A Comprehensive Evaluation of Mutual Information Analysis Using a Fair Evaluation Framework

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 resistance of cryptographic implementations to side-channel analysis is a matter of considerable interest to those concerned with information security. It is particularly desirable to identify the attack methodology (e.g. differential power analysis using correlation or distance-of-means as the distinguisher) able to produce the best results. Such attempts are complicated by the many and varied factors contributing to attack success: the device power consumption characteristics, an attacker's power model, the distinguisher by which measurements and model predictions are compared, the quality of the estimations, and so on. Previous work has delivered partial answers for certain restricted scenarios. In this paper we assess the effectiveness of mutual information-based differential power analysis within a generic and comprehensive evaluation framework. Complementary to existing work, we present several notions/characterisations of attack success with direct implications for the amount of data required. We are thus able to identify scenarios in which mutual information offers performance advantages over other distinguishers. Furthermore we observe an interesting feature—unique to the mutual information based distinguisher—resembling a type of stochastic resonance, which could potentially enhance the effectiveness of such attacks over other methods in certain noisy scenarios.
Keywords: side channel analysis, mutual information

1 Introduction

Side-channel analysis (SCA) refers to a collection of cryptanalytic techniques for extracting secret information from the physical leakage of a device as it executes a cryptographic algorithm. Of the various types, one of the most popularly studied is differential power analysis (DPA); it involves applying some type of statistic (the *distinguisher*) to identify a correct hypothesis about (part of) the secret key from the set of all possible hypotheses about this key. Popular distinguisher choices are the Pearson correlation coefficient and the distance-of-means test. Mutual information (MI) measures the total dependency between two random variables, and was first proposed for use in DPA at CHES 2008 ([6]). *A priori* it was expected to display certain advantages over other distinguishers, loosely summarized by three (informal) conjectures:

1. By comprehensively exploiting *all* of the information contained within trace measurements it could have an efficiency advantage over existing side-channel distinguishers such as correlation (which measures linear dependencies only).
2. By capturing total dependency between the true device leakage and the modeled leakage it could prove effective in scenarios where an accurate model for the data-dependent leakage of the device is not known, thereby serving as a 'generic' distinguisher.
3. By natural extension to multivariate statistics it might be adapted to the context of higher-order attacks against (for example) protected implementations. Existing distinguishers operate on univariate data only and therefore require trace data to be pre-processed, resulting in loss of information.

Subsequent investigations such as [1,17,20,23] have found little evidence of the first two expectations being met in practice (there is rather more support for the third—see, for example, [1,5,17]). However, the literature has not been comprehensive in explaining why this might be. We must bear in mind that many factors influence DPA outcomes: not only the choice of distinguisher, but also the target intermediate function, the form of the data-dependent device leakage and how well this can be modeled, and the precision with which the distinguishing vector can be estimated using the resources and capabilities available. It is often unclear whether the observed underperformance of MI-based DPA is an inherent theoretical weakness of the distinguisher, a result of sub-optimal estimation procedures, or simply a failure to identify scenarios (i.e. combinations of target functions and power consumption patterns) where it offers a useful advantage: see Batina et al. [1] for an overview of these issues.

In this paper we introduce a framework for assessing and comparing DPA attacks in any given scenario on a theoretical basis, abstracting away from the problem of practical estimation. We use this to gain fresh insight into the findings of the existing literature and to clarify when and in what sense the *a priori* intuition regarding MI-based DPA *does* hold. Moreover, we are able to identify and describe attack scenarios in which MI-based DPA is theoretically successful whilst other distinguishers fail, or in which it displays a theoretic advantage large enough to potentially translate to a practical advantage. Further, we demonstrate that the (standardised) MI-based distinguishing vector exhibits the property of stochastic resonance as the noise levels in the power consumption vary. This feature, which is not shared by correlation-based DPA, could potentially be exploited to enhance MI-based attacks via noise injection.

In what follows, we first give the relevant preliminary information on DPA attacks, including details of particular distinguishers and a discussion of previous work in Sect. 2. In Sect. 3 we describe our methodology, whilst Sect. 4 reports on our findings as they relate to various attack scenarios. We conclude in Sect. 5.

2 DPA Attacks

We consider a 'standard DPA attack' scenario such as defined in [13]: The power consumption $\mathcal{L}$ of the target device depends on some internal value (or state)

$f_{k^*}(x)$: a function of some part of the plaintext $x \in \mathcal{X}$, as well as some part of the secret key $k^* \in \mathcal{K}$. Hence, we have that $\mathcal{L} = L \circ f_{k^*}(x) + \varepsilon$, 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. 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 model M.

DPA exploits the fact that the modeled power traces corresponding to the correct key hypothesis should bear more resemblance to the true power traces than do the modeled traces corresponding to incorrect hypotheses. An attacker is thus concerned with quantifying and comparing the degree of similarity between the true and modeled traces for each key hypothesis. A range of comparison tools—'distinguishers'—are available, of which mutual information and Pearson's correlation coefficient are popular examples. We introduce these formally and examine them in more detail in the remaining parts of this section. We use the shorthands CPA and MIA to refer (respectively) to correlation-based and MI-based DPA attacks.

2.1 Reasoning about the Success and Efficiency of DPA Attacks

Previous work has made some progress towards providing meaningful and practically relevant definitions for the 'success' and 'efficiency' of DPA attacks. Standaert's work [21] formalised the notion of key-recovery success (and, correspondingly, success rate), which we adopt for our purposes here: 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 $\#\{k \in \mathcal{K} : \hat{D}_N(k^*) \leq \hat{D}_N(k)\} < o$.

Since we are particularly interested in the impact of L on attack outcomes, it is desirable to abstract away from the impact of noise, as well as from the estimation process. We define a distinguisher as *ideally successful* if it is theoretically successful in a noise-free scenario.

Ideal success thus depends on the target intermediate function, the form of the data-dependent device leakage L, the set $\mathcal{X}' \subseteq \mathcal{X}$ of plaintexts being encrypted, and the choice of power model and distinguisher. Theoretic success is further determined by the size and distribution of the noise ε whilst practical success depends additionally on the choice of estimator for the distinguisher and

the number of trace measurements N. That is, given an attack which *theoretically* distinguishes the correct key (by a margin of a certain size), the practical outcome will be determined by whether or not an attacker has adequate resources to estimate $\hat{D}$ with sufficient precision to detect a difference of that size.

2.2 Distinguishers for DPA Attacks

Standaert *et al.* [20] provide a good overview of the many distinguishers that have been employed in the literature since DPA was first introduced in the late 1990s [9]. In this paper, we focus on mutual information and compare it with one other distinguisher of interest: Pearson's correlation coefficient.

In recent work, Mangard *et al.* [13] have shown that in the scenario of standard DPA attacks, the three most popular distinguishers, Pearson correlation, distance of means, and Bayes, are equally successful. Under additional, strong assumptions such that the MI can be estimated parametrically as a Gaussian mixture, they are even able to demonstrate a mapping between a correlation-based and an MI-based distinguisher. Our work relates to rather more general distributional assumptions.

Mutual Information Mutual information measures, in bits, the total information shared between two random variables X and Y. It is most intuitively expressed in terms of entropies via Shannon's formula: $\mathrm{I}(X;Y) = H(X) - H(X|Y)$.[1]

Mutual information is a functional of probability distributions, and estimation is a much studied problem with no simple answers ([3,8,14,19,22]). All estimators are biased, and further no 'ideal' estimator exists; 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. [23]). 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. In practice, due to the large sample space and small datasets we usually estimate the densities via an m-bin regularisation of the space. By an important data processing inequality[2] this means we are always estimating a lower bound on the mutual

[1] The original (but equivalent) definition is $\mathrm{I}(X;Y) = \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p_{X,Y}(x,y)\log_2\left(\frac{p_{X,Y}(x,y)}{p_X(x)\,p_Y(y)}\right)$, where $p_{X,Y}$ is the joint probability density of X and Y and p_X, p_Y are the marginal densities.

[2] $I(S(X);T(Y)) \leq I(X;Y)$ for any random variables X and Y and any functions S and T on the range of X and Y.

information—as the binning or mesh becomes finer the estimate approaches the true mutual information monotonically from below [14].

In security evaluations we often would like to be able to talk about the number of traces needed for an attack to be successful. This requires knowing the sampling distribution for the distinguisher under reasonable assumptions. Unfortunately, estimators for MI do not 'behave nicely' as do other statistics (such as the correlation coefficient—see below); in fact, there are no universal rates of convergence [14], so that whatever estimator we pick, we can always find a distribution for which the error vanishes arbitrarily slowly.

The relationship between the ideal MI and the theoretic MI in the presence of noise is complex (see, for example, [11]). In particular, whilst $\mathrm{I}(X+\varepsilon;Y) \leq \mathrm{I}(X;Y)$ (X, ε independent), nonetheless $\mathrm{I}(X;Y)-\mathrm{I}(X+\varepsilon;Y) \neq \mathrm{I}(X;Z)-\mathrm{I}(X+\varepsilon;Z)$. Thus, the elements of the theoretic MIA vector are differentially affected so that ideal outcomes do not directly generalise to theoretic outcomes in the presence of noise.

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

It is estimated from samples $\{x_i\}_{i=1}^N$, $\{y_i\}_{i=1}^N$ via the sample correlation coefficient: $r(X,Y) = \frac{\sum_{i=1}^N (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^N (x_i-\bar{x})^2}\sqrt{\sum_{i=1}^N (y_i-\bar{y})^2}}$. This is a consistent estimator for $\rho(X,Y)$ and, moreover, is asymptotically unbiased and efficient if X and Y have a joint Normal distribution. Under the same assumptions, we can even approximate the sampling distribution which leads to 'nice' results such as the number of trace measurements required for attacks to be successful (see Chap. 6.4 of [12]).

The relationship between the ideal correlation and the theoretic correlation in the presence of noise is straightforward. In fact, as derived in Chap. 6.3 of [12], $\rho(L+\varepsilon, M_k) = \frac{\rho(L,M_k)}{\sqrt{1+\frac{\sigma_\varepsilon^2}{\mathrm{Var}(L)}}}$. 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 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 [12])).

3 A Comprehensive Evaluation Framework

We compute and examine ideal/theoretic CPA and MIA vectors for a broad spectrum of possible leakage scenarios in unprofiled attacks where the true leak-

age L is unknown and modeled via the Hamming weight (HW) or the raw value (ID) of the target function output. For CPA, this is the same as assuming that the leakage is *proportional* to the HW or ID of the target, whereas for MIA this is the same as allowing the leakage to be *different* for each distinct HW or ID value, without any restriction on the nature of that dependency (for example, it needn't be a monotonic relationship). These vectors provide insight into the relative strengths and weaknesses of the distinguishers. We are particularly interested in finding scenarios where MIA has an ideal/theoretic advantage over CPA because we hope that a sufficiently large theoretic advantage would translate into a practical advantage. To do this we need to formulate an appropriate notion of "advantage".

An extremely desirable metric for security evaluation is 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* [10], provided the sampling distribution can be approximated—but this is not achievable in general (see Sect. 2.2), besides which we are seeking to avoid estimator-specific comparisons. Our solution 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:

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).
2. *Average distinguishing score*: The number of standard deviations above (or below) the mean for the distinguisher value corresponding to the correct key. This matches the "DPA signal-to-noise ratio" described by [7] and indicates the sensitivity of the attack in isolating the correct key: A very sensitive attack may be able to succeed in practice with only a few trace measurements, as even imprecise estimates will detect a large difference. A theoretically 'unsuccessful' attack may still be able to return a small candidate subset containing the correct key if the average distinguishing power is high.
3. *Nearest-rival distinguishing score*: The distance from the 'nearest rival' (i.e. the difference between the correct key distinguisher value and the value for the highest ranked alternative), normalised by the standard deviation. This represents, more directly than the average distinguishing power, the margin to be detected by a practical attack.

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 restricted information by inspecting ideal/theoretic attack vectors for reduced subsets of the plaintext space. These vectors depend not only on the size but also on the

composition of the input set; we cannot perform the computation exhaustively over the entire space of possible subsets (it is too large), but by repeated random draws of increasing size we can estimate the average support size needed for attack success. Thus we add the following measures as further clues to the "how many traces" problem:

5. *Average minimum 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. *Support required for x% success rate*: The support size for which the rate of success (of the appropriate order) is at least x 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 minimum support in the presence of high noise.

In some parts of this study it is more desirable to measure the *average* behaviour of an attack in a class of scenarios than to describe results under a specific scenario. This is relevant, for example, when considering functions of sufficient arbitrariness that we cannot detail each case exhaustively. In such cases, as with the analysis of restricted input support, we estimate average behaviour by using randomly sampled examples (note that the distinguishing vectors themselves are still *computed*, not estimated).

We acknowledge that data complexity is not the only measure of cost and that considerations such as computational complexity also play a role in determining the practicality of an attack. A formal study is outside the scope of this paper, but we do try to comment where appropriate.

Ideal/Theoretic vs. Practical Attacks. Recall that we define theoretic (as well as ideal, i.e. noise-free) attacks to abstract away from the impact of the estimation process (and from noise). As such, theoretic outcomes depend on the target intermediate function, the device leakage (including how much noise is present), the set of plaintexts used as inputs, the attackers choice/knowledge about the power model, and the theoretical distinguisher (which is in this case the estimand). Practical outcomes depend on an additional, crucial factor, namely the estimator—the quality of which, and the sensitivity to the underlying population parameters and noise, will ultimately determine whether an observed ideal/theoretical advantage is translated into a real advantage in a practical attack.

We consider several outcome measures to allow for a nuanced analysis of the distinguisher qualities contributing to practical outcomes. For example, the notion of ranking order is needed in addition to correct key ranking because, whilst ties are highly unlikely in the estimated vectors arising from practical

attacks, the underlying theoretic values may well rank keys equally. The approach of studying the distinguishing quality of the estimands separately from the qualities of the estimators is new and, as we will demonstrate in latter parts of the paper, provides fresh insight into the strengths and weaknesses of different distinguishers in practice.

4 Results

We now evaluate MIA and CPA distinguishers using the framework and considerations w.r.t. leakage models as spelled out before. For the sake of clarity and conciseness, we first show one detailed example (Hamming-weight leakage of a device implementing the DES algorithm), and then briefly report outcomes for some other leakage models. The choice for our focus is motivated by previous practical work which has focused on DES implementations [1], and the fact that DES is still used as predominant algorithm in the banking world. Note though that our framework could be used in the same way in a different context, and that the results of our evaluation of MI as a distinguisher are not strongly dependent on our specific choice.

4.1 Hamming-Weight Leakage

We begin with an ideal evaluation of MIA relative to CPA in the simplest and most popularly studied scenario: the first S-Box in a DES implementation (short: DES_{S1}) with a Hamming-weight (HW) leakage. As attacker power models we consider HW and the identity (ID) power model. For the sake of simplicity we use the following abbreviations: CPA(HW) as short-hand for correlation-based DPA with a HW power model, MIA(HW)/MIA(ID) as short-hand for MI-based DPA with a HW/ID power model, and MMIA for multivariate MI-based DPA. Using the notation as introduced before we first evaluate

$$\mathrm{CPA(M)} : \{\rho(L \circ \mathrm{DES}_{S1}(x, k^*), M \circ \mathrm{DES}_{S1}(x, k))\}_{k \in \mathcal{K}}, \tag{1}$$

$$\mathrm{MIA(M)} : \{\mathrm{I}(L \circ \mathrm{DES}_{S1}(x, k^*); M \circ \mathrm{DES}_{S1}(x, k))\}_{k \in \mathcal{K}} \tag{2}$$

assuming that both the attacker's power model, as well as the device's power model is the Hamming weight, i.e. $L = M = HW$.

This is a scenario in which we expect CPA(HW) to perform well: the use of the true power model enables perfect prediction of the data-dependent leakage under the correct key hypothesis, whilst the choice of the S-Box as target ensures that the alternative hypotheses will each give rise to substantially different predictions (see [16]).

Figure 1 shows the ideal distinguisher values for a CPA(HW) and an MIA(HW) attack. Since the target function has the Equal Images under different Subkeys (EIS) property [18] and the plaintexts are assumed uniformly distributed, attack outcomes are key independent [13]: the correct hypothesis yields the same distinguisher value under any key, and only the arrangement of the remaining vector entries changes.

It is evident that both attacks are first-order successful by a clear margin, but that MIA(HW) has a substantial ideal advantage, with a nearest-rival distinguishability score of 5.61 compared with just 2.14 for CPA(HW). This simple result confirms that it must instead be a combination of the impact of noise and the relative efficiency of estimating the correlation coefficient which enables CPA to consistently outperform MIA in practical attacks with a good power model.

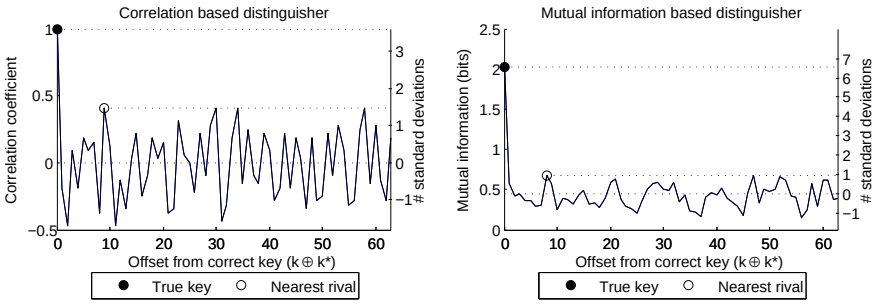

Fig. 1: *Ideal distinguishing vectors using the HW power model against the output of the first DES S-Box.*

As a partial insight into the quantity of data needed we next look at the minimum input support size required for the distinguishers to approach their full ideal potential. The space of possible plaintext combinations is too large to explore exhaustively, so we look at the average behaviour of the attacks in repeated random draws from the plaintext space. We find that CPA is able to identify the correct key from a far smaller support than MIA, requiring just 6 inputs on average, and achieving 100% success with just 12, compared with an average of 8 and threshold of 14 for MIA. Note as well that even once a high ideal success rate is achieved, it may be that a broader support is required before MIA regains the distinguishing advantage it displays with respect to the full distribution.

We next investigate the enhancement of MIA via the incorporation of an additional data point in a multivariate attack on AddRoundKey (short: DES_{ARK}) and the first S-Box jointly. Figure 2 plots the ideal outcome[3]. First observe that the distinguisher values are greater in size (by a factor of about two) than that of the single point attack—that is, we *are* capturing a larger amount of information. However, the increase applies across the range of key hypotheses so does not

[3] Note that what we are proposing here is to use the mutual information between two *bivariate* variables; since joint entropy is well-defined this is entirely consistent with the formulation of MI described in Section 2.2. However, there are other notions of 'multivariate mutual information' which become more interesting and relevant in the context of higher-order attacks against protected implementations—see [1] for a full discussion.

automatically raise the distinguishing power. In fact the true key is less strongly distinguished than in the attack against the S-Box alone: the nearest-rival distinguishability is reduced from 5.61 to 3.66. Moreover, the attack requires a larger input support—13 on average compared with 8 for MIA(HW).

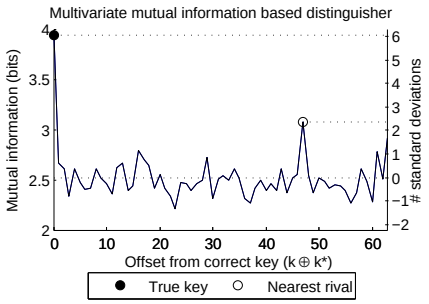

Fig. 2: *Ideal MIA vector against the DES AddRoundKey and the first S-Box jointly.*

Table 1 summarises outcomes for a wider selection of attacks, including MIA(ID): the proposed 'generic' attack of [6]. Unsurprisingly, in this first example where the leakage *is* proportional to the HW, MIA(ID) displays a disadvantage relative to MIA(HW). The generic capabilities of MIA will be of more relevance in leakage scenarios where the attacker is *not* able to correctly model the true leakage.

The attacks against AddRoundKey well illustrate the role of the target function: distinguishing power is greatly reduced in the case that incorrect key hypotheses give rise to outputs closely resembling the correct key outputs. Greater precision (and therefore a greater number of measured traces) will be required in order to detect a difference of this size in a practical attack, and moreover in the case of MIA there will remain an ambiguity between the true key k^* and its bitwise complement $\bar{k}^*$.

Stochastic Resonance We conclude this section by briefly considering the impact of (Gaussian) noise on theoretic outcomes. Figure 3 plots distinguishing scores against an increasing signal-to-noise ratio (SNR, defined as $\frac{\mathrm{var}(L \circ f_{k^*}(X))}{\mathrm{var}(\varepsilon)}$), confirming that (standardised) MIA outcomes are not constant. Moreover, the relationships are not monotonic: in each case there seems to be an optimal SNR at which the distinguishing scores reach a maximum, after which they diminish to that of the ideal (as depicted by the dashed lines). Such a phenomenon is a type of *stochastic resonance* [2], which can (in principle) occur in any nonlinear measurement system. Perhaps surprisingly, the required support sizes for both MIA(HW) and MIA(ID) match the ideal requirements and remain constant—though in general, such measures could also be subject to similar effects.

Table 1: *Ideal strength of CPA and MIA attacks against DES with Hamming weight leakage.*

	AddRoundKey		**First S-Box**			**Multivariate**
DES with a HW leakage	**CPA (HW)**	**MIA (HW)**	**CPA (HW)**	**MIA (HW)**	**MIA (ID)**	**MMIA (HW)**
Correct key ranking (order)	1 (1)	1 (2)	1 (1)	1 (1)	1 (1)	1 (1)
Average score	2.45	4.48	3.61	6.59	6.35	6.04
Nearest-rival score	0.82	0.00	2.14	5.61	5.08	3.66
Average minimum support	6	9	6	8	16	13
Support required for 90% SR	8	11	8	11	19	15
Support required for 100% SR	11	15	12	14	22	21

Recall, from Sect. 2.2, that by the properties of correlation, (standardised) CPA outcomes are unaffected by the level of noise. Hence the opportunity to enhance MIA (at least theoretically) by varying the noise is not available in the context of CPA.

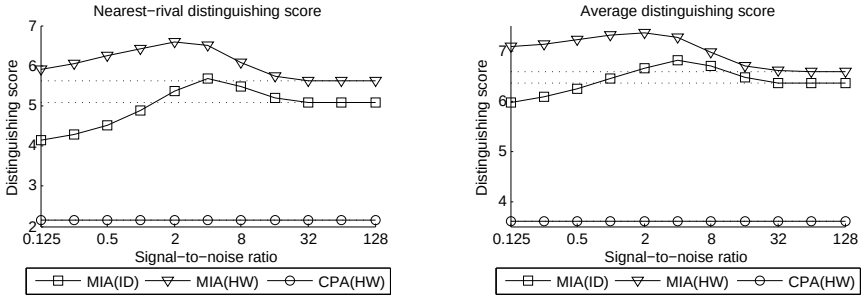

Fig. 3: *The effect of Gaussian noise on HW and ID attacks against HW leakage of the first DES S-Box.*

4.2 Hamming-Distance Leakage

Whilst the Hamming weight model is very popular in the literature, Hamming distance leakage can be widely observed in practical devices using CMOS logic. Broadly speaking there are three scenarios which may be encountered. Firstly, the previous state is known to the attacker, in which case the attacks are equivalent to Hamming weight attacks. Secondly, the previous state is unknown to the attacker but fixed. Thirdly, the previous state is unknown to the attacker and can vary. The latter two scenarios are the focus of the following discussion.

Constant Reference State Now let us suppose, as in [4], that the reference state is a constant but unknown machine word R. The device no longer leaks $\mathrm{L}(f_{k^*}(X))$ but rather $\mathrm{L}(R \oplus f_{k^*}(X))$.

First observe that no attack against a linear target function such as AddRoundKey can achieve first-order success, because the 'true key' values are perfectly replicated under an incorrect key hypothesis, namely $k^* \oplus R$. The power consumption for a plaintext X will be proportional to $\mathrm{HD}((k^* \oplus X), R) = \mathrm{HW}((k^* \oplus X) \oplus R) = \mathrm{HW}((k^* \oplus R) \oplus X)$, so that when our hypothesis is $k = k^* \oplus R$ we get maximum correlation/MI (for both HW and ID models) and in fact the theoretical distinguishing vector is identical to that of a successful attack against HW leakage with a key of $k^* \oplus R$.

Targeting the S-Box avoids this predicament thanks to the high nonlinearity of the S-Box. In particular, there is no R' such that $\text{S-Box}(k^* \oplus X) \oplus R = \text{S-Box}((k^* \oplus R') \oplus X)\ \forall X \in \mathcal{X}$, so no incorrect key will produce the correct predictions. It remains to be seen whether the resemblance between the imperfect predictions (with naive power models) and the true power consumption remains strong enough for the correct key and weak enough for the alternative hypotheses for any sort of attack to be successful.

Ideal CPA(HW) succeeds precisely in those scenarios where the HW of the reference is 1 (or 0) and fails whenever it is 2 (see Table 2). Further, were we to use the absolute value of the correlation to distinguish (denoting this strategy as |CPA(HW)|) the resulting ideal attack would succeed whenever the HW of the reference state is 3 or 4; however, there is a substantial reduction in theoretic strength when the HW is 1 or 3, and for some reference states |CPA(HW)| requires almost the entire plaintext set to determine the correct key.

MIA(HW) succeeds in all scenarios and gains a considerable advantage both in terms of the ideal distinguishing scores with full information (nearest-rival scores are in the range of 3.6 to 4.5 for MIA(HW) but just 0.5-2.7 for |CPA(HW)|) and also in terms of the minimum input support required for success (on average, 14 to 15 for MIA(HW) compared with 17 to 18 for |CPA(HW)|). We can take advantage of the non-injectivity of the DES S-Box to launch generic MIA(ID) attacks. As the authors of [6] observed, these are essentially unaffected by a constant reference state so that the nearest-rival distinguishing score is always around 5 for MIA(ID) and average support requirement around 16. This means that when $R \in \{0000_{(2)}, 1111_{(2)}\}$ (i.e. L is the HW function) the generic attacks are less effective than the equivalent methods combined with a HW power model, but in all other reference state scenarios they gain an advantage. The consistency and ideal strength of these attacks might be sufficient to translate into a practical advantage—a possibility which we will investigate in a latter section.

We have thus shown that MIA applied with little consideration for or knowledge about the true leakage can be effective even when that leakage actually depends on an unknown reference state. CPA, applied equally blindly, is far less likely to yield a successful attack. However, Brier et al. ([4]) showed how to adapt it in order to determine R as an unknown of the problem in addition to $f_{k^*}(X) \oplus R$, which together reveals the secret key k^*. Whilst this simultane-

Table 2: *Theoretical strength of CPA and MIA attacks against DES with Hamming distance leakage from a constant reference state.*

4 LSBs of reference state	**CPA (HW)**	**|CPA| (HW)**	**MIA (HW)**	**MIA (ID)**
Hamming weight 1				
Correct key ranking	1	1	1	1
Average score	2.04-4.05	2.56-4.94	5.48-5.97	5.81-6.46
Nearest-rival score	0.38-2.28	0.53-2.65	3.60-4.47	4.57-5.20
Average minimum support	17-25	20-34	14-15	16-17
Support required for 90% SR	31-49	33-53	20-22	19-20
Support required for 100% SR	40-59	44-61	28-32	21-24
Hamming weight 2				
Correct key ranking	27-32	54-63	1	1
Average score	0.00	-1.94-0.00	5.06-5.53	5.98-6.43
Nearest-rival score	-2.31-0.00	-5.62-0.00	3.05-3.16	4.49-5.42
Average minimum support	-	-	17-18	16-16
Support required for 90% SR	-	-	26-29	19-20
Support required for 100% SR	-	-	33-36	22
Hamming weight 3				
Correct key ranking	64	1	1	1
Average score	-2.44-0.00	2.56-4.94	5.48-5.97	5.81-6.46
Nearest-rival score	-4.58-0.00	0.53-2.65	3.60-4.47	4.57-5.20
Average minimum support	-	20-34	14-15	16-17
Support required for 90% SR	-	33-53	20-22	19-20
Support required for 100% SR	-	44-61	28-32	21-24
Hamming weight 4				
Correct key ranking	64	1	1	1
Average score	0.00	5.14	6.59	6.35
Nearest-rival score	0.00	3.56	5.61	5.08
Average minimum support	-	6	8	16
Support required for 90% SR	-	8	11	19
Support required for 100% SR	-	12	14	22

ous search process is more computationally costly than a standard CPA(HW) attack, MIA with an ID power model can itself be computationally costly in addition to the likely data complexity overheads. Further work (and broader cost considerations) would be required to establish which of the two methods is most practical.

A Note on DRP logic. We observe an important and useful parallel between HD leakage and the expected behaviour of DPA-resistant dual-rail precharge (DRP) logic. In fact, an imperfect realisation of the logic style can be shown to exhibit data-dependent power consumption of a similar form to the HD from a constant reference state, enabling us to clarify its vulnerability to the 'generic' MIA(ID) attack described by Gierlichs et al. in [6].

DRP logic attempts to eradicate the data-dependency of the power consumption by making it equal in each clock cycle. This is achieved insofar as the capacitances of the complementary output wires in each logic gate can be balanced, a difficult feat in practice ([15]). Suppose the i^{th} bit of an m-bit word x is carried by a DRP logic gate driving two differential outputs with imperfectly balanced capacities (α_i, β_i), so that $\alpha_i = \beta_i + \gamma_i$. The power consumption of such a circuit can be shown to be equivalent to leakage scenarios with which we are more familiar, enabling us to comment on theoretical attack capabilities.

Let us initially consider the simplified case that both capacitances are the same throughout the circuit: $\beta_i = \beta$, $\alpha_i = \beta + \gamma$, $\forall i \in \{0, \ldots, m-1\}$. Then the data-dependent leakage is proportional to:

$$\begin{aligned} \mathrm{HW}(x)\alpha + \mathrm{HW}(\bar{x})\beta &= \mathrm{HW}(x)(\beta+\gamma) + \mathrm{HW}(\bar{x})\beta \\ &= (\mathrm{HW}(x) + \mathrm{HW}(\bar{x}))\beta + \mathrm{HW}(x)\gamma \\ &= m\beta + \mathrm{HW}(x)\gamma \end{aligned}$$

The constant $m\beta$ is absorbed into the non-data-dependent component and we thus obtain the result that the leakage is proportional to the Hamming weight. Both CPA(HW) and MIA(HW) will be theoretically capable of returning the correct key; practical success will depend on ability and resources to estimate the distinguishing vectors with sufficient precision, in which case CPA(HW) is likely to have an advantage, as we have already seen.

Now suppose that the capacitances are the same throughout the circuit but that the order changes, i.e. so that some gates have capacitances (α, β) and others (β, α), where $\alpha = \beta + \gamma$. We can express this by introducing $R = (r_0, \ldots, r_{m-1}) \in \{0,1\}^m$ such that gate i is (β, α) if $r_i = 1$ and (α, β) otherwise. Then the data-dependent leakage is:

$$\begin{aligned} \mathrm{HW}(x \oplus R)\alpha + \mathrm{HW}(x \oplus \bar{R})\beta &= \mathrm{HW}(x \oplus R)(\beta+\gamma) + \mathrm{HW}(x \oplus \bar{R})\beta \\ &= (\mathrm{HW}(x \oplus R) + \mathrm{HW}(x \oplus \bar{R}))\beta + \mathrm{HW}(x \oplus R)\gamma \\ &= m\beta + \mathrm{HW}(x \oplus R)\gamma \end{aligned}$$

That is, the data-dependent leakage is proportional to the Hamming distance from R, which equates to the scenario of a more conventional logic style (such

as CMOS) consuming power proportional to the number of transitions from a constant, unknown reference state. We have already shown that MIA(ID) remains ideally successful against such leakage, whilst CPA(HW) is (depending on the state) either unsuccessful or greatly reduced in distinguishing power. This confirms that DRP logic gives rise to leakage scenarios under which first-order MIA(ID) could be useful, in particular, shedding light on the experimental result of [6].

In the most general case, the size of the capacitances and not just the direction of the differences may vary over the circuit. Suppose the gates corresponding to bits $i = 1, \ldots, m$ have capacitances (α_i, β_i) such that $\alpha_i = \beta_i + \gamma_i$ where γ_i can be positive or negative. Letting $\mathbf{x} = (x_1, \ldots, x_m)$ and $\alpha = (\alpha_1, \ldots, \alpha_m)$, $\beta = (\beta_1, \ldots, \beta_m)$, $\gamma = (\gamma_1, \ldots, \gamma_m)$ we get a leakage function of $\mathbf{x} \cdot \alpha + (\mathbf{x} \oplus \mathbf{1}) \cdot \beta = (\mathbf{x} + \mathbf{x} \oplus \mathbf{1}) \cdot \beta + \mathbf{x} \cdot \gamma = \mathbf{1} \cdot \beta + \mathbf{x} \cdot \gamma$, so that the data-dependent power consumption is proportional to a weighted combination of the bits of $\mathbf{x}$, where the weights can take negative values. Further investigation is needed to establish the expected behaviour of our distinguishers as the relative weights become increasingly disproportionate.

Data-Dependent Reference State We next investigate ideal performance against Hamming distance leakage allowing for R to take two or more different values depending on the plaintext, unknown to the attacker, but restricting it to be constant in repeated runs. In practice this could happen due to an incorrect implementation of a masking scheme.

In the (commonly studied) case of an 8-bit micro-controller, the reference states (or masks) take values in $\{0,1\}^8 = \{0, \ldots, 255\}$. Since our attacks on the first DES S-Box target 6-bit key portions, our plaintext inputs are drawn from $\{0,1\}^6 = \{0, \ldots, 63\}$—there could be up to 64 different input-dependent reference states. The number of possible ways that r reference states could be associated with the 64 input values is given by the Stirling number of the second kind: $\left\{ {64 \atop r} \right\} = \frac{1}{r!} \sum_{j=0}^{r} (-1)^{r-j} \binom{r}{j} j^{64}$, so it is no longer possible to exhaustively explore every scenario. Instead, we calculate the success rates in 1,000 random experiments for increasing numbers of different reference states, randomly assigned to approximately equal-sized subsets of the input space (see Figure 4). [4]

We find that MIA is much better able to succeed than |CPA|, particularly when provided with an ID power model—although even then it does not achieve 100% success for attacks with more than 2 different states and for more than 6 states success rates drop to below 50%. The success of |CPA(HW)| degrades rapidly; for attacks with about 20 different states it is no better than a random

[4] When the reference state is constant, only the 4 bits which are replaced by the S-Box output contribute to the data-dependent leakage whilst the contribution of the remaining bits is absorbed into the static component of the power consumption. However, when the state depends on the data in the manner described here, the contribution of the remaining bits *does* need to be taken into consideration as it becomes part of the data-dependent power consumption.

guess, whilst MIA(ID) and even MIA(HW) appear to retain some advantage over guessing.

Thus, when very little is known about the leakage an attacker may well be able to recover a great deal of information just by applying a 'blind' MIA—though even ideal success will be partially determined by chance, and the number of traces required for adequate estimation may be prohibitive. Such an approach may not be the best way of exploiting the available data: where resources permit, it may prove more effective or efficient to refine a CPA based approach (or similar), investing greater effort in understanding the leakage to begin with, perhaps through profiling.

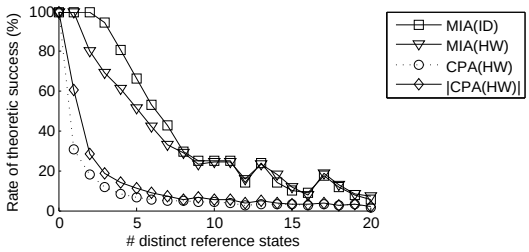

Fig. 4: *Ideal success against the first DES S-Box in the presence of data-dependent reference states of length 8 bits, as the number of different states increases.*

4.3 Theoretical vs. Practical Success

We now return to a scenario which was identified as a candidate for MIA to hold an advantage over CPA in practice: Hamming-distance leakage from a reference state unknown to the attacker (taken to be $0100_{(2)}$ for the purposes of our example). We wish to investigate whether the observed ideal advantages generalise (theoretically) in the presence of noise and hence whether they can be translated into practical advantages. Figure 5 shows the impact of Gaussian noise on theoretic attack effectiveness, both in terms of nearest-rival distinguishability and in terms of the minimum support size required for first-order success. MIA(HW) distinguishability is not very robust to the addition of noise, even falling below that of CPA(HW). Moreover, there is a hefty penalty in terms of required support size. By contrast, MIA(ID) distinguishability is more robust and even exhibits some evidence of stochastic resonance-type behaviour, whilst required support size remains constant in the tested range.

Our simulated attacks use histogram-based estimators where bin counts are chosen equal to the cardinality of the power model domain, according to the heuristic which has emerged from the literature (see, for example, [1]). In a pure-signal scenario (see the dashed lines in Figure 6) the 5-bin estimator for

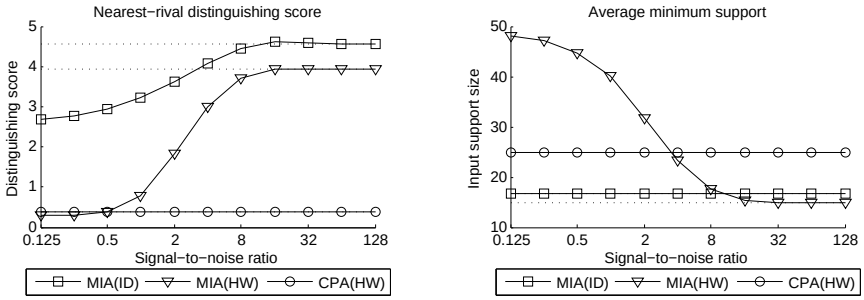

Fig. 5: *Nearest-rival distinguishability and required support size of theoretic attacks against Hamming distance leakage (with a reference state of* $0100_{(2)}$*) for varying levels of Gaussian noise.*

MIA(HW) requires fewer traces than CPA(HW) to identify the correct key, but the introduction of even the smallest amount of noise incurs a burden so that across the tested range it is substantially less efficient. By contrast, the 16-bin estimator for MIA(ID) approaches the efficiency achieved in the pure-signal scenario as the SNR increases, and moreover substantially outperforms CPA(HW) once the SNR is at least 1. We have thus confirmed that—in this instance at least—ideal MIA advantages *can* be translated into practical advantages.

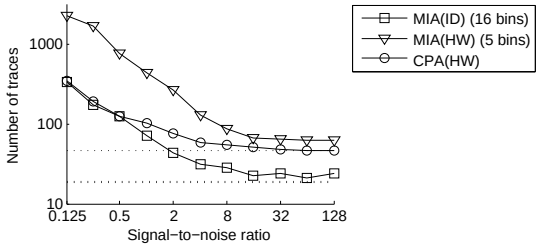

Fig. 6: *Average number of traces required for key recovery in simulated practical attacks against Hamming-distance leakage (with a reference state of* $0100_{(2)}$*), for varying levels of Gaussian noise.*

5 Conclusions

In this paper we have presented a framework for evaluating and comparing DPA methodologies on a like-for-like, ideal/theoretic basis. Our outcome measures allow for a nuanced assessment of the relative strengths and weaknesses of particular distinguishers as employed under different leakage scenarios. We have

thus been able to compare MIA and CPA as abstracted away from the confounding problem of estimation, gaining valuable insight into the empirical results of existing literature which tends to focus on practical instantiations of the attacks. We have identified scenarios in which MIA offers a substantial theoretic advantage over CPA, and demonstrated that such theoretic advantages can be translated into practical advantages. Particular candidate scenarios for MIA to be useful arise when the leakage takes the form of the Hamming distance from an unknown reference state or in implementations using dual-rail precharge logic—and, in fact, we are able to demonstrate a relationship between these two cases. The generic capabilities of MIA are found to be an advantage as the HW model degrades relative to the true leakage, but multivariate extensions do not exhibit much if any advantage over univariate attacks in the first-order 'unprotected' setting. Lastly, we observe for the first time (to our knowledge) the noise-sensitivity of the (standardised) MIA distinguishing vector, which exhibits an effect which can be likened to stochastic resonance and which could possibly be exploited in certain noisy scenarios to enhance the distinguishing ability of MIA attacks. This is a question for further research. Another open problem—persistently arising in the context of MIA—is that of finding estimators which most effectively translate theoretical advantages into practical ones.

6 Acknowledgements

The authors would like to thank Dave Cliff for pointing them towards the concept of stochastic resonance, and the anonymous referees for their comments. The first author of this paper has been funded via an EPSRC studentship. The second author has been supported by an EPSRC Leadership Fellowship I005226.

References

1. Akkar, M., Bevan, R., Dischamp, P., Moyart, D.: Power analysis, what is now possible ... In: Okamoto, T. (ed.) Advances in Cryptology ASIACRYPT 2000, Proceedings. pp. 489–502. Lecture Notes in Computer Science (2000)
2. Batina, L., Gierlichs, B., Prouff, E., Rivain, M., Standaert, F.X., Veyrat-Charvillon, N.: Mutual Information Analysis: A comprehensive study. Journal of Cryptology pp. 1–23 (2010)
3. Benzi, R., Parisi, G., Sutera, A., Vulpiani, A.: Stochastic resonance in climatic change. Tellus 34(1), 10–16 (1982)
4. Bonachela, J., Hinrichsen, H., Munoz, M.: Entropy estimates of small data sets. Journal of Physics A – Mathematical and Theoretical 41(20) (2008)
5. Brier, E., Clavier, C., Olivier, F.: Correlation power analysis with a leakage model. In: Joye, M., Quisquater, J.J. (eds.) Cryptographic Hardware and Embedded Systems – CHES 2004, Proceedings. Lecture Notes in Computer Science, vol. 3156, pp. 135–152. Springer Berlin / Heidelberg (2004), http://dx.doi.org/10.1007/978-3-540-28632-5_2
6. Gierlichs, B., Batina, L., Tuyls, P., Preneel, B.: Mutual information analysis: A generic side-channel distinguisher. In: Oswald, E., Rohatgi, P. (eds.) Cryptographic

Hardware and Embedded Systems – CHES 2008, Proceedings. Lecture Notes in Computer Science, vol. 5154, pp. 426–442. Springer-Verlag Berlin (2008)
7. Guilley, S., Hoogvorst, P., Pacalet, R.: Differential power analysis model and some results. Smart Card Research and Advanced Applications Vi pp. 127–142 (2004)
8. Hutter, M.: Distribution of mutual information. Advances in Neural Information Processing Systems 14, 399–406 (2002)
9. Kocher, P., Jaffe, J., Jun, B.: Differential power analysis. In: Proceedings of CRYPTO 1999. pp. 388–397. Springer-Verlag (1999)
10. Madiman, M.: On the entropy of sums. In: 2008 IEEE Information Theory Workshop (2008)
11. Mangard, S., Oswald, E., Popp, T.: Power Analysis Attacks: Revealing the Secrets of Smart Cards. Springer (2007)
12. Mangard, S., Oswald, E., Standaert, F.X.: One for all - all for one: Unifying standard DPA attacks. IET Information Security (2011), to appear, preprint available from `http://eprint.iacr.org/2009/449`
13. Paninski, L.: Estimation of entropy and mutual information. Neural Computation 15(6), 1191–1253 (2003)
14. Popp, T., Mangard, S.: Masked dual-rail pre-charge logic: DPA-resistance without routing constraints. In: Rao, J., Sunar, B. (eds.) Cryptographic Hardware and Embedded Systems – CHES 2005, Proceedings. Lecture Notes in Computer Science, vol. 3659, pp. 172–186. Springer Berlin / Heidelberg (2005), `http://dx.doi.org/10.1007/11545262_13`
15. Prouff, E.: DPA attacks and S-boxes. Fast Software Encryption 3557, 424–441 (2005)
16. Prouff, E., Rivain, M., Bevan, R.: Statistical analysis of second order differential power analysis. Computers, IEEE Transactions on 58(6), 799 –811 (june 2009)
17. Shiga, M., Yokota, Y.: An optimal entropy estimator for discrete random variables. In: Proceedings of the International Joint Conference on Neural Networks. pp. 1280–1285. IEEE International Joint Conference on Neural Networks (IJCNN), IEEE, New York (2005)
18. 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. ICISC 2008 5461, 253–267 (2009)
19. Standaert, F.X., Malkin, T.G., Yung, M.: A unified framework for the analysis of side-channel key recovery attacks. In: EUROCRYPT '09: Proceedings of the 28th Annual International Conference on Advances in Cryptology. pp. 443–461. Springer-Verlag, Berlin, Heidelberg (2009)
20. Treves, A., Panzeri, S.: The upward bias in measures on information derived from limited data samples. Neural Computation 7(2), 399–407 (1995)
21. Veyrat-Charvillon, N., Standaert, F.X.: Mutual information analysis: How, when and why? In: Clavier, C., Gaj, K. (eds.) Cryptographic Hardware and Embedded Systems – CHES 2009, Proceedings. Lecture Notes in Computer Science, vol. 5747, pp. 429–443 (2009)