys in scenarios (and CPA is not) where the device leakage is that of the Hamming distance to an unknown reference state. We have also shown how this relates to leakage exhibited from dual-rail pre-charge logic.
4. 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.
5. We also observed that (in the context of MI or KS based attacks) 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.
6. We found that theoretic MI distinguishing vectors are profoundly influenced by noise. 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.

3.2 Results concerning Practical Distinguishing Vectors

The first systematic empirical comparison was [26] and they looked at distance-of-means based DPA (using single or several bits), correlation, MIA, Bayes, and a variance based attack. The empirical study involved two leakage scenarios (an optimistic and a realistic by our terminology). Noise levels were not varied for the scenarios, but were different between the two scenarios. As measures of performance the success rate for a given success order was used as well as the guessing entropy (which we do not use in the context of this article). General conclusions from the paper were the heavy reliance of MIA performance on the choice of estimator, the fact that many-bits attacks favour different distinguishers in different scenarios, and that whenever a good power model is available correlation outperforms any other distinguisher in standard unprofiled settings (when profiling is allowed Bayes attacks are superior as to be expected).

A different approach to evaluating distinguishers was taken in [16]. Rather than focusing on differences between distinguishers, the authors aim to show that in the so-called standard DPA scenario (also assumed in this article and in [26] and [8]), distance-of-means, correlation and Bayes based distinguishers are in fact equally efficient when supplied with the same power model. The notion of efficiency used is that of the average minimum number of power traces needed to achieve a certain success rate. The proof in the paper relies on a number of key assumptions (that the leakages can be modeled by Gaussian distributions, and that the average minimum traces

needed is sufficiently large). The article provides evidential support for this result even when the assumptions are not perfectly fulfilled (e.g. several of the supplied examples are for Hamming weight leakage), by performing attacks with the analysed distinguishers in optimistic, realistic, and challenging scenarios. Interestingly, the article also shows that under the strict assumption of Gaussian leakages and models, the mutual information can be expressed in terms of the correlation between models and predictions. From this it clearly follows that in the best case, the theoretic MI-based and correlation-based distinguishers are equally effective; however, as estimating MI incurs an overhead, it can be expected that correlation based DPA will outperform MI based DPA in those scenarios in practice. Another interesting side-note in this article is related to key and algorithm independence of DPA attacks: assuming the so-called EIS property [24], the success rate of an attack against, e.g., the AES SubBytes operation on a certain device should be no different to that of any other cipher using a cryptographically equivalent substitution box. Summarising the key message from this work though is that the most important factor contributing to the success of a DPA attack is that of the relationship between the power model of the attacker and the true leakage of the device.

This idea/insight was further developed in [8] where it is shown that under the assumption of having 'balanced' input sets, several distance-of-means based distinguishers can in fact be expressed as correlation based DPA with an appropriately chosen power model. The proofs in the article are constructive, hence these appropriate power models are derived in the paper. This result seems to strengthen [16] as the requirement for 'balanced' input sets seems slightly weaker than that of having 'a high enough average minimum sample size' (i.e. noisy enough measurements).

The key results of these three articles all give the same message: the choice of distinguisher is less important than that of the power model, and in many standard cases the distinguishers' performances are in fact equivalent.

4 Conclusions

After having defined standard DPA attacks and a number of distinguishers in this context, we have reviewed several key results about distinguisher performance in the literature. Two approaches to evaluating distinguisher performance can be followed: in the first approach one examines several key properties of so-called theoretic distinguishing vectors which are relevant for practical distinguisher performance. Results using this approach strongly suggest that only in certain scenarios (e.g. devices leaking the Hamming distance from an unknown reference state, or a highly non-linear function, or a function that is sufficiently different from the attackers power model, or is built in a not perfectly balanced dual-rail pre-charge logic style) one can expect distinguishers such as MI or KS to outperform correlation in practice. Results based on investigating practical distinguishing vectors have repeatedly shown that distinguisher performance is mainly depend on the power model. Hence these results echo the conclusions drawn from the study of theoretic vectors.

We promised at the outside of this article to give some guidance to choosing distinguishers. As can be expected from the preceding analysis, the key points to observe when choosing distinguishers are the leakage from the device and the availability of a good (or not) power model of the attacker. Table 2 gives the best choices in a number of relevant scenarios to the best of our current knowledge[5]. It shows that in the absence of an adequate power model, KSA with the identity power model is the best choice whenever noisy measurements are encountered. MIA with the identity power model is superior when high quality measurements are available. For both these cases we assume that the leakage model of the device is not 'simple' Hamming weight such that the attacker (in the unprofiled scenario) lacks a good power model. Importantly one should keep in mind that if an attacker is able to derive a 'good' power model (e.g. by first deriving a stochastic model if 'proper' profiling is not possible) then a correlation attack with this good model (assuming that there is a strong enough linear component in the power model) will most likely be the most efficient (in terms of number of power traces) option.

[5] This table is based on our data from [31], [32], and [33]

Table 2. Best distinguisher choices

Scenario	Leakage	SNR	Favoured Distinguisher
Optimistic	HW	Low/High	Correlation (HW)
Realistic	Weighted sum of bits	Low/High	KSA (ID)(/MIA (ID)
Challenging	non-linear	Low/High	KSA (ID)/MIA(ID)

References

1. Akkar, M., Bevan, R., Dischamp, P., Moyart, D.: Power Analysis, What is Now Possible... In ASIACRYPT 2000, pages 489–502, Springer LNCS 1976, Springer, 2000
2. L. Batina, B. Gierlichs, E. Prouff, M. Rivain, F.-X. Standaert, and N. Veyrat-Charvillon. Mutual Information Analysis: A Comprehensive Study. *Journal of Cryptology*, pages 1–23, 2010.
3. Benzi, R., Parisi, G., Sutera, A., Vulpiani, A.: Stochastic Resonance in Climatic Change. Tellus **34**(1), 10–16 (1982)
4. Regis Bevan, and Erik Knudsen Ways to Enhance Differential Power Analysis. In ICISC, pages 327–342, Springer LNCS 2587, 2002
5. E. Brier, C. Clavier, F. Olivier. Correlation Power Analysis with a Leakage Model. In *CHES*, pages 16–29, Springer, 2004.
6. Chari, S., Jutla, C. S., Rao, J.R., and Rohatgi, P. A note regarding evaluation of AES candidates on smart-cards. In Second AES Candidate Conference, pp. 133–147, 1999
7. Chari, S., Rao, J., Rohatgi, P.: Template Attacks. In CHES 2002, pages 51–62, Springer LNCS 2523, 2003
8. M. Rivain, J. Doget, E. Prouff and F.-X. Standaert. Univariate side channel attacks and leakage modeling (extended version). In *Journal of Cryptographic Engineering*, Springer, 1(2): 145-160 (2011).
9. B. Gierlichs, L. Batina, P. Tuyls, and B. Preneel. Mutual Information Analysis: A Generic Side-Channel Distinguisher. In *CHES*, pages 426–442. Springer, 2008.
10. Guilley, S., Hoogvorst, P., Pacalet, R.: Differential Power Analysis Model and Some Results. Smart Card Research and Advanced Applications Vi (2004) 127–142
11. P. C. Kocher, J. Jaffe, and B. Jun: Differential Power Analysis. In Crypto 1999, pages 338-397, Springer LNCS 1666, 1999
12. Kraemer, H.C., Thiemann, S.: How Many Subjects?: Statistical Power Analysis in Research, 1 edn. Sage Publications, Inc (1987)
13. T.-H. Le, J. Clediere, C. Canovas, B. Robisson, C. Serviere, J.-L. Lacoume. A Proposition for Correlation Power Analysis Enhancement. In CHES, pages 174-186, Springer LNCS 4249, 2006
14. S. Mangard, N. Pramstaller, E. Oswald Successfully Attacking Masked AES Hardware Implementations. In CHES 2005, pages 157–171, Springer LNCS 3659, Springer, 2005
15. Mangard, S., Oswald, E., Popp, T.: Power Analysis Attacks: Revealing the Secrets of Smart Cards. Springer (2007)
16. S. Mangard, E. Oswald, and F-X. Standaert. One for All - All for One: Unifying Standard DPA Attacks. In *IET Information Security*, 5(2), pages 100–110, 2011.
17. , T. S. Messerges and E. A. Dabbish and R. H. Sloan, Investigations of Power Analysis Attacks on Smartcards. In USENIX Workshop on Smartcard Technology (Smartcard '99), pages 151–162, 1999.
18. T. S. Messerges Using Second-Order Power Analysis to Attack DPA Resistant Software. In CHES 2000, pages 27–78, Springer LNCS 1965, Springer, 2000
19. T. S. Messerges, E. A. Dabbish, and R. H. Sloan Examining Smart-Card Security under the Threat of Power Analysis Attacks. In IEEE Trans. Computers 51(5): 541-552 (2002)
20. E. Oswald. On Side-Channel Attacks and the Application of Algorithmic Countermeasures. PhD dissertation, TU Graz, 2003
21. E. Prouff DPA Attacks and S-Boxes. In Fast Software Encryption, pages 424–441, Springer LNCS 355, Springer, 2005
22. E. Prouff, M. Rivain, and R. Bevan. Statistical Analysis of Second Order Differential Power Analysis. In IEEE Trans. Computers 58(6): 799-811 (2009)
23. M. Renauld, F.-X. Standaert, N. Veyrat-Charvillon, K. Kamel, D. Flandre A formal study of power variability issues and side-channel attacks for nanoscale devices. In EUROCRYPT 2011, pages 109–128, Springer LNCS 6632, Springer, 2011
24. W. Schindler, K. Lemke, and C. Paar. A Stochastic Model for Differential Side Channel Cryptanalysis. In CHES 2005, pages 30–46, Springer LNCS 3659, 2005
25. F.-X. Standaert, B. Gierlichs, and I. Verbauwhede. Partition vs. Comparison Side-Channel Distinguishers: An Empirical Evaluation of Statistical Tests for Univariate Side-Channel Attacks against Two Unprotected CMOS Devices. In ICISC 2008, pages 253–267, Springer LNCS 5461, 2009

26. F.-X. Standaert, B. Gierlichs, I. Verbauwhede Partition vs. Comparison Side-Channel Distinguishers: An Empirical Evaluation of Statistical Tests for Univariate Side-Channel Attacks against Two Unprotected CMOS Devices. In ICISC 2008, pages 253–267, Springer LNCS 5461, Springer 2009
27. F.-X. Standaert, T.G. Malkin, M. Yung A Unified Framework for the Analysis of Side-Channel Key Recovery Attacks. In EUROCRYPT 2009, pages 443–461, Springer LNCS 5479, Springer, 2009
28. F.-X. Standaert, N. Veyrat-Charvillon, E. Oswald, B. Gierlichs, M. Medwed, M. Kasper, and S. Mangard. The world is not enough: Another look on second-order dpa. In *ASIACRYPT*, pages 112–129, 2010.
29. Standaert, F.-X., Gierlichs, B., and Verbauwhede, I. Partition vs. Comparison Side-Channel Distingsuishers: An Empirical Evaluation of Statistical Tests for Univariate Side-Channel Attacks. In ICISC 2008, pages 253–267, Springer LNCS 5461, 2008
30. N. Veyrat-Charvillon, F.-X. Standaert Mutual Information Analysis: How, When and Why? In CHES 2009, pages 429–443, Springer LNCS 5747, Springer, 2009
31. C. Whitnall and E. Oswald A Comprehensive Evaluation of Mutual Information Analysis Using a Fair Evaluation Framework. In *CRYPTO*, 2011, LNCS 6841, pages 316–334, 2011
32. C. Whitnall and E. Oswald A fair evaluation framework for comparing side-channel distinguishers. J. Cryptographic Engineering 1(2): 145-160 (2011)
33. C. Whitnall, E. Oswald, L. Mather An Exploration of the Kolmogorov-Smirnov Test as Competitor to Mutual Information Analysis. CARDIS 2011, accepted for publication Cryptology ePrint Archive, Report 2011/380 (2011).