# One for All - All for One: Unifying Standard DPA Attacks*

Stefan Mangard[1], Elisabeth Oswald[2], François-Xavier Standaert[3]

[1] Infineon Technologies AG, Security Innovation, Germany.
[2] University of Bristol, Department of Computer Science, UK.
[3] Université catholique de Louvain, Crypto Group, Belgium.

**Abstract.** In this paper, we examine the relationship between and the efficiency of different approaches to standard (univariate) DPA attacks. We first show that, when feeded with the same assumptions about the target device (*i.e.* with the same leakage model), the most popular approaches such as using a distance-of-means test, correlation analysis, and Bayes attacks are essentially equivalent in this setting. Differences observed in practice are not due to differences in the statistical tests but due to statistical artifacts. Then, we establish a link between the correlation coefficient and the conditional entropy in side-channel attacks. In a first-order attack scenario, this relationship allows linking currently used metrics to evaluate standard DPA attacks (such as the number of power traces needed to perform a key recovery) with an information theoretic metric (the mutual information). Our results show that in the practical scenario defined formally in this paper, both measures are equally suitable to compare devices in respect to their susceptibility to DPA attacks. Together with observations regarding key and algorithm independence we consequently extend theoretical strategies for the sound evaluation of leaking devices towards the practice of side-channel attacks.

## 1 Introduction

Just over a decade ago the publication of *Differential Power Analysis* (DPA) attacks [10] excited the cryptographic community because of their unexpected simplicity and effectiveness in practical settings. Their introduction sparked off research in different directions: attacks *vs.* countermeasures and theory *vs.* practice. Obviously, practice informs the development of theory, and countermeasures can neither be formalized nor tested without a sound understanding of attacks.

Important steps towards the theoretical analysis of countermeasures (based on formal descriptions of adversaries) have been made in [4, 17]. Dedicated work on theoretical models for side-channel attacks has been published in [13], and then further developed in [18]. The beauty of these theoretical approaches is that they formalize a large number of attacks. The drawback is they have not bridged the gap to practical works such as [10]. In particular, the work in [18] establishes the conditional entropy as a measure of information leakage and discusses several

* This paper is a postprint of a paper submitted to and accepted for publication in IET Information Security and is subject to Institution of Engineering and Technology Copyright. The copy of record is available at IET Digital Library.

properties of it. But while these metrics are expected to serve as a comparison basis for any device, their connection to the efficiency of standard attacks was let as an open problem. Hence it is not clear if the evaluation of side-channel attacks requires an information theoretic approach anyway or if simpler evaluation tools (DPA, typically) could be exploited in certain meaningful scenarios. In parallel, research such as [2, 9], which is discussed in more practically orientated communities such as CHES (Cryptographic Hardware and Embedded Systems), has developed an interest in finding the "best" way to conduct DPA attacks. In other words, given the multitude of approaches to DPA attacks (*e.g.* correlation attacks, distance-of-means attacks, template attacks, *etc.* ), which one requires the least number of leakage traces to break a given algorithm?[1]

**Our contributions.** In this article we tackle this question of what is the "best" standard DPA attack. For this purpose, we first provide concise definitions for attacks and *discuss to which extent DPA attacks are key and algorithm independent*. Our definitions capture a large class of DPA attacks whilst being specific enough to allow us to make concrete statements later on. This is an important contribution towards putting such attacks on a sound theoretical basis.

Second, we show that for standard univariate DPA attacks (*i.e.* attacks based on assumptions such as made in [2, 10], precise definitions are given in Section 2) the *most popular methods are in fact equally efficient*. Specifically, we show that these different distinguishers mainly optimize the same criteria. Hence, differences observed in actual experiments are only due to statistical artifacts (*i.e.* imprecise estimations in case of too low numbers of leakages).

Third and under certain reasonable physical assumptions, we relate the correlation coefficient to the concept of conditional entropy (or mutual information), which has been established as a theoretical measure for side-channel leakage in [18]. This relationship has the important practical implication that *the leakage of a device* can sometimes be *directly related to the efficiency of standard DPA attacks* mounted against this device. Linking this with our discussion about key and algorithm independence, it turns out that the leakage can be measured independent of algorithms and related to the efficiency of DPA attacks.

Summarizing, our research solves a long-standing discussion about the efficiency of different types of DPA attacks. We show that in a first-order scenario and when provided with the same leakage models, the most popular attack methods essentially require the same number of leakage traces to extract keys in practice. Furthermore, we show that in this setting, the conditional entropy as a measure for the leakage of a device can be related to the efficiency of standard DPA attacks. Hence, by applying one standard DPA method, designers are even able to quantify the leakage of a device using the theoretical measure introduced in [18]. These results consequently establish a simple link between theory and practice, for a restricted but important class of side-channel attacks.

[1] Another possible definition of "best" could be: which attack requires the least precise assumptions about the leakage behaviour of the devices they target? In this paper we look at the number of leakages which is important for practical adversaries.

**Preliminary remarks.** While we believe that the following results highlight useful intuitions for the evaluation of leaking devices, it is essential to underline that they are also limited, in order to avoid misinterpretations. Note that these limitations are inherent to our goal of stating general results for physical attacks, which requires the definition of a well-defined context in which our claims hold.

First, this paper considers statistical distinguishers that all exploit the same leakage model. But in practice, the choice of a distinguisher can go together with the choice of a leakage model. For example, correlation attacks often exploit approximated models, based on the Hamming weights of the manipulated data. By contrast, Bayesian attacks are usually applied after a careful profiling of the target device (*i.e.* the construction of templates). This does not reduce the impact of our conclusions but suggests another reading. Namely, in standard first-order DPA attacks, the selection of a leakage model has more impact on the success rate than the selection of the statistical tool to exploit this model.

Second, the equivalence that we claim for different distinguishers is asymptotic. That is, we show that once all the terms in a statistical test used in a DPA attack are properly estimated, then the only key-dependent terms are identical for the equivalent distinguishers. Since side-channel attacks use statistical tests to discriminate different key candidates, and this discrimination usually happens before the perfect estimation of all the terms in the statistics, this means that different distinguishers may indeed perform differently, up to a certain extent. But since these differences are only due to statistical artifacts, there is not much to be learned from them. Also, this difference vanishes when the number of queries required to perform a successful key recovery increases.

Third, our results rely on a number of physical assumptions that may not be perfectly observed in practice. Most importantly, the equivalence between correlation attacks and template attacks only holds in case of so-called Gaussian leakages. This nicely captures the context of an attack against an unprotected CMOS device. But as demonstrated in [21], the equivalence between different distinguishers does not generalize to attacks against implementations protected with masking. More generally, the observations in this paper do not prevent the information theoretic metric of [18] to remain necessary in more advanced scenarios. For example, when exploiting multivariate statistics against protected devices, the generality of an information theoretic metric remains useful to capture any possible leakage dependency (in contrast with the correlation coefficient).

**Organization of this paper.** We define our notations, the attacks we consider and the necessary assumptions for our analysis to hold in Sect. 2. Our contributions are organised in three main sections. Sect. 3 discusses key and algorithm independence issues in side-channel attacks. In Sect. 4 we show that (two of) the most popular methods for DPA attacks are equally efficient. Sect. 5 investigates the relationship between correlation and conditional entropy in a practical implementation setting. Finally, Sect. 6 welds the different pieces of this work into a whole and discusses implications and future research directions. There are several appendices to this paper that provide definitions, more details for proofs, and cover an additional distinguisher (the distance-of-means test).

## 2 Background

### 2.1 Notations

In this work, we use an $n$-bit block cipher as an example of a cryptographic algorithm that is implemented in a device. It aims to illustrate concepts and attacks, but our analyzes and definitions hold also for other cryptographic primitives. Let $x$ be a plaintext selected at random from a set $\mathcal{X}$: $x \xleftarrow{R} \mathcal{X}$ and let $k$ be a key selected at random from a set $\mathcal{K}$: $k \xleftarrow{R} \mathcal{K}$ such that $\mathcal{X} = \mathcal{K} = \{0,1\}^n$. For such $x$ and $k$, let $\mathsf{E}_k(x)$ be the encryption of the plaintext $x$ under a key $k$. In classical cryptanalytic attacks, an adversary is able to query the block cipher (or any algorithm) in order to obtain pairs of plaintexts and ciphertexts $[x_i, \mathsf{E}_k(x_i)]$. In side-channel attacks, the adversary is additionally provided with the output of a leakage function $\mathsf{L}$. Following [13], this function is used to represent the actual output of a leaking device's measurement setup. Let now $\mathbf{x}_q = [x_1, x_2, \ldots, x_q]$ be a vector containing a sequence of $q$ input plaintexts to a target implementation. The measurements resulting from the observation of the encryption of these $q$ plaintexts are stored in a leakage vector denoted as $\mathbf{l}_q = [l_1, l_2, \ldots, l_q]$. Each element $l_i$ corresponds to the encryption of an input $x_i$ under key $k$. These elements are often referred to as leakage traces and contain many points in practice.

### 2.2 Definition of the attacks

Side-channel attacks are usually based on a divide-and-conquer strategy in which different parts of a secret key are recovered separately. In general, the attacks define a function $\gamma : \mathcal{K} \rightarrow \mathcal{S}$ which maps each key $k$ to a subkey $s = \gamma(k)$, such that $|\mathcal{S}| \ll |\mathcal{K}|$. Note that [16] uses the term "subkeys" where [18] uses the term "key classes". We use "subkeys" in the remainder of this article. In [18], a side-channel key recovery adversary is defined as an algorithm with a certain time, data and memory complexity that can query a target implementation with inputs $\mathbf{x}_q$ and exploit answers containing both the ciphertexts $\mathsf{E}_k(\mathbf{x}_q)$ and the leakages $\mathbf{l}_q$ corresponding to their encryption. A more practical definition is also proposed in which the adversary is described as a statistical procedure that compares key-dependent predictions of the leakages with actual measurements. The success of a side-channel attack essentially depends on the extent to which the best prediction actually corresponds to the leakage of the correct subkey. As a matter of fact, there are many different ways to predict the leakages and compare the predictions with physical measurements. In this paper, we focus on a specific category of attacks that is widely used in practical settings.

Specifically, we will focus on standard DPA attacks that are intuitively pictured in Figure 1 using the key addition and S-box layers of a block cipher as concrete example. These DPA attacks follow three main steps:

1. For different plaintexts $x_i$ and subkey candidates $s^*$, the adversary predicts some intermediate values in the target implementation. For example, one could predict S-box outputs $z_i$ in Figure 1 and get values $v_i^{s^*} = \mathsf{S}(x_i \oplus s^*)$.

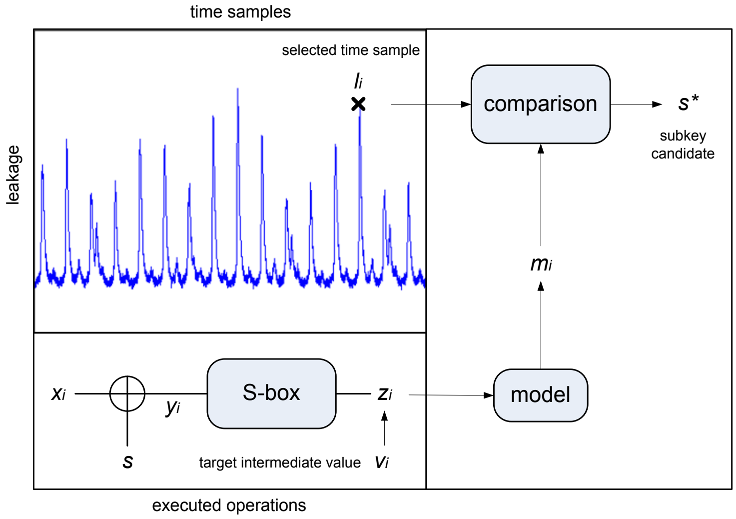


**Fig. 1.** Notations and illustrative univariate DPA attack.

2. For each of these predicted values, the adversary models the leakages. For example, if the target block cipher is implemented in a CMOS-based 8-bit microcontroller, the model can be the Hamming weight ($HW$) of the predicted values. One then obtains modeled leakages $m_i^{s^*} = HW(v_i^{s^*})$.
3. For each subkey candidate $s^*$, the adversary compares the modeled leakages with actual measurements, produced with the same plaintexts $x_i$ and a secret subkey $s$. In first-order DPA attacks, each $m_i^{s^*}$ is compared with a single point in the traces. This comparison is independent of all other points. Consequently, these attacks are referred to as *univariate* attacks. In practical attacks, this comparison is applied to many points in the leakage traces and the subkey candidate that performs best is selected by the adversary.

As detailed in [12], different statistical tests can be considered to perform the comparison and our goal is to analyze them. We will investigate two frequently considered ones in detail, namely Pearson's correlation coefficient [2] and Bayes (used, *e.g.* in template attacks [5]). We also deal with a third one (namely the distance-of-means test) in Appendix E (that should be read after Section 4). In a so-called correlation attack, the adversary selects the subkey candidate as:

$$\tilde{s} = \underset{s^*}{\mathrm{argmax}}\ \hat{\rho}(\mathbf{l}_q, \mathbf{m}_q^{s^*}), \tag{1}$$

where $\hat{\rho}$ denotes Pearson's sample correlation coefficient. In an attack using Bayes, the adversary directly exploits an approximated probability density function for the leakages and selects the subkey candidate with maximum likelihood:

$$\tilde{s} = \underset{s^*}{\mathrm{argmax}} \prod_{i=1}^{q} \hat{\Pr}[l_i | m_i^{s^*}] \tag{2}$$

Note that in practice, comparing different statistical tests requires to provide them with the same leakage samples. The points which lead to good attacks are related to the intermediate values selected in step 1 of the analysis. Among these points, there is typically one that stands out. Different methods can be used to identify it. In this paper, we only assume that it is somehow selected with an arbitrary method of choice (*i.e.* we do *not* assume that it is known a priori). Because we show in Section 4 that all the statistical tests under investigation optimize the same criteria, our results hold irrespective of the selection of the point: any sample that is good (or bad) for one test will be equally good (or bad) for the other tests. Finally and as far as adversarial capabilities are concerned, we consider known and uniformly generated random inputs for our attacks. In this context, we work with adversaries exploiting a known message leakage model.

**Definition 1.** *A side-channel adversary exploits a known message leakage model if he can predict the leakage generated by any input/output of his target device.*

This definition is needed to avoid (unusual) situations in which, *e.g.* an adversary would have built templates only for a few plaintexts and key bytes.

### 2.3 Characteristics of the leakage function

**Definition 2.** *A leakage sample $l_i$ is said to have additive noise if this sample can be written as the sum of a deterministic part $d_i$ and a random part $r_i$. In addition, the random part is independent of the deterministic part and it is identically distributed for all messages and subkeys.*

Additive noise is a standard assumption in side-channel attacks. For example, it was used in [1], [4]. Another common assumption is related to the existence of symmetry properties for certain pairs of leakages and subkeys. It is denoted as the EIS property and we use the definition provided in [16].

**Definition 3.** *Let $\mathcal{A}$ be an arbitrary set and let $\phi : \mathcal{X} \times \mathcal{S} \to \mathcal{A}$ be a mapping for which the images $\phi(\mathcal{X} \times s) \subset \mathcal{A}$ are equal for all subkeys $s \in \mathcal{S}$. We say that a leakage sample has property of Equal Images under different Subkeys (EIS) if it can be written as $l_i = \delta(x_i, s) + r_i$ with $\delta = \overline{\delta} \circ \phi$ for a suitable mapping $\overline{\delta} : \mathcal{A} \to \mathcal{L}$, i.e. $\delta(x_i, s)$ can be written as a function of $\phi(x_i, s)$.*

We illustrate the EIS property using our block cipher example. Consider an implementation of a key addition mapping $\phi(x_i, s) = x_i \oplus s$ that leaks the Hamming weight. The leakage of this intermediate value then has the EIS property because all pairs $(x_i, s)$ map to the same set of images $x_i \oplus s$.

Note that the EIS property can be similarly defined for a leakage model $m_i^{s^*}$. In fact, if one assumes EIS for the target leakage samples, then the underlying model also has the EIS property. Combined with the additive noise assumption, this further implies that the models have equal distributions (EDS) for all subkeys. As a last property for the leakages, we fix the probability distribution of the random part to be a normal distribution. In practice, these properties are not supposed to be perfectly respected but to hold to a sufficient degree.

**Definition 4.** *A leakage sample $l_i$ is Gaussian if it has additive noise and its random part follows a normal distribution with mean zero and variance $\sigma_R^2$.*

We note that this last assumption is very frequently considered and verified in the practice of side-channel attacks [12]. For example, it was already used in the seminal papers on template attacks [5] and stochastic models [16].

### 2.4 Attack scenario

Taking advantage of the previous definitions, we now define our attack scenario.

**Definition 5.** *A standard univariate DPA attack (short: standard DPA attack) is an attack that follows the 3-step procedure outlined in Sect. 2.2 under the following conditions. The adversary exploits uniform inputs with a known message leakage model, the leakage samples are Gaussian, models and leakages have the EIS property, and all subkey candidates $s$ have equal a-priori probabilities.*

The previous definitions capture the conditions observed in many practical DPA attacks, in particular [2, 5, 10]. In the next section, we discuss to which extent these attacks are independent of a cryptographic algorithm and its key.

## 3 Key and algorithm independence issues

It is a general intuition that up to a certain extent, DPA attacks are key and algorithm independent. Again, we use a block cipher as concrete example. One expects that if an attack against this cipher implemented on a given platform succeeds, then an attack against another block cipher implemented on the same platform should succeed as well. It is also frequently assumed that block ciphers have no weak or strong keys with respect to DPA attacks. In this section, we discuss the conditions upon which these intuitions hold. In particular, we show that standard DPA attacks imply a certain level of key and algorithm independence.

### 3.1 Key independence

Essentially, the EIS property combined with the additive noise assumption implies that all keys potentially lead to the same leakages. If inputs are chosen uniformly at random, this implies that standard DPA attacks apply equally to all subkeys values and hence are independent of them. Theorem 1 formalizes this observation, using the definition of success rate recalled in Appendix A.

**Theorem 1.** *On average over its input vectors $\boldsymbol{x}_q$ and given a target subkey defined by a function $\gamma : \mathcal{K} \rightarrow \mathcal{S}$, the success rate of a standard DPA attack against a cryptographic implementation is the same for all subkey values $s \in \mathcal{S}$.*

A proof sketch is given in Appendix B. Note that the assumption about uniformly distributed plaintexts is important - it is easy to find an example of attack with non uniform plaintexts leading to key dependencies (see Appendix C). When applicable, Theorem 1 indicates that all key values are equally difficult to recover. A similar claim holds for the information theoretic metric of [18], namely that all the lines in the entropy matrix (recalled in Appendix A) are identical, up to a permutation (*i.e.* the so-called weak template attack context).

### 3.2 Algorithm independence

**Theorem 2.** *If for each input plaintext $x_i$, there is a bijective relation between the target subkey and the intermediate value exploited in a single-query standard DPA attack against a cryptographic implementation, then the success rate of this attack averaged over its inputs is independent of the algorithm it targets.*

The proof of this theorem trivially derives from the fact that the distribution of a random variable is only permuted if this variable goes through a bijection. Hence, it does not change the cardinalities of the subkey candidates in a side-channel attack. Taking the example of Figure 1, we can imagine an attack in which the S-box outputs $z_i$ are the target intermediate values. Intuitively, a first leakage sample $l_1$ corresponding to a plaintext $x_1$ reduces the set of intermediate values $\mathcal{Z}$ to the dark grey subset in Figure 2. A subset of the same cardinality can be defined in the set of subkeys $\mathcal{S}$ since for a given plaintext $x_1$, going through the inverse S-box and XORing with $x_1$ is a bijection. We note that this statement holds even if only a part of the leakage produced by the intermediate values is modeled by the adversary (*e.g.* if only a few bits of $z_i$ are predicted).

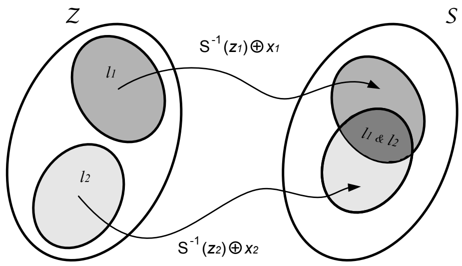


**Fig. 2.** Combination of leakages and algorithm independence.

More interesting is the observation that Theorem 2 does not hold anymore for two leakage samples. Indeed, if a second plaintext $x_2$ is used to generate a second sample $l_2$, then the correct key $s$ also has to be part of the light grey subset in Figure 2. The probability $\Pr[s|l_1, l_2]$ consequently depends on how the intersection between the two subsets in $\mathcal{S}$ is distributed. As already pointed out, *e.g.* in [15, 19], this depends on the algebraic structure of the S-box. Note that in practice, this second sample could correspond both to a new plaintext $x_2$ or to the leakage of another intermediate value, *e.g.* an attack targeting the inputs and outputs of the S-box in Figure 1 would suffer from the same issue. However, this algorithm dependency vanishes when the number of queries increases. That is, the intersection in the set $\mathcal{S}$ eventually only contains the correct subkey candidate $s$ (if a sound leakage model in the sense of [18] is used).

In summary, if there is a bijective relation between subkeys and target values, side-channel attacks are independent of the block cipher on which they operate, when one or all leakages are used. Intermediate amounts of queries lead to algorithm-dependencies[2]. This fact naturally derives from the asymptotic behavior of the attacks that will be considered in detail in the next section. Combined with the result of the previous section, it means that the evaluation of standard DPA attacks can indeed be performed key and algorithm independent.

## 4 Relation between statistical methods for standard DPA

In this section, we formalize the similar efficiency of the correlation and template attacks in a univariate DPA attack scenario. In practice there are several popular methods. The covariance was first mentioned in Chari *et al.* [3] and later on re-introduced in [2], and Bayes was introduced by Chari *et al.* in [5]. We show that in a standard DPA attack, these methods are equally efficient in practice.

**Definition 6.** *The efficiency of a side-channel attack* $\mathsf{A}$ *to reach a success rate* $sr$ *is the minimum average number of queries* $q_{sr}^{\mathsf{A}}$ *such that the success rate (defined in Appendix A.1) of this attack reaches* $sr$*, i.e.* $\boldsymbol{Succ}_{\mathsf{A}}^{sc\text{-}kr}(q_{sr}^{\mathsf{A}}) \geq sr$*.*

**Theorem 3.** *In a standard DPA attack where the leakage samples have noise variance* $\sigma_R^2$*, the statistical closeness between correlation and Bayesian attacks* $|q_{sr}^{\mathsf{corr}} - q_{sr}^{\mathsf{bayes}}|$ *is a monotonously decreasing function of* $\sigma_R^2$*.*

*Proof sketch.* Without restricting generality, we assume that an attacker subtracts the mean $\hat{\mathbf{E}}(\mathbf{L}_q)$ from the leakages and models at the beginning of an attack. An attacker using the correlation coefficient then selects the key according to (1), which can be simplified and written as described in Appendix D. Because the attacker first subtracts the mean $\hat{\mathbf{E}}(\mathbf{L}_q)$ from all leakages, the term $\hat{\mathbf{E}}(\mathbf{L}_q) \cdot \hat{\mathbf{E}}(\mathbf{M}_q^{s^*})$ equals zero, and the correlation is given in (3):

$$\tilde{s} = \underset{s^*}{\operatorname{argmax}} \frac{\hat{\mathbf{E}}(\mathbf{L}_q \cdot \mathbf{M}_q^{s^*})}{\hat{\mathbf{E}}((\mathbf{M}_q^{s^*})^2) - (\hat{\mathbf{E}}(\mathbf{M}_q^{s^*}))^2} \tag{3}$$

Similarly, an attacker using Bayes' method selects the key according to (2), which can be simplified and written as detailed in Appendix D:

[2] In case of surjective S-boxes, even single queries lead to algorithm dependencies. We let the careful investigation of this context as a scope for further research.

$$\tilde{s} = \underset{s^*}{\text{argmax}} \; \frac{\hat{\mathbf{E}}(\mathbf{L}_q \cdot \mathbf{M}_q^{s^*})}{\hat{\mathbf{E}}\left((\mathbf{M}_q^{s^*})^2\right)} \tag{4}$$

Theorem 3 essentially results from the three following observations:

1. Due to the EIS property and for any given number of leakage traces $q$, the distribution of the model $\mathbf{M}_q^{s^*}$ is equal for all subkey candidates $s^*$. It directly implies that the sampling distribution of the denominator terms in equations (3) and (4) is independent of the subkeys in this context.
2. For any model used by the adversary, the variance of the sampling distribution of the denominator terms in equations (3) and (4) is a monotonously decreasing function of $q$. In other words, the more leakage traces are used in an attack, the better these denominator terms are estimated.
3. Eventually, the product term in the numerators of equations (3) and (4) has a different sampling distribution for different subkey candidates. Hence, only this product term allows discriminating the correct subkey. Furthermore, the sampling distribution of the variance of this product term is a monotonously increasing function of the noise variance $\sigma_R^2$. It implies that the more noise there is in the leakages, the higher is the variance of the sampling distribution of this product term and the more difficult it is to estimate.

Putting these observations together leads to Theorem 3 as follows. First, increasing $\sigma_R^2$ leads to a higher variance for the sampling distribution of the product term in the nominators of equations (3) and (4). It implies a drop in the success rate of the attacks. In order to achieve the same success rate for the increased noise level as for the original one, it is necessary to increase $q$. But by increasing both $\sigma_R^2$ and $q$, one can only improve the estimation of the denominator terms in equations (3) and (4). Indeed, the variance of the sampling distribution of these denominator terms does not depend on $\sigma_R^2$ and decreases with $q$. Hence, the difference between these denominator terms for different subkey candidates is becoming smaller in this case. Since these denominators are also the only place where the correlation and Bayes attacks differ, it implies that these statistical tests are becoming more similar. Summarising, the numerator becomes more decisive the higher $\sigma_R^2$ and $q$, because the denominator becomes subkey independent. As the numerator is the same for Bayes and correlation, the absolute difference $|q_{sr}^{\mathsf{corr}} - q_{sr}^{\mathsf{bayes}}|$ monotonously decreases as a function of $\sigma_R^2$. □

The previous theorem states that the efficiency of Bayesian and correlation attacks gets close as soon as the number of queries required to perform a successful attack is "large enough" and this number depends on the variance of the product term in equations (3) and (4). Of course, in practice the important question is to determine how this requirement fits to practical scenarios and to quantify it. In the following, we first discuss this problem theoretically, by introducing an additional empirical assumption. Then, we provide simulated and actual experiments that both validate our claims.

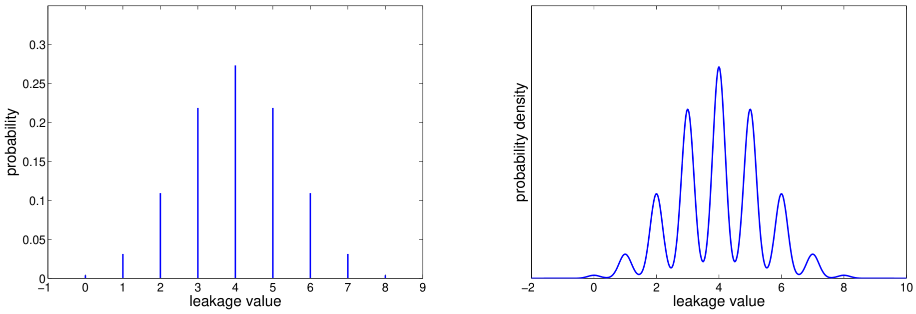


**Fig. 3.** Examples of overall leakage probability distribution.

**Empirical assumption.** *We assume that the distributions of the leakage samples' deterministic part $d_i$ and the models $m_i^{s^*}$ in a standard DPA attack are close to Gaussian, with respective variances $\sigma_D^2$ and $\sigma_M^2$.*

We first note that this empirical assumption considers the leakages' deterministic part and the models, by opposition to Definition 4 that only considers the leakages' random part. Intuitively, it can be simply explained using the example of Figure 3. The left part of the figure shows the binomial distribution that corresponds to a Hamming weight leakage function with noise variance $\sigma_R^2 = 0$. The right part of the figure shows the same leakage function with some additive noise. Clearly, none of these distributions is strictly Gaussian. On the other hand, such Gaussian mixtures constructed from a binomial distribution reasonably fulfill our empirical requirement. And in fact, this observation holds for a lot of leakage models that are considered in practice (in particular, all the models that capture a weighted sum of certain bits in an implementation). Furthermore, this empirical assumption holds the better the higher $\sigma_R^2$. Hence, it appears as a reasonable starting point to discuss the quantitative aspects of Theorem 3. It directly leads to the following corollary:

**Corollary 1.** *In a standard DPA attack where the leakages' deterministic part and the models are close to Gaussian, the variance of the sampling distribution of the terms $\hat{\boldsymbol{E}}(\boldsymbol{L}_q \cdot \boldsymbol{M}_q^{s^*})$ and $(\hat{\boldsymbol{E}}(\boldsymbol{M}_q^{s^*}))^2$ in equations (3) and (4) equals $(1+r^2) \cdot (\sigma_D^2 + \sigma_R^2) \cdot \sigma_M^2/q$ and $2\sigma_M^4/q$, respectively, where the coefficient $r = \rho(\boldsymbol{M}_q^{s^*}, \boldsymbol{L}_q)$ denotes the correlation between the model and the leakages.*

These variances can be directly obtained from [6]. They allow putting forward an important shortcut to avoid in the application of Theorem 3. Namely, the corollary shows that the number of queries required to perform a successful attack that can be considered as "large enough" for Theorem 3 to hold actually depends on several parameters, since the condition to be respected is:

$$(1 + r^2) \cdot (\sigma_D^2 + \sigma_R^2) \cdot \sigma_M^2/q > 2\sigma_M^4/q$$

For example, the more the model and leakages are correlated, the smaller this number will be. In other words, the corollary shows that Theorem 3 strictly holds for a given device and model. But it does not indicate any improved closeness between the distinguishers when different devices and models are considered.

We finally emphasize that the empirical assumption in this section is not necessary for Theorem 3 to hold but it is useful to analyze it quantitatively. The next two sections confirm the previous claims empirically.

### 4.1 Validation of the results using simulated experiments

We first simulated attacks against the exemplary implementation of Figure 1 with the AES S-box, assuming a Hamming weight leakage function and model for which we have $\sigma_M^2 = 2$. Figure 4 illustrates the success rates of the correlation and Bayesian attacks in a standard DPA scenario, with respective noise standard deviations $\sigma_R = 1$ and $\sigma_R = 5$. It clearly illustrates that correlation and Bayes attacks have a very similar efficiency in this context, even with a low number of leakages. This can be explained by the perfect matching of (*i.e.* high correlation between) the actual leakages and the adversary's model.

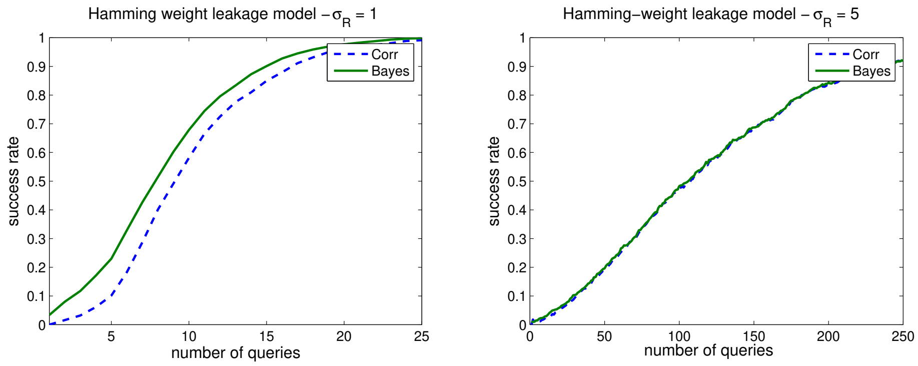


**Fig. 4.** Success rate of the correlation and Bayesian attacks - simulated attacks.

### 4.2 Validation of the results using real experiments

In order to further validate Theorem 3, this section shows that our results also hold in practice, for a range of different devices. For this purpose, we have selected: an 8-bit microcontroller such as found in typical low-end smart cards, a 32-bit microprocessor such as found in more expensive smart cards and embedded devices and finally a 128-bit ASIC coprocessor dedicated to the computation of the AES. The different attacks we performed exactly follow the standard DPA procedure described in this paper and exploit setups such as described, *e.g.* in [12]. Figure 5 illustrates the success rates of the correlation and Bayesian attacks for our three different devices. We again observe that by only adding Gaussian noise to the measurements, the two attacks get closer (as in the two upper parts of the figure). Also, we see that changing a device implies a different condition for the "large enough" number of leakages. For example, the Hamming weight model does not perfectly capture the leakage variations of our 32-bit microprocessor, which explains a larger difference between the two statistical tests in this case.

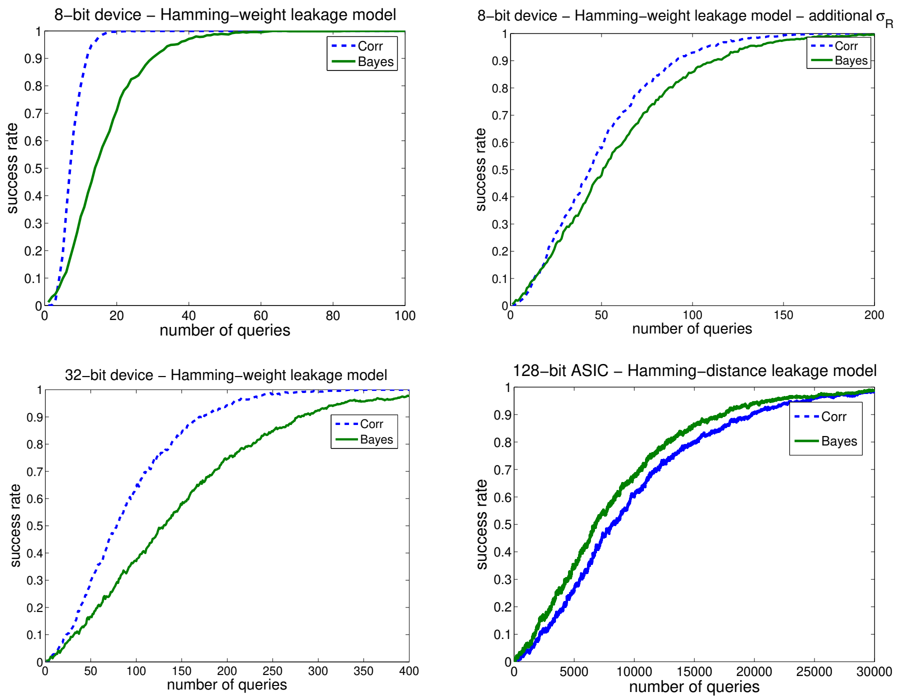


**Fig. 5.** Success rate of the correlation and Bayesian attacks - real measurements.

Summarizing, the results in this section show that from an attacker's point of view, the most important question when performing a standard DPA attack is the selection of a good leakage model. But once this model is given, using a correlation coefficient, a Bayesian distinguisher (or a distance-of-means test, see Appendix E) in parallel is redundant to a certain extent and will not give much additional insights about the security of a leaking device. One statistical test will essentially do the job. These results confirm empirical evaluations such as [20] in which different univariate side-channel attacks are experimented. We note that the analysis of circuits protected with dual-rail logic styles such as [22, 23] would lead to similar practical observations. By contrast, moving to masked implementations (in which the Gaussian leakages are turned into mixtures of Gaussians) would lead to significantly different conclusions, as detailed in [21].

## 5 Relation between correlation and mutual information

In [18], the mutual information is suggested as a metric to compare different leaking implementations. The intuition is that if two devices $A$ and $B$ run an algorithm $\mathsf{E}_k(x)$ and the same subkey is targeted by an adversary such that $\mathrm{H}[S|\mathbf{L}_1]$ respectively equals $h_A, h_B$ for these devices and $h_A > h_B$, then the success rate of a Bayesian adversary in recovering this subkey should be higher for device $B$. It results from the observation that side-channel attacks generally require several queries to be successful, which allows the intuition "more entropy implies less success rate" to be experimented in practice. But although empirically confirmed by different applications, this relation between information theoretic and security metrics was only shown for a Bayesian adversary recovering a single-bit subkey in previous work. In this section, we investigate how the mutual information relates to the correlation coefficient used in standard DPA attacks. We show how this relation confirms the proofs of [18] in another specific realistic implementation scenario. For this purpose, we use the following theorem.

**Theorem 4.** *The mutual information between two normally distributed random variables $X, Y$, with means $\mu_X, \mu_Y$ and variances $\sigma_X^2, \sigma_Y^2$ can be expressed as:*

$$\mathrm{I}(X;Y) = -\frac{1}{2} \cdot \log_2 \left(1 - \rho(X,Y)^2\right) \tag{5}$$

The proof of this theorem is in Appendix F. We now discuss the extent to which it applies to standard DPA attacks. For this purpose, it is worth recalling the four main variables that we consider, namely the subkeys $S$, intermediate values $\mathbf{V}_q$, models $\mathbf{M}_q$ and leakages $\mathbf{L}_q$. Overall, the quantity we are interested in to evaluate a leaking device is $\mathrm{H}[S|\mathbf{L}_q]$. A first observation is that, assuming a bijective relation between $S$ and $\mathbf{V}_q$, we have $\mathrm{H}[S|\mathbf{L}_1] = \mathrm{H}[\mathbf{V}_1|\mathbf{L}_1]$, as discussed in Section 3.2. Hence, we can identically evaluate the leaking device with these two quantities. Let us now assume that the adversary's model exactly corresponds to the leakages' deterministic part. Then, a second observation is that, because of the additive noise property, we have $\rho(\mathbf{V}_q, \mathbf{L}_q) = \rho(\mathbf{V}_q, \mathbf{M}_q) \cdot \rho(\mathbf{M}_q, \mathbf{L}_q)$. Similarly, we have for the conditional entropy:

$$\mathrm{H}[\mathbf{V}_1|\mathbf{L}_1] = \mathrm{H}[\mathbf{V}_1|\mathbf{M}_1] + \mathrm{H}[\mathbf{M}_1|\mathbf{L}_1] \tag{6}$$

(see Appendix G). Eventually, if we additionally assume that the leakage's deterministic part and the models have a close to Gaussian distribution (*i.e.* our empirical assumption in the previous section), we can directly use Theorem 4 and approximate $\mathrm{I}(\mathbf{M}_1;\mathbf{L}_1)$ with $-\frac{1}{2} \cdot \log_2 \left(1 - \rho(\mathbf{M}_q, \mathbf{L}_q)^2\right)$ and $\mathrm{H}[\mathbf{V}_1|\mathbf{M}_1]$ (that does not depend on the actual measurements $\mathbf{L}_1$) separately. The additive law of Equation 6 can then be used to compute $\mathrm{H}[\mathbf{V}_1|\mathbf{L}_1]$.

In practice, the quality of this approximation depends on how well the "close to Gaussian" assumption is respected. In order to confirm this assumption, we again simulated experiments with the example of Figure 1, assuming a Hamming weight leakage function. We considered both a Hamming weight and single-bit leakage model. The results are illustrated in Figure 6. In both cases, we see that:

– H$[\mathbf{V}_1|\mathbf{M}_1]$ is fixed and does not depend on the measurements $\mathbf{L}_1$,
– The mutual information $\mathrm{I}(\mathbf{M}_1; \mathbf{L}_1)$ varies from 0 and $\approx 2.5$ (*resp.* $\approx 0.1$) for the Hamming weight (*resp.* single-bit) leakage models
– The estimation of this mutual information with the correlation coefficient is good as long as the value of this coefficient is not too close to one (*e.g.* when the noise standard deviation is lower than 0.25 in the left part of Figure 6).

As already mentioned, none of the investigated models is Gaussian. But the Hamming weight leakage model follows a binomial distribution that reasonably approximates the normal distribution. And for the single-bit leakage model, the noise generated by the 7 other bits of the target intermediate values is sufficient for Theorem 4 to be observed in practice. In this respect, it is worth noting that the correlation never reaches one with the single-bit leakage model because even for low measurement noises, the algorithmic noise (made of 7 bits out of 8) is such that the maximum value for $\rho(\mathbf{M}_q, \mathbf{L}_q)$ equals $\sqrt{1/8} \approx 0.35$.

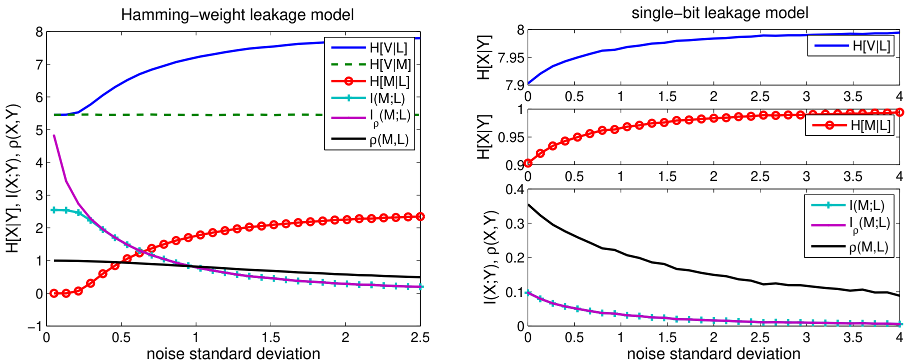


**Fig. 6.** Estimation of the conditional entropy with a correlation coefficient.

Summarizing, Theorem 4 states that in the context of standard DPA attacks, when the leakages' deterministic parts and models are "close to Gaussian", both an information theoretic metric (such as the conditional entropy proposed in [18]) and the correlation coefficient previously used in the side-channel literature measure the extent to which a key dependent model captures the actual leakage variations. We note that this result also applies in the case of unperfect leakage models. But as discussed in [24], the adversary will then underestimate the information leakages. Again, the good selection (or profiling) of a leakage model is the main element to allow a proper evaluation of its leakages. We finally mention that, as for Theorem 3, it is important to avoid shortcuts in the application of Theorem 4. In particular, the "close to Gaussian assumption" applies reasonably well to standard DPA attacks. But it does not apply in advanced scenarios, *e.g.* when countermeasures such as masking are considered. Because of its genericity, the conditional entropy remains necessary in such contexts.

## 6 Implications for practice and further research

Our results first imply that in a standard DPA scenario, the efficiency of attacks using the correlation coefficient and a Bayesian distinguisher is statistically close if both attacks use the same leakage model. An amusing consequence is that in this context, even Kocher's original single bit DPA is close to a single bit Bayesian attack that is usually assumed to be much more powerful. Consequently, our work provides a mathematical foundation for the observations made in the "DPA Contest" [8] and other empirical evaluations such as [20]. It explains that the small differences between the investigated distinguishers that are observed in practice are not due to the statistical tests but to statistical artifacts (which may nevertheless be significant in certain scenarios, *e.g.* if the number of queries to the target device is strictly limited). Further research could tackle the question of how these conclusions apply to distinguishers that we did not discuss, *e.g.* [9].

Our second conclusion is that the amount of information leaked by a cryptographic device measured with an information theoretic metric is connected to the correlation coefficient used in standard DPA attacks. Hence, under certain reasonable physical assumptions that we discuss in this paper, both metrics can be used as a figure of merit of the target devices with respect to these attacks. Connecting this result with the practical security analysis of [11], we can even relate these quantities to the security of the implementations (*i.e.* the number of traces required to recover the keys with high success rate). Recent work has shown that this conclusion does not generalize to higher-order attacks [21]. Further research could tackle more complex situations, combining implementations randomized with masking and advanced technologies in which leakage functions significantly deviate from traditional (Hamming weigh, distance, . . . ) models.

Third, this paper shows that up to a certain extent, the evaluation of a leaking device can be done independently of the algorithms and keys that are targeted in side-channel attacks. With this respect, further research could investigate situations where the relation between the intermediate value from which information is extracted and the subkey that is to be recovered is not bijective, situations where the inputs are not uniformly generated, because of chosen plaintext leakage models or adaptively selected plaintexts in the attacks, or situations in which side-channel attacks are combined with advanced cryptanalysis tools.

Overall, the results discussed in this article are an important step towards a sound mathematical investigation of side-channel attacks. They are also of practical importance for developers and evaluators of cryptographic devices because they show that in the important scenario of standard DPA attacks, all distinguishers are equally efficient. Hence, testing does not always require to investigate all distinguishers exhaustively: it is sometimes sound to use "one distinguisher for all". Last, we hope that our work will provide a germ for further research and raise more fundamental questions in the field of side-channel analysis.

**Acknowledgements.** Elisabeth Oswald is supported in part by the EPSRC grant EP/F039638/1. François-Xavier Standaert is an associate researcher of the Belgian Fund for Scientific Research (FNRS - F.R.S.).

## References


1. M.-L. Akkar, R. Bevan, P. Dischamp, D. Moyart, *Power Analysis, What Is Now Possible...*, in the proceedings of ASIACRYPT 2000, Lecture Notes in Computer Science, vol 1976, pp 489-502, Kyoto, Japan, December 2000.
2. E. Brier, C. Clavier, F. Olivier, *Correlation Power Analysis with a Leakage Model*, CHES 2004, LNCS, vol 3156, pp 16-29, Boston, MA, USA, August 2004.
3. S. Chari C.S. Jutla, J.R. Rao, P. Rohatgi, *A note regarding evaluation of AES candidates on smart-cards*, Second AES Candidate Conference, p. 133-147, 1999
4. S. Chari, C.S. Jutla, J.R. Rao, P. Rohatgi, *Towards Sound Approaches to Counteract Power Analysis Attacks*, in the proceedings of CRYPTO 1999, Lecture Notes in Computer Science, vol 1666, pp 398-412, Santa Barbara, CA, USA, August 1999.
5. S. Chari, J. Rao, P. Rohatgi, *Template Attacks*, CHES 2002, Lecture Notes in Computer Science, vol 2523, pp 13-28, CA, USA, August 2002.
6. C. Craig, *On the Frequency Function of xy*, Ann. Math. Statist., vol.7, p. 115, 1936
7. T. Cover, J. Thomas, *Elements of Information Theory*, Wiley Interscience, 2006.
8. DPA Contest, http://www.dpacontest.org
9. B. Gierlichs, L. Batina, P. Tuyls, B. Preneel, *Mutual Information Analysis*, CHES 2008, LNCS, vol 5154, pp 426-442, Washington DC, USA, 2008
10. P. C. Kocher, J. Jaffe, B. Jun, *Differential Power Analysis*, CRYPTO 1999, LNCS, vol 1666, pp 388-397, Santa Barbara, CA, USA, August 1999.
11. S. Mangard *Hardware Countermeasures against DPA – A Statistical Analysis of Their Effectiveness*, in the proceedings of CT-RSA 2004, Lecture Notes in Computer Science, vol 2964, pp 222-235, San Francisco, CA, USA,Feburary 2004
12. S. Mangard, E. Oswald, T. Popp, *Power Analysis Attacks*, Springer, 2007.
13. S. Micali, L. Reyzin, *Physically Observable Cryptography*, TCC 2004, LNCS, vol 2951, pp 278-296, Cambridge, MA, USA, February 2004.
14. S. M. Ross, *Introduction to Probability Theory and Statistics for Engineers and Scientists*, Second Edition, Academic Press, ISBN: 0-12-598472-3
15. E. Prouff, *DPA Attacks and S-Boxes*, in the proceedings of FSE 2005, Lecture Notes in Computer Science, vol 3557, pp 424-441, Paris, France, Fevruary 2005.
16. W. Schindler, K. Lemke, C. Paar, *A Stochastic Model for Differential Side-Channel Cryptanalysis*, in the proceedings of CHES 2005, Lecture Notes in Computer Science, vol 3659, pp 30-46, Edinburgh, Scotland, September 2005.
17. N. P. Smart, D. Page, and E. Oswald, *Randomised Representations*, IET Information Security, June 2008, Volume 2, Issue 2, p. 19-27
18. F.-X. Standaert, T.G. Malkin, M. Yung, *A Unified Framework for the Analysis of Side-Channel Key Recovery Attacks*, in the proceedings of Eurocrypt 2009, LNCS, vol 5479, pp 443-461, Cologne, Germany, April 2009, extended version available on the Cryptology ePrint Archive, Report 2006/139.
19. F.-X. Standaert, E. Peeters, C. Archambeau, J.-J. Quisquater, *Towards Security Limits in Side-Channel Attacks*, in the proceedings of CHES 2006, Lecture Notes in Computer Science, vol 4249, pp. 30–45, Yokohama, Japan, October 2006.
20. F.-X. Standaert, B. Gierlichs, I. Verbauwhede, *Partition vs. Comparison Side-Channel Distingsuishers: An Empirical Evaluation of Statistical Tests for Univariate Side-Channel Attacks*, in the proceedings of ICISC 2008, LNCS, vol 5461, pp 253-267, Seoul, Korea, December 2008.
21. F.-X. Standaert, N. Veyrat-Charvillon, E. Oswald, B. Gierlichs, M. Medwed, M. Kasper, S. Mangard, *The World is Not Enough: Another Look on Second-Order DPA*, in the proceedings of Asiacrypt 2010, Lecture Notes in Computer Science, vol 6477, pp 112-129, Singapore, December 2010.

22. K. Tiri, I. Verbauwhede, *Securing Encryption Algorithms against DPA at the Logic Level: Next Generation Smart Card Technology*, in the proceedings of CHES 2003, Lecture Notes in Computer Science, vol 2779, pp 125-136, Cologne; Germany, September 2003.
23. K. Tiri, I. Verbauwhede, *A Logic Level Design Methodology for a Secure DPA Resistant ASIC or FPGA Implementation*, in the proceedings of DATE 2004, pp 246-251, Paris, France, February 2004.
24. N. Veyrat-Charvillon, F.-X. Standaert, *Mutual Information Analysis: How, When and Why?*, in the proceedings of CHES 2009, Lecture Notes in Computer Science, vol 5747, pp 429-443, Lausanne, Switzerland, Spetember 2009.

# A Definition of the metrics

## A.1 Security metric

We define the success rate of a side-channel adversary as follows. First, the adversary $\mathsf{A}_{\mathsf{E}_K,\mathsf{L}}$ is an algorithm with limited time complexity $\tau$, memory complexity $m$ and queries $q$ to the target implementation $(\mathsf{E}_K, \mathsf{L})$. Its goal is to guess a subkey $s = \gamma(k)$ with non negligible probability. For this purpose, we assume that the adversary $\mathsf{A}_{\mathsf{E}_K,\mathsf{L}}$ outputs a guess vector $\mathbf{g} = [g_1, g_2, \ldots, g_{|\mathcal{S}|}]$ with the different subkey candidates sorted according to the attack result: the most likely candidate being $g_1$. A success rate of order 1 (*resp.* 2, ...) relates to the probability that the correct subkey is sorted first (*resp.* among the two first ones, ...) by the adversary. More formally, we define the experiment:

$$
\begin{array}{ll}
\text{Experiment } \mathbf{Exp}^{\text{sc-kr-}o}_{\mathsf{A}_{\mathsf{E}_K,\mathsf{L}}} & \\
k \xleftarrow{R} \mathcal{K}; & \\
s = \gamma(k); & \\
\mathbf{g} \leftarrow \mathsf{A}_{\mathsf{E}_k,\mathsf{L}}; & \\
\textbf{if } s \in [g_1, \ldots, g_o] & \textbf{then } \text{return } 1; \\
 & \textbf{else } \text{return } 0;
\end{array}
$$

The $o^{\text{th}}$-order success rate of the side-channel key recovery adversary $\mathsf{A}_{\mathsf{E}_K,\mathsf{L}}$ against a subkey variable $S$ is straightforwardly defined as:

$$\mathbf{Succ}^{\text{sc-kr-}o,S}_{\mathsf{A}_{\mathsf{E}_K,\mathsf{L}}}(\tau, m, q) = \Pr\ [\mathbf{Exp}^{\text{sc-kr-}o}_{\mathsf{A}_{\mathsf{E}_K,\mathsf{L}}} = 1] \tag{7}$$

We also define the success rate for a given subkey value $s$ with the same experiment, where $s$ is not picked up at random but fixed to an arbitrary value. Theorem 1 considers this second definition. Note also that [18] defines an alternative security metric (namely, the guessing entropy) that can be used to measure the efficiency of a side-channel adversary in a more flexible fashion: it measures the average position of the correct subkey candidate in the guess vector $\mathbf{g}$.

### A.2 Information theoretic metric

Using the notations of Section 2.1, let $\Pr[s|\mathbf{l}_q]$ be the conditional probability of a subkey $s$ given a leakage vector $\mathbf{l}_q$. We first define a conditional entropy matrix:

$$\mathbf{H}^q_{s,s^*} = -\sum_{\mathbf{l}_q} \Pr[\mathbf{l}_q|s] \cdot \log_2 \Pr[s^*|\mathbf{l}_q],$$

where $s$ and $s^*$ denote the correct key and a candidate out of $|\mathcal{S}|$ possible ones in a side-channel attack. A conditional entropy matrix typically looks like:

$$\mathbf{H}^q_{s,s^*} = \begin{pmatrix} h_{1,1} & h_{1,2} & ... & h_{1,|\mathcal{S}|} \\ h_{2,2} & h_{2,2} & ... & h_{2,|\mathcal{S}|} \\ ... & ... & ... & ... \\ h_{|\mathcal{S}|,1} & h_{|\mathcal{S}|,2} & ... & h_{|\mathcal{S}|,|\mathcal{S}|} \end{pmatrix}$$

For each line of the matrix, we denote the diagonal element $h_{s,s}$ as the residual entropy of a key $s$. Then, we define Shannon's conditional entropy:

$$\mathrm{H}[S|\mathbf{L}_q] = -\sum_{s} \Pr[s] \sum_{\mathbf{l}_q} \Pr[\mathbf{l}_q|s] \cdot \log_2 \Pr[s|\mathbf{l}_q] = \mathop{\mathbf{E}}_{s} \ \mathbf{H}^q_{s,s} \tag{8}$$

## B Proof sketch of Theorem 1

*Proof sketch.* We only give an argument of independence for a Bayesian adversary and assuming an EIS proposerty with $\phi(x_i, s)$ a group operation. A similar argument can be used for correlation attacks, distance-of-means tests and other functions $\phi$. Let us pick two subkeys $s$ and $s'$ and show that the conditional probabilities $\Pr[s|\mathbf{L}_q]$ and $\Pr[s'|\mathbf{L}_q]$ are equal. In practice, the leakages $\mathbf{L}_q$ are generated by a sequence of plaintexts $\mathbf{X}_q$. Because of the EIS property with $\phi$ a group operation, we have that $\forall x_i \in \mathcal{X}$, there is only one $x'_i \in \mathcal{X}$ such that $\phi(x_i, s) = \phi(x'_i, s')$ and for this $x'_i$, we have that $\mathsf{L}(x_i, s) = \mathsf{L}(x'_i, s')$. Hence, from each sequence of plaintexts $\mathbf{x}_q$ that is used to identify $s$, one can build a corresponding sequence $\mathbf{x}'_q$ to identify $s'$ such that the leakage function outputs corresponding to $(s, \mathbf{x}_q)$ and $(s', \mathbf{x}'_q)$ are identical and the number of different sequences equals $|\mathcal{X}|^q$ for both $\mathbf{x}_q$ and $\mathbf{x}'_q$. Let finally $\mathbf{L}'_q$ be the corresponding random leakage vector. Since the plaintext vectors $\mathbf{x}_q$ are uniformly distributed, we have that $\Pr[s|\mathbf{L}_q] = \Pr[s'|\mathbf{L}'_q] = \Pr[s'|\mathbf{L}_q]$. And since the same equalities holds for any pair of correct key candidates $(s, s')$ or incorrect pair of key candidates $(s^*, s'^*)$ we directly have the claimed key independence.

## C Key-dependent attack with non uniform plaintexts

Let us consider Figure 1 with the 8-bit S-box of the AES Rijndael and a Hamming weight leakage model without noise. Say that an adversary can only observe 2 plaintexts out of the 256 possible ones. Then, we can find a subkey value $s_1$

for which one of these two plaintexts leads to $z_i = 0$ or $z_i = 255$. Hence, this subkey can be recovered with probability one during the attack (since Hamming weights 0 and 8 exactly identify the S-box output $z_i$). But we can also find a subkey value $s_2$ for which these two plaintexts lead to different $z_i$ values such that when performing the attack, $s_2$ remains undistinguishable from a few other subkey candidates. Hence, the success rates against $s_1$ and $s_2$ are different.

## D Simplification of Bayes and correlation

A correlation attack selects the subkey based on (1), which we rewrite as follows.

$$\begin{aligned}
\tilde{s} &= \underset{s^*}{\operatorname{argmax}} \frac{\hat{\mathbf{E}}(\mathbf{L}_q \cdot \mathbf{M}_q^{s^*}) - \hat{\mathbf{E}}(\mathbf{L}_q) \cdot \hat{\mathbf{E}}(\mathbf{M}_q^{s^*})}{\hat{\sigma}(\mathbf{L}_q) \cdot \hat{\sigma}(\mathbf{M}_q^{s^*})} \\
&= \underset{s^*}{\operatorname{argmax}} \frac{\hat{\mathbf{E}}(\mathbf{L}_q \cdot \mathbf{M}_q^{s^*}) - \hat{\mathbf{E}}(\mathbf{L}_q) \cdot \hat{\mathbf{E}}(\mathbf{M}_q^{s^*})}{\hat{\sigma}(\mathbf{M}_q^{s^*})} \\
&= \underset{s^*}{\operatorname{argmax}} \frac{\hat{\mathbf{E}}(\mathbf{L}_q \cdot \mathbf{M}_q^{s^*}) - \hat{\mathbf{E}}(\mathbf{L}_q) \cdot \hat{\mathbf{E}}(\mathbf{M}_q^{s^*})}{\hat{\mathbf{E}}\left(\left(\mathbf{M}_q^{s^*}\right)^2\right) - \left(\hat{\mathbf{E}}(\mathbf{M}_q^{s^*})\right)^2}
\end{aligned}$$

A Bayesian attack selects the subkey based on (2), which we rewrite as follows.

$$\begin{aligned}
\tilde{s} &= \underset{s^*}{\operatorname{argmax}} \prod_{i=1}^{q} \frac{1}{\sqrt{2 \cdot \pi} \cdot \sigma_L} \cdot \exp\left(-\frac{1}{2} \cdot \left(\frac{l_i - m_i^{s^*}}{\sigma_L}\right)^2\right) \\
&= \underset{s^*}{\operatorname{argmin}} \sum_{i=1}^{q} \left(\frac{l_i - m_i^{s^*}}{\sigma_L}\right)^2 \\
&= \underset{s^*}{\operatorname{argmin}} \sum_{i=1}^{q} l_i^2 + (m_i^{s^*})^2 - 2 \cdot l_i \cdot m_i^{s^*} \\
&= \underset{s^*}{\operatorname{argmin}}\, \hat{\mathbf{E}}\left((\mathbf{L}_q)^2\right) - 2 \cdot \hat{\mathbf{E}}(\mathbf{L}_q \cdot \mathbf{M}_q^{s^*}) + \hat{\mathbf{E}}\left((\mathbf{M}_q^{s^*})^2\right) \\
&= \underset{s^*}{\operatorname{argmax}}\, 2 \cdot \hat{\mathbf{E}}(\mathbf{L}_q \cdot \mathbf{M}_q^{s^*}) - \hat{\mathbf{E}}\left((\mathbf{M}_q^{s^*})^2\right) \\
&= \underset{s^*}{\operatorname{argmax}} \frac{\hat{\mathbf{E}}(\mathbf{L}_q \cdot \mathbf{M}_q^{s^*})}{\hat{\mathbf{E}}\left((\mathbf{M}_q^{s^*})^2\right)} - 1 \\
&= \underset{s^*}{\operatorname{argmax}} \frac{\hat{\mathbf{E}}(\mathbf{L}_q \cdot \mathbf{M}_q^{s^*})}{\hat{\mathbf{E}}\left((\mathbf{M}_q^{s^*})^2\right)}
\end{aligned}$$

Notes. The simplification of Bayes and relation to the correlation coefficient strongly relates to the Gaussian assumption of the leakage samples. Also, the leakage standard deviation term $\hat{\sigma}(\mathbf{L}_q)$ in a correlation attack, although asymptotically independent of the subkey candidates, can have an impact in attacks where several leakage samples (potentially having different standard deviations) have to be tested in parallel, *e.g.* in order to detect ghost peaks [2].

## E Connection with Kocher's original DPA attack

The original DPA attack, as it has been presented by Kocher *et al.* in [10] assumes a binary leakage model. Using our notation this means that $m_i^{s^*} \in [0,1]$. In the attack the adversary selects his key candidate as:

$$\tilde{s} = \underset{s^*}{\operatorname{argmax}} \; \frac{1}{q_1^{s^*}} \sum_{i=1}^{q} l_i \cdot m_i^{s^*} - \frac{1}{1 - q_1^{s^*}} \sum_{i=1}^{q} l_i \cdot (1 - m_i^{s^*}),$$

where $q_1^{s^*} = \sum_{i=1}^{q} m_i^{s^*}$. This classical definition can be rewritten as follows:

$$\tilde{s} = \underset{s^*}{\operatorname{argmax}} \; \sum_{i=1}^{q} l_i \cdot \left( \frac{m_i^{s^*}}{q_1^{s^*}} - \frac{1 - m_i^{s^*}}{q - q_1^{s^*}} \right) = \frac{1}{q} \sum_{i=1}^{q} l_i \cdot \tilde{m}_i^{s^*},$$

where $\tilde{m}_i^{s^*} = m_i^{s^*} \frac{q}{q_1} + (m_i^{s^*} - 1) \cdot \frac{q}{q - q_1}$ and therefore $\tilde{m}_i^{s^*} \in [-\frac{q}{q-q_1}, \frac{q}{q_1}]$. These two values equal the probabilities that $m_i^{s^*}$ is zero or one, respectively. In case of a correlation attack, the adversary selects the key according Equation (1). The correlation between two random variables is invariant under linear transformations of the variables and hence $\hat{\rho}(\mathbf{l}_q, \mathbf{m}_q^{s^*}) = \hat{\rho}(\mathbf{l}_q, \tilde{\mathbf{m}}_q^{s^*})$. Using almost exactly the same proof as in case of Theorem 3 we therefore have that the difference of success rate between the original DPA attack and the correlation attack is a monotonously decreasing function of $\sigma_R^2$. Thanks to Theorem 3 this also holds for the difference between a Bayesian attack and the original DPA attack.

## F Proof of Theorem 4

From [7], we write the joint entropy of a multivariate gaussian distribution as:

$$\mathrm{H}(X_1, \ldots, X_n) = \frac{1}{2} \cdot \log_2 \left( (2\pi e)^n \cdot |C| \right), \tag{9}$$

where $|C|$ is the covariance matrix corresponding to $(X_1, \ldots, X_n)$. Considering a bivariate distribution $(X, Y)$, it yields (we use $\rho(X,Y) = \rho$ for short):

$$C = \begin{pmatrix} \sigma_X^2 & \rho \cdot \sigma_X \sigma_Y \\ \rho \cdot \sigma_Y \sigma_X & \sigma_Y^2 \end{pmatrix}$$

Add filling this covariance matrix in (9), we directly find:

$$\begin{aligned} \mathrm{I}(X;Y) &= \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y) \\ &= \frac{1}{2} \cdot \log_2(2\pi e \cdot {\sigma_X}^2) + \frac{1}{2} \cdot \log_2(2\pi e \cdot {\sigma_Y}^2) \\ &\quad - \frac{1}{2} \log_2 \left( (2\pi e)^2 \cdot (\sigma_X^2 \sigma_Y^2 - \rho^2 \sigma_X^2 \sigma_Y^2) \right) \\ &= -\frac{1}{2} \cdot \log_2(1 - \rho^2) \end{aligned}$$

## G $\mathbf{H[V_1|L_1] = H[V_1|M_1] + H[M_1|L_1]}$

This relation comes from the observation that, in our scenario:

1) $\mathrm{H}[\mathbf{V}_1|\mathbf{M}_1, \mathbf{L}_1] = \mathrm{H}[\mathbf{V}_1|\mathbf{M}_1]$, *i.e.* knowing the model $\mathbf{M}_1$, there is nothing to learn about the target values by observing the leakage $\mathbf{L}_1$.
1) $\mathrm{H}[\mathbf{M}_1] = \mathrm{I}(\mathbf{M}_1; \mathbf{V}_1)$, *i.e.* the only randomness in $\mathbf{M}_1$ comes from $\mathbf{V}_1$.

Then, by applying standard relations from [7], we find:

$$\begin{aligned}
\mathrm{H}[\mathbf{V}_1|\mathbf{L}_1] &= \mathrm{H}[\mathbf{V}_1|\mathbf{M}_1, \mathbf{L}_1] + \mathrm{I}[\mathbf{V}_1; \mathbf{M}_1|\mathbf{L}_1] \\
&= \mathrm{H}[\mathbf{V}_1|\mathbf{M}_1] + \mathrm{I}[\mathbf{V}_1; \mathbf{M}_1|\mathbf{L}_1] \\
&= \mathrm{H}[\mathbf{V}_1|\mathbf{M}_1] + \mathrm{H}[\mathbf{M}_1|\mathbf{L}_1]
\end{aligned}$$