The World is Not Enough: Another Look on Second-Order DPA

François-Xavier Standaert[1], Nicolas Veyrat-Charvillon[1], Elisabeth Oswald[2], Benedikt Gierlichs[3], Marcel Medwed[4], Markus Kasper[5], Stefan Mangard[6].

[1] Université catholique de Louvain, Crypto Group, Belgium.
[2] University of Bristol, Department of Computer Science, UK.
[3] K.U. Leuven, ESAT/SCD-COSIC and IBBT, Belgium.
[4] Graz University of Technology, IAIK, Austria.
[5] Ruhr University Bochum, Horst Görtz Institute for IT Security, Germany.
[6] Infineon Technologies AG, Security Innovation, Germany.

Abstract. In a recent work, Mangard *et al.* showed that under certain assumptions, the (so-called) standard univariate side-channel attacks using a distance-of-means test, correlation analysis and Gaussian templates are essentially equivalent. In this paper, we show that in the context of multivariate attacks against masked implementations, this conclusion does not hold anymore. While a single distinguisher can be used to compare the susceptibility of different unprotected devices to first-order DPA, understanding second-order attacks requires to carefully investigate the information leakages and the adversaries exploiting these leakages, separately. Using a framework put forward by Standaert *et al.* at Eurocrypt 2009, we provide the first analysis that explores these two topics in the case of a masked implementation exhibiting a Hamming weight leakage model. Our results lead to refined intuitions regarding the efficiency of various practically-relevant distinguishers. Further, we also investigate the case of second- and third-order masking (*i.e.* using three and four shares to represent one value). This evaluation confirms that higher-order masking only leads to significant security improvements if the secret sharing is combined with a sufficient amount of noise. Eventually, we show that an information theoretic analysis allows determining this necessary noise level, for different masking schemes and target security levels, with high accuracy and smaller data complexity than previous methods.

1 Introduction

Masking (as described, *e.g.* in [2, 7, 19]) is a very frequently considered solution to thwart side-channel attacks. The basic idea is to randomize all the sensitive variables during a cryptographic computation by splitting them into d shares. The value $d-1$ is usually denoted as the order of the masking scheme. As most countermeasures against side-channel attacks, masking does not totally prevent the leakages but it is expected to increase the difficulty of performing a successful key-recovery. For example, masking can be defeated because of technological issues such as glitches [9]. Alternatively, an adversary can always perform a higher-order DPA (*e.g.* [10, 13, 23]) in which he "combines" the leakages corresponding to the d shares in order to extract key-dependent information. From a performance point of view, masking a block cipher implies significant performance overheads, because it requires to compute the encryption of the different shares separately. As a result, an important problem is to determine the exact security level that it provides in function of the order of the scheme $d-1$.

In order to solve this problem, Prouff *et al.* proposed a comprehensive study of first-order masking (*i.e.* second-order power analysis) in [17]. In their paper, the two leakage samples corresponding to the different shares are first mingled with a combination function. Next, a (key-dependent) leakage model is used to predict the output of this function. Eventually, the combined physical leakages are compared with the key-dependent predictions, thanks to Pearson's correlation coefficient [1]. Different combination functions are analyzed regarding the efficiency of the resulting attacks, leading to the following conclusions:

1. For every device and combination function, an optimal prediction function (or model) can be exhibited, that leads to the best attack efficiency.
2. Following an analysis based on Pearson's coefficient and assuming a "Hamming weight leakage model", the "normalized product combining function" (both to be detailed in this paper) is the best available in the literature.

The first observation is in fact quite natural. Since every device is characterized by its leakage function, there is one optimal model to predict these leakages that perfectly captures their probability density function (pdf for short). And for every optimal model, there is one way to combine the leakage samples that leads to the best possible correlation. But the idea of *optimal combination function* also leads to a number of issues. On the one hand, as acknowledged by the authors of [17], their analysis is carried out for a fixed (Hamming weight) leakage function. Therefore, how the observations made in this context would be affected by a different leakage function is an open question. On the other hand, their analysis is also performed for a given statistical tool, *i.e.* Pearson's correlation coefficient. Hence, one can wonder about the extent to which this statistical tool is generic enough for evaluating second-order DPA.

This second question is particularly interesting in view of the recent results of [12]. This reference shows that in the context of (so-called) standard first-order DPA and when provided with the same leakage model, the most popular distinguishers such as using distance-of-means tests [8], correlation analysis and Gaussian templates [3] require approximately the same number of traces to extract keys. Differences observed in practice are only due to statistical artifacts. In addition, it is shown that the correlation coefficient can be related to the concept of conditional entropy which has been established as a measure for side-channel leakage in [20]. Therefore, a natural question is to ask if these observations still hold in the second-order case. For example, can the correlation coefficient be used to evaluate the information leakage of a masked implementation?

In this paper, we answer this question negatively. We show that second-order DPA attacks are a typical context in which the two parts of the framework for the analysis of side-channel key-recovery of Eurocrypt 2009 lead to different intuitions. First, an information theoretic analysis measures the amount of leakage provided by the masked implementation. It quantifies its security limits and relates to the success rate of an adversary who can perfectly profile the leakage pdf. Second, a security analysis measures the efficiency of one particular distinguisher. By applying this framework, we exhibit refined intuitions regarding the behavior of different second-order DPA attacks and combination functions. We then discuss the impact of these observations in profiled and non-profiled attack scenarios and confirm our theoretical investigations with practical experiments. We note that our results do not contradict [17] but rather emphasize that a single distinguisher cannot capture all the specificities of a leakage function. Eventually, we extend our analysis towards higher-order masking. This allows us to confirm that, from an information theoretic point of view, increasing the number of shares in a masking scheme only leads to an improved physical security if a sufficient amount of noise is limiting the quality of the adversary's measurements [2]. Higher-order masking also provides a case for the information theoretic metric introduced in [20]. We show that this metric can be used to determine the exact amount of shares and noise required to reach a certain security level (against worst-case template attacks, exploiting intensively profiled leakage models), with smaller data complexity than previous methods.

Summarizing, first-order side-channel attacks are a quite simple context in which (under certain conditions) most popular distinguishers behave similarly, if they are fed with the same leakage models. As a consequence, it can be sound to use "one distinguisher for all" in this context. By contrast, second-order (or higher-order) DPA can be confronted with leakage probability distributions that can take very different forms (mixtures, typically). Hence, given a certain amount of information leaked by a masked implementation, and even if fed with the same leakage models (and combination functions), different statistical tools will take advantage of the key-dependencies in very different manners. In other words, depending on the devices and countermeasures, one or another attack may perform better, hence suggesting our title "the world is not enough".

2 Boolean masking and second-order attacks

Many different masking schemes have been proposed in the literature. Although they can result in significantly different performances, the application of second-order attacks generally relies on the same principles, independent of the type of masking. In the following, we decided to focus on the Generalized Look Up Table (GLUT for short) that is described, e.g. in [16]. Such a scheme is represented in the lower left part of Figure 1, using the key addition and S-box layer of a block cipher as a concrete example. It can be explained as follows. For an input plaintext x_i, a random mask a_i is first generated within the device. The value $x_i \oplus a_i$ is generally denoted as the masked variable. Then, the encryption algorithm (here, the key addition and S-box) are applied to the masked variables, where s denotes a secret key byte (we will use the term subkey in the following). Concurrently, some correction terms are also computed such that anytime during the cryptographic computation, the XOR between a masked variable and its corresponding mask produces the original variable. In the case of the GLUT proposal, a precomputed function Sbox' is used for this purpose. For example in Figure 1, the masked S-box output $\mathsf{Sbox}(x_i \oplus a_i \oplus s)$ can be written as $\mathsf{Sbox}(x_i \oplus s) \oplus b_i$, where b_i denotes an output mask produced by Sbox'.

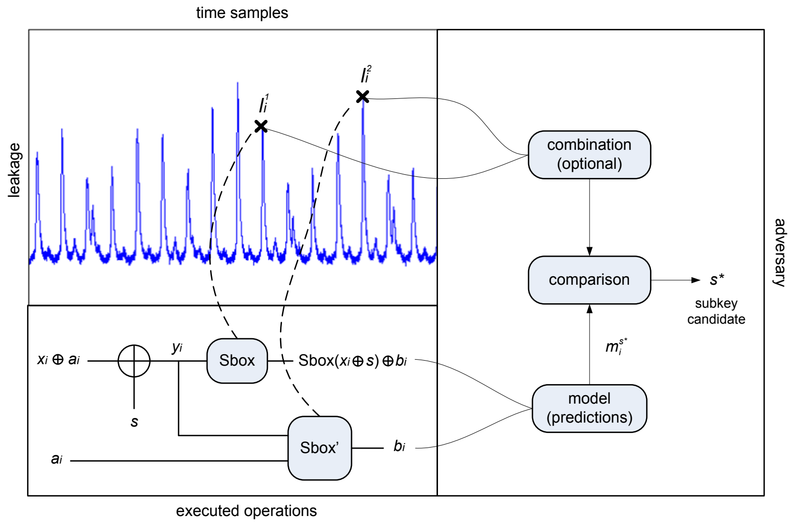

Fig. 1. Illustrative second-order DPA.

In practice, the GLUT countermeasure can be implemented in different manners. Mainly, the two S-box computations can be performed sequentially (as typical for software implementations) or in parallel (as typical for hardware implementations). In order to describe the second-order DPA that we investigate in this paper, we first use the sequential approach (the parallel one will be discussed in the next section). Also, we rely on the terminology introduced in [20]. Essentially, the idea of second-order DPA is to take advantage of the joint leakage of two intermediate computations during the encryption process (*i.e.* the masked value and its mask). In the software approach, the computation of these intermediate variables will typically be performed in two different clock cycles. Hence, two leakage samples l_i^1 and l_i^2 corresponding to these computations can be found in the leakage traces, as in the top of Figure 1. Following the standard DPA described in [12], the adversary will then work in three (plus one optional) 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 the S-box outputs $\mathsf{Sbox}(x_i \oplus s)$ in Figure 1.
2. For each of these predicted values, the adversary models the leakages. Because of the presence of a mask in the implementation, this prediction can use a pdf (where the probability is taken over the masks and leakage noise) or some simpler function *e.g.* capturing only certain moments of this pdf.

3. Optionally, the adversary combines the leakage samples into a single variable.
4. For each subkey candidate s^*, the adversary finally compares the modeled leakages with actual measurements, produced with the same plaintexts x_i and a secret subkey s. In a second-order DPA, each model is compared with two samples in the traces. This comparison is independent of all other points. Consequently, these attacks are referred to as *bivariate*. In practice, this comparison is applied to many pairs of points in the leakage traces and the subkey candidate that performs best is selected by the adversary.

As for the analysis of first-order attacks, comparing different distinguishers requires to provide them with the same leakage samples. However, contrary to the first-order case and as will be discussed in the following sections, the best pair of leakage samples is not necessarily the same for all distinguishers. This is because different distinguishers can take advantage of different leakage pdf with different efficiencies in this case. In practice, this requires to test all pairs of samples in the traces (but this means $N(N-1)/2$ statistical tests to perform if the traces have N samples). In this paper, we will generally assume that this best pair of samples is provided to the attacks we perform (which can be done easily when simulating experiments and requires significant - but tractable - computational power when performing attacks based on real measurements).

Finally, we will use the following notations:

- $\mathbf{x}_q = [x_1, x_2, \ldots, x_q]$: a vector of plaintext bytes.
- $\mathbf{a}_q = [a_1, a_2, \ldots, a_q]$: a vector of random input mask bytes.
- $\mathbf{b}_q = [b_1, b_2, \ldots, b_q]$: a vector of random output mask bytes.
- $v_i^1 = \mathsf{Sbox}(x_i \oplus s) \oplus b_i$: an intermediate value in the encryption of x_i.
- $v_i^2 = b_i$: another intermediate value in the encryption of x_i.
- $\mathbf{l}_q^1 = [l_1^1, l_2^1, \ldots, l_q^1]$: a vector of leakage samples corresponding to the first intermediate values v_i^1 during the encryption process.
- $\mathbf{l}_q^2 = [l_1^2, l_2^2, \ldots, l_q^2]$: a vector of leakage samples corresponding to the second intermediate values v_i^2 during the encryption process.
- $\mathbf{m}_q^{s^*} = [m_1^{s^*}, m_2^{s^*}, \ldots, m_q^{s^*}]$: a vector containing leakage models (*i.e.* predictions) corresponding to a subkey candidate s^* and the plaintexts $\mathbf{x}_q$.

In the rest of the paper, these notations (in small caps) will represent sampled values, while their counterpart in capital letters will represent random variables.

3 Second-order attacks with Pearson's coefficient

In theory, second-order DPA is possible if the joint probability distributions $\Pr[\mathbf{L}_q^1, \mathbf{L}_q^2 | \mathbf{X}_q, s]$ are different for different subkey values s. This can be illustrated, *e.g.* for a Hamming weight leakage function which is frequently considered in the practice of side-channel attacks [11] and has been the running example in [17]. It means assuming that the leakage samples l_i^1 and l_i^2 can be written as:

$$l_i^1 = \mathrm{W_H}(v_i^1) + n_i^1, \quad (1)$$
$$l_i^2 = \mathrm{W_H}(v_i^2) + n_i^2, \quad (2)$$

where $\mathrm{W_H}$ is the Hamming weight function and n_i^1, n_i^2 are normally distributed noise values with mean 0 and standard deviation σ_n. In the context of an 8-bit S-box (*e.g.* the AES one), it leads to 9 possible leakage distributions, corresponding to the 9 Hamming weight values of a secret state $\Sigma_i = \mathsf{Sbox}(x_i \oplus s)$, as observed in [21]. The left parts of Figures 11, 12 and 13 in Appendix A show the joint leakage distributions in this setting and clearly illustrate that they are key-dependent. As detailed in the previous section, taking advantage of these dependencies requires a comparison tool. In their statistical evaluation of second-order DPA, Prouff *et al.* use Pearson's correlation coefficient. In the context of first-order attacks exploiting a single leakage sample l_i, it implies computing:

$$\hat{\rho}(\mathbf{M}_q^{s^*}, \mathbf{L}_q) = \frac{\hat{\mathbf{E}}\Big(\big(l_i - \hat{\mathbf{E}}(\mathbf{L}_q)\big) \cdot \big(m_i^{s^*} - \hat{\mathbf{E}}(\mathbf{M}_q^{s^*})\big)\Big)}{\hat{\sigma}(\mathbf{L}_q) \cdot \hat{\sigma}(\mathbf{M}_q^{s^*})},$$

where $\hat{\mathbf{E}}$ and $\hat{\sigma}$ denote the sample means and standard deviations of a random variable, respectively. In order to extend this tool towards the second-order case, the classical approach is to first combine the two leakage samples l_i^1 and l_i^2 with a *combination function* C. For example, Chari *et al.* proposed to take the product of two centered samples [2]: $\mathsf{C}(l_i^1, l_i^2) = (l_i^1 - \hat{\mathbf{E}}(\mathbf{L}_q^1)) \cdot (l_i^2 - \hat{\mathbf{E}}(\mathbf{L}_q^2))$ and Messerges used the absolute difference between them [13]: $\mathsf{C}(l_i^1, l_i^2) = |l_i^1 - l_i^2|$. As illustrated in the right parts of Figures 11, 12 and 13, those combining functions also lead to key-dependencies. In addition to these standard examples, we finally plotted the distribution of the sum combining function $\mathsf{C}(l_i^1, l_i^2) = l_i^1 + l_i^2$ because it can be used to emulate the behavior of the GLUT masking in a hardware setting, where the two S-boxes of Figure 1 are computed in parallel.

3.1 Choice of a model and leakage-dependency of C

Given the above descriptions and assuming that the adversary knows a good leakage model for the samples l_i^1 and l_i^2, it remains to determine which model to use when computing $\hat{\rho}(\mathbf{M}_q^{s^*}, \mathsf{C}(\mathbf{L}_q^1, \mathbf{L}_q^2))$. That is, we do not need to predict the leakage samples separately, but their combination. In addition and contrary to the first-order case, there is an additional variable (*i.e.* the mask) that is unknown to the adversary. But given a model for the separate samples, it is possible to derive one for their combination. For example, assuming a Hamming weight model that perfectly corresponds to the leakages of Equations (1) and (2), we can use the mean of the combination function, taken over the masks. For each subkey candidate s^*, the model is is then given by:

$$m_i^{s^*} = \mathop{\mathbf{E}}_{b_i}\Big(\mathsf{C}\big(\mathrm{W_H}(\Sigma_i^* \oplus b_i), \mathrm{W_H}(b_i)\big)\Big).$$

This is in fact similar to what is proposed in [17], where the mean is additionally taken over the leakage noise (which is more general, but implies additional profiling, *i.e.* a sufficiently precise knowledge of the noise distribution). As an illustration, Figure 2 shows the leakage models corresponding to the absolute difference and normalized product combination functions. They again only depend on the 9 Hamming weight values of the secret state, as opposed to the ones of a sum combining function for which the mean value (over the masks) is constant for all secret states. Hence, as already observed in [11], this sum combining function will not lead to successful second-order correlation attacks.

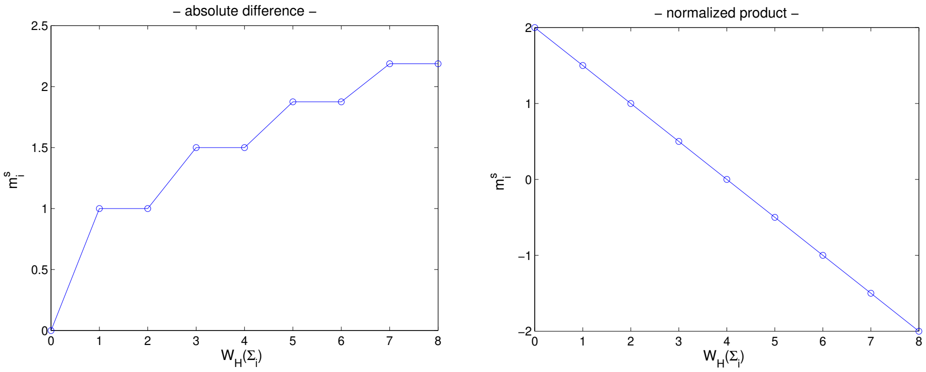

Fig. 2. Leakage models for second-order DPA using the correlation coefficient.

The figure intuitively confirms the previous theoretical analysis of Prouff *et al.* where it is demonstrated that the normalized product combining function leads to the most efficient second-order side-channel attacks when using Pearson's coefficient and assuming a Hamming weight leakage model for the separate samples. Indeed, this particular setting gives rise to nicely linear dependencies of the models m_i^s in the Hamming weight of the secret states $\mathrm{W_H}(\Sigma_i)$. Also, and contrary to the absolute difference combining function, all the 9 possible Hamming weights correspond to a different model m_i^s in this particular case.

Interestingly, the efficiency of the normalized product combining function can be simply explained when looking at the equations since it computes:

$$\hat{\rho}(\mathbf{M}_q^{s^*}, \mathsf{C}(\mathbf{L}_q^1, \mathbf{L}_q^2)) = \frac{\hat{\mathbf{E}}\Big(\big(\mathsf{C}(l_i^1, l_i^2) - \hat{\mathbf{E}}(\mathsf{C}(\mathbf{L}_q^1, \mathbf{L}_q^2))\big) \cdot \big(m_i^{s^*} - \hat{\mathbf{E}}(\mathbf{M}_q^{s^*})\big)\Big)}{\hat{\sigma}(\mathsf{C}(\mathbf{L}_q^1, \mathbf{L}_q^2)) \cdot \hat{\sigma}(\mathbf{M}_q^{s^*})}.$$

As the product is normalized, we have that $\hat{\mathbf{E}}(\mathsf{C}(\mathbf{L}_q^1, \mathbf{L}_q^2)) = 0$, which leads to:

$$\hat{\rho}(\mathbf{M}_q^{s^*}, \mathsf{C}(\mathbf{L}_q^1, \mathbf{L}_q^2)) = \frac{\hat{\mathbf{E}}\Big(\big(l_i^1 - \hat{\mathbf{E}}(\mathbf{L}_q^1)\big) \cdot \big(l_i^2 - \hat{\mathbf{E}}(\mathbf{L}_q^2)\big) \cdot \big(m_i^{s^*} - \hat{\mathbf{E}}(\mathbf{M}_q^{s^*})\big)\Big)}{\hat{\sigma}(\mathsf{C}(\mathbf{L}_q^1, \mathbf{L}_q^2)) \cdot \hat{\sigma}(\mathbf{M}_q^{s^*})}. \quad (3)$$

And this formula is in fact very close to the straightforward generalization of Pearson's correlation coefficient to the case of three random variables:

$$\hat{\rho}(\mathbf{M}_q^{s^*}, \mathbf{L}_q^1, \mathbf{L}_q^2) = \frac{\hat{\mathbf{E}}\Big(\big(l_i^1 - \hat{\mathbf{E}}(\mathbf{L}_q^1)\big) \cdot \big(l_i^2 - \hat{\mathbf{E}}(\mathbf{L}_q^2)\big) \cdot \big(m_i^{s^*} - \hat{\mathbf{E}}(\mathbf{M}_q^{s^*})\big)\Big)}{\hat{\sigma}(\mathbf{L}_q^1) \cdot \hat{\sigma}(\mathbf{L}_q^2) \cdot \hat{\sigma}(\mathbf{M}_q^{s^*})}. \qquad (4)$$

The only difference between Equations (3) and (4) is in the leakage samples' standard deviation terms, which are key-independent. Hence, when applied to the same pair of samples, attacks using Equations (3) or (4) are equivalent. Intuitively, these equations provide a simple explanation of the normalized product combining function. That is, such a combining function will efficiently take advantage of pairs of leakage samples that are linearly correlated conditioned on the key. As illustrated in Figures 11, 12 and 13, this is nicely achieved in the case of a Hamming weight leakage function for the two samples l_i^1 and l_i^2.

4 Evaluating second-order leakage: IT analysis

In general, the evaluation of second-order side-channel attacks is not straightforward to capture. More precisely, it is easy to see that an analysis based only on the correlation coefficient may suffer from certain limitations. For example:

- Given Pearson's correlation coefficient as a distinguisher and a Hamming weight leakage function, there exist (trivial) combination functions for the samples (*e.g.* the sum) that do not lead to successful key recoveries.
- Given Pearson's coefficient as a distinguisher and the normalized product combination function, there exist leakage functions (*e.g.* with no linear dependencies between the samples) that don't lead to successful key recoveries.

These observations suggest that the simple situation in the first-order context, where the correlation coefficient could (under certain physical assumptions detailed in [12]) be used both as a distinguisher and as a measure of side-channel leakage, does not hold here. In second-order side-channel attacks, this correlation is only a distinguisher. Hence, it is a typical context in which the evaluation framework of Eurocrypt 2009 is interesting to put into practice:

1. First, an information theoretic analysis is performed, in order to evaluate the physical leakages, independently of the adversary who exploits them. When applied to a countermeasure (*e.g.* masking), this step allows to quantify how much the security of the device has been improved against an adversary who can perfectly profile the leakage pdf. In other words, it can be used as an objective measure of the quality of the countermeasure, in a worst case scenario (*i.e.* best adversary, large number of queries - see [20] for the details).
2. Second, a security analysis is performed, in order to evaluate how efficiently a particular distinguisher (*e.g.* Pearson's correlation coefficient with a given combining function) can exploit the available leakage. This step is useful to translate the previous information theoretic analysis into a "number of measurements required to extract the key", in a given scenario.

In this section, we tackle the first part of the analysis. For this purpose, and in order to compare our conclusions with previous works, we use exactly the same assumptions as [17], *i.e.* a Hamming weight leakage function for the two samples, just as described in Section 3. Following the definitions in [20], we compute:

$$\mathrm{H}[S|\mathbf{L}_1^1,\mathbf{L}_1^2,\mathbf{X}_1] = -\sum_s \Pr[s] \sum_{x_1} \Pr[x_1] \int_{l_1^1}\int_{l_1^2} \Pr[l_1^1,l_1^2|s,x_1] \log_2 \Pr[s|l_1^1,l_1^2,x_1]\, dl_1^1 dl_1^2.$$

Since the leakage samples are assumed to be normally distributed, this can be quite easily done in function of the noise standard deviation σ_n. Some simplifications allow to speed up the computations, *e.g.* by observing that only nine distributions are possible, corresponding to the nine Hamming weights of the secret states Σ_i. Also, in order to evaluate the information loss caused by the different combination functions, we similarly evaluated $\mathrm{H}[S|\mathsf{C}(\mathbf{L}_1^1,\mathbf{L}_1^2),\mathbf{X}_1]$. This implies slightly more complex integrals since, *e.g.* the product combining gives rise to mixtures of normal product distributions. Figure 9 in Appendix A illustrates these distributions for two secret states and two σ_n's. The mutual information values corresponding to these different information leakages (*i.e.* $\mathrm{I}(S;(\mathbf{L}_1^1,\mathbf{L}_1^2,\mathbf{X}_1)) = \mathrm{H}[S] - \mathrm{H}[S|\mathbf{L}_1^1,\mathbf{L}_1^2,\mathbf{X}_1]$) are then plotted in Figure 3, in function of the noise variance σ_n^2 (in log scale). From this figure, we can observe:

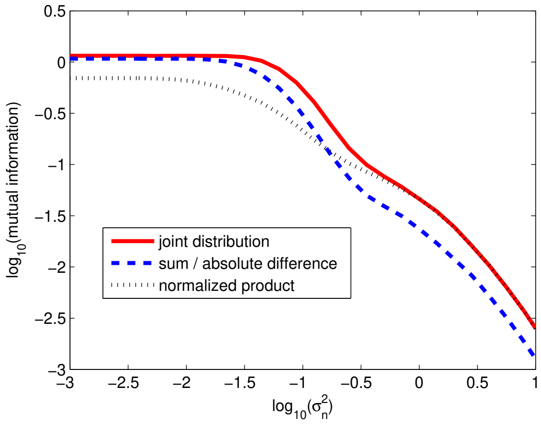

Fig. 3. Information leakage for different combination functions.

1. All combination functions imply a loss of information that can be avoided by dealing directly with the 2-dimensional joint leakage distribution.
2. The sum and absolute difference combining functions give rise to exactly the same information leakage. This can be understood from the shape of their distributions: the distribution of the absolute difference combining can be seen as the one of the sum combining that has been folded up.
3. For small σ_n^2, the normalized product is the least informative combining function. By contrast, when increasing the noise, the information leakage of the normalized product combining gets close to the one of the joint distribution.

4. The respective efficiency of different combining functions varies with the amount of noise. In particular, after a certain noise threshold, the product combining carries more information on S than the sum/absolute difference.

Note that the leakage of the sum combining's output clearly relates to the previous evaluation of [21] in which masking is analyzed in the hardware setting.

5 Implications for profiled attacks: security analysis (I)

The previous information theoretic analysis provides a new perspective to understand the relation between a masking scheme, its physical leakages and the exploitation of this information by a side-channel attack. For example, it exhibits that the sum combining function leads to significant information leakages (as can also be seen from the different pdf in appendix), although they cannot be directly exploited with Pearson's correlation coefficient. Previous works such as the one of Waddle and Wagner [23] showed how to overcome this limitation of the correlation coefficient, by squaring the combined samples. But our analysis raises the question whether these information leakages can be directly exploited (*i.e.* without squaring) by other distinguishers. In order to tackle this question, we apply the second part of the framework in [20], *i.e.* security analysis. This section starts with the evaluation of profiled (template) attacks, for which a strong relation with the previous information theoretic analysis should hold.

The results of various template attacks performed against the same masked AES S-box as in the previous sections are given in Figure 4, for two different noise standard deviations. We mention that these attacks do *not* use Gaussian templates as in [3] but the exact leakage distributions as in the previous information theoretic analysis (*e.g.* attacks using the joint distributions exploit Gaussian mixtures; attacks using the normalized product combining function exploit normal product distribution mixtures, *etc.* as plotted in appendix A). The different success rates are computed over 1000 independent experiments and nicely confirm the theoretical predictions of Theorem 2 in [20].

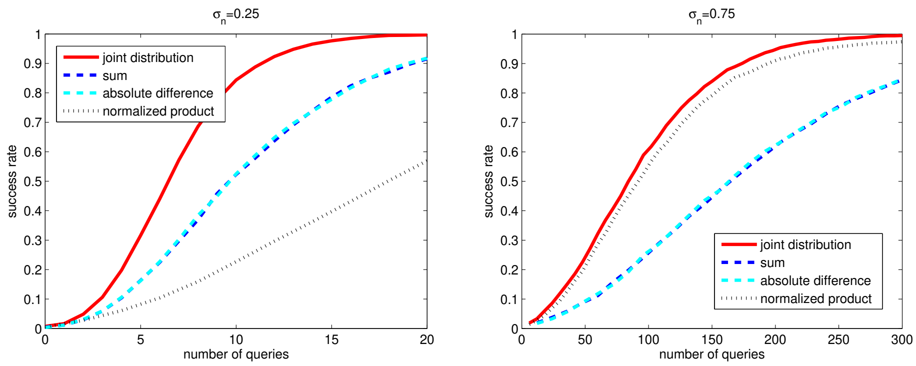

Fig. 4. Success rate of (simulated) profiled attacks against a masked AES S-box.

First, we see that the sum and absolute difference combining functions lead to the same attack efficiency in this profiled case (since their outputs lead to the same information leakages). Second, we see that the point in Figure 3 where the sum / absolute difference and the normalized product curves intersect is meaningful. Left of the intersection (*e.g.* for $\sigma_n = 0.25$), the sum / absolute difference combining functions allow more efficient attacks than the normalized product one. Right of the intersection (*e.g.* for $\sigma_n = 0.75$), the opposite conclusion holds. And as shown in Appendix A, Figure 10, these attacks have a similar efficiency at the intersection that falls around $\sigma_n = 0.4$ (that is, $\log_{10}(\sigma_n^2) \approx -0.8$).

Of course, these experiments are partially artificial since in practice, an adversary who can profile the leakages will generally use the templates based on the joint distribution only. At least, this is the best strategy if the adversary has enough data and time to profile the multivariate leakage pdf. However, our results confirm that an information theoretic analysis provides an objective evaluation of the quality of a countermeasure against the "best-available" template adversaries in the DPA setting. Hence, they emphasize that such an analysis is an important part in the evaluation of side-channel countermeasures. Also, these results lead to the same conclusions as [14], and show that resistance against sufficiently profiled template attacks cannot be achieved by masking only.

6 Implications for non-profiled attacks: security analysis (II)

The previous section showed that for carefully profiled template attacks, there is a strong connection between the information leakage of a device and the success rate of the adversary. By contrast, we know that in the non-profiled context of correlation attacks, this observation does not hold in general. For example, Pearson's coefficient cannot be used to exploit the leakages corresponding to the sum combining of Section 3.1. Hence, it is natural to check whether there exist other non-profiled distinguishers that can be successful in this case. We answer this question positively, using the Mutual Information Analysis (MIA) introduced in [5]. It can be seen as the counterpart of template attacks, in which the leakage distributions are estimated "on-the-fly" rather than prior to the attacks.

The success rates of correlation and MIA attacks (here, and in the rest of the paper, computed over 500 independent experiments), using different combining functions, are given in Figure 5, again using the (simulated) setting described in the previous section. In our experiments, MIA estimates the pdf using histograms with N_b linearly-spaced bins, and N_b corresponding to the number of possible values for the models, as proposed in [5]. That is, we use 9 bins per leakage sample and we partition the leakage samples according to the 9 Hamming weights of the secret state Σ_i. The following observations can be emphasized:

1. In the low noise scenario, MIA with the sum and absolute difference combining functions works best, as similarly observed for template attacks.
2. By contrast, and contrary to template attacks, MIA without combining function (*i.e.* using the joint distribution directly, as in [6, 18]), is not the most

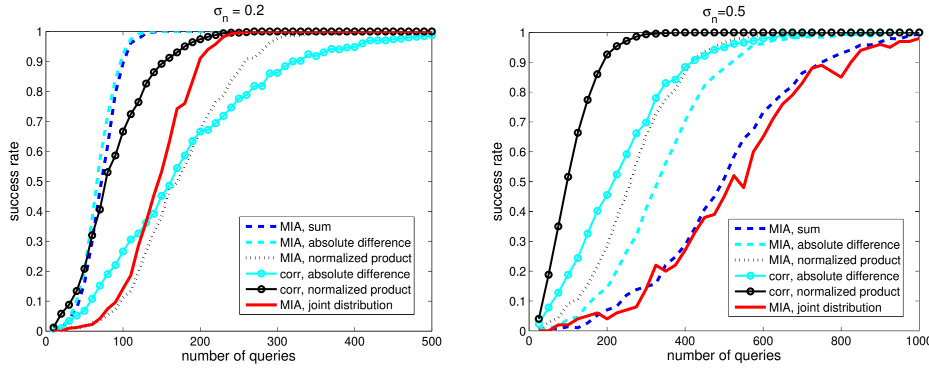

Fig. 5. Success rate of (simulated) non-profiled attacks - masked AES S-box.

efficient solution in our simulations. This is caused by the need to estimate two-dimensional distributions, which turns out to require more data.

3. For similar reasons (*i.e.* also related to the different efficiency of the "on-the-fly" pdf estimation), when increasing the noise, MIA with the sum and absolute difference combining functions are not equivalent anymore.
4. Finally, attacks using Pearson's correlation coefficient perform well, specially when combined with the normalized product (which is natural since our simulated leakages perfectly fulfill the requirements of Section 3.1).

Importantly, we note that all these non-profiled distinguishers lead to significantly lower efficiencies than the profiled ones in the previous section.

7 Experimental results

The previous sections evaluated the impact of masking an S-box with respect to various side-channel distinguishers, based on simulations. But as for most investigations in physically observable cryptography, it is important to verify that our conclusions are reasonably confirmed by practical measurements performed against a real chip. For this purpose, we also carried out a set of attacks against a masked implementation of the DES in an 8-bit RISC microcontroller from the Atmel AVR family. Considering the DES (rather than the AES) was motivated by practical facilities. Since the output of the DES S-box is 4-bit wide, it allows considering different contexts: in a first (low noise) scenario, the 4 remaining bits on the bus are kept constant; in a second scenario, these 4 bits are used in order to produce some additional algorithmic noise, by concatenating (secret) random strings to the two target values of Figure 1. This is interesting since the noise level was an important parameter, *e.g.* in our simulations of Figure 5. Hence, the different scenarios can be used to adapt the noise level in our experimental setting as well. The results in Figure 6 bring an interesting complement to our previous simulations and lead to the following observations:

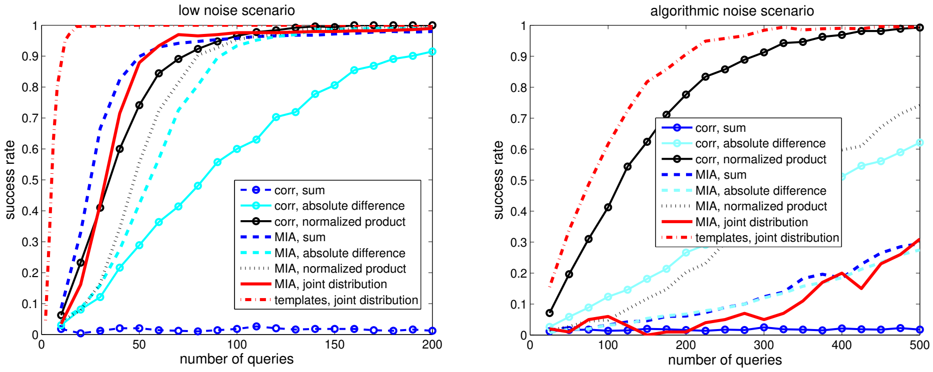

Fig. 6. Success rate of various experimental attacks against a masked DES.

1. The excellent efficiency of template attacks[1] and the good behavior of correlation attacks using the normalized product combining function are again exhibited. Interestingly, their respective efficiency gets closer when increasing the amount of algorithmic noise in the measurements, as it is suggested by the information theoretic analysis of Section 4.
2. MIA using the joint distribution is much more efficient than in the AES case. This is in fact related to the reduced number of bins that the 4-bit DES S-box allows in the pdf estimations (*i.e.* 25 rather than 81).
3. The presence of algorithmic noise (in the right part of Figure 6), affects the different distinguishers in a very different manner. To give a single example, MIA with the absolute difference combining function is strongly affected by this noise addition, compared to its counterpart using Pearson's coefficient.

Summarizing, these experiments confirm the "world is not enough" nature of second-order DPA that was already underlined in the previous simulations. The only strong statement that can be made in this context is that an information theoretic metric estimated with perfect templates captures the security against the best possible profiled adversary. As for all the other distinguishers, their efficiency highly depends on the actual shape of the leakage pdf and the engineering knowledge that can be exploited when mounting an attack. And contrary to the first-order case discussed in [12], the Gaussian assumption for the leakage samples does not hold anymore from the adversary's point of view (*e.g.* masking typically imply mixtures of Gaussians - or other - distributions).

8 Generalization to higher-orders

In order to improve the security of masking schemes further, one approach is to increase their order. For this purpose, this final section analyzes the cost *vs.* security tradeoff that can be obtained by generalizing the GLUT countermeasure in such a way, and details the second- and third-order cases for illustration. That

[1] We profiled our templates as described in the template-based DPA of [14].

is, rather than using one input mask per S-box, we now use two or three masks per S-box. In terms of cost, this implies using one or two additional tables Sbox'' and Sbox''', as described, *e.g.* in [15]. Conveniently, all the tools used in second-order DPA can be easily generalized to these third- and fourth-order attack cases. In particular, the information theoretic analysis of Section 4 just requires to integrate over three or four leakage samples l_i^1, l_i^2, l_i^3 and l_i^4.

The information leakage of these different masking schemes is represented in Figure 7, in function of the noise variance. On the same plot, we represented the average number of queries to the target device required for a perfectly profiled attack (similar to the ones in Section 5) to reach a success rate of 90%. These figures provide a quantitative insight to the observations in [2], where it is demonstrated that, given a large enough noise variance, the data complexity of a side-channel attack increases exponentially with the amount of shares in the masking scheme. That is, given a noise variance σ_n^2 in the leakage samples and k shares, the data complexity required to attack a masking scheme successfully is proportional to $(\sigma_n^2)^{k/2}$. The linear regions of the (log scale) curves that are observed in the right part of the figure suggest that this expectation is fulfilled in our experiments. Importantly, it also shows that the impact of (higher-order) masking can be extremely small in terms of security increases, for small σ_n^2's.

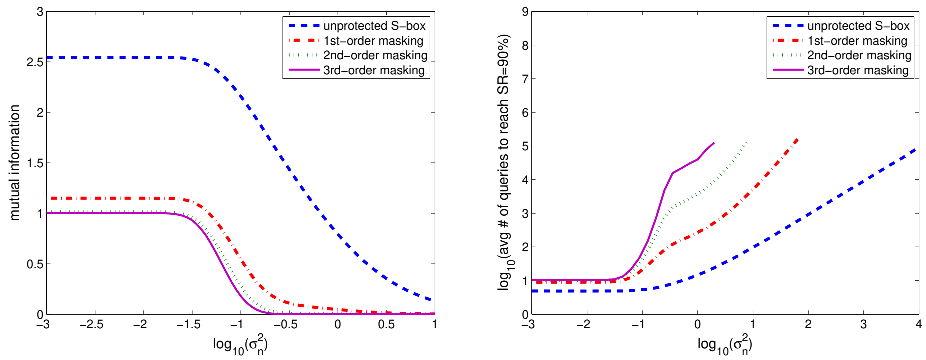

Fig. 7. Information leakage and success rates for 1^{st}, 2^{nd} and 3^{rd}-order masking.

Note finally that these results give a practical counterpart to the recent theoretical analysis of [4], where it is shown that masking schemes based on secret sharing techniques lead to secure implementations if the number of shares is adjusted to be large enough with respect to the noise in the measurements.

8.1 A case for the information theoretic metric

Looking at Figure 7, the main question for a designer (or evaluation laboratory) is to best trade the amount of shares and the amount of noise that he has to add to his implementation, in order to reach a certain security level. This is essential since increasing these parameters has a strong impact on the performance of the implementation. Unfortunately, for high security levels, the proper estimation of the number of traces required to reach a certain success rate becomes intensive (because of statistical sampling issues). Already in simulations, running 1000 attacks, each of them using 10^5 queries, is time consuming. And when moving to the analysis of real traces (taking much more time to be generated and space to be stored), this limitation becomes even more critical. Interestingly, this is exactly the context where an information theoretic analysis becomes useful. Given a leakage model, the mutual information $\mathrm{I}(S; \mathbf{L}_1^1, \mathbf{L}_1^2, \ldots)$ can be estimated with less data than the success rate of the corresponding template attack. And following [20], Theorem 2, it should hold that this mutual information is reasonably correlated with the number of traces required to reach a certain success rate. In order to confirm this expectation, we plotted an estimation of this number, based on the inverse of the mutual information multiplied with a constant factor c. As illustrated in Figure 8, this approximation holds nearly perfectly, with the same constant c for all attacks, essentially depending on the success rate to reach (here 90%). Summarizing, these simulations confirm the relevance of an information theoretic analysis when designing countermeasures against side-channel attacks.

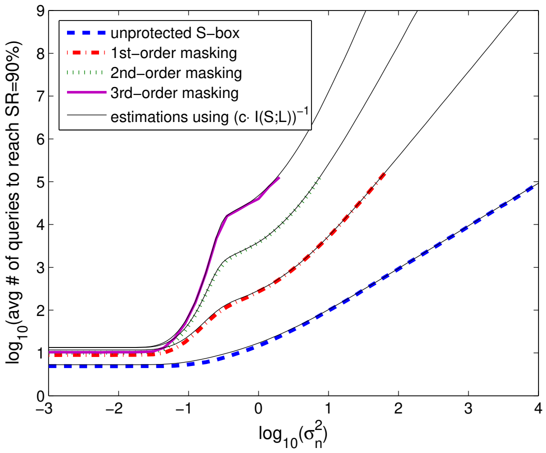

Fig. 8. Information leakage and success rates for 1^{st}, 2^{nd} and 3^{rd}-order masking.

Before to conclude, we note again that such an information theoretic analysis only captures the most powerful adversaries for which the profiling of the leakage distributions is perfect. But in practice, the reduction of the information leakage is not the only effect that increases the security in masked implementations. Namely, the pdf estimation of multidimensional distributions may become too complex for allowing the exploitation of all the information in the traces. And the number of pairs, triples, *etc.* of samples to test in the attacks also increases their time complexity considerably (up to N^2, N^3, *etc.*). However, we believe that the formal analysis of a worst-case scenario as in this paper is an important step towards a better understanding of the masking countermeasure.

9 Conclusions

The results in this paper provide a first complete and unifying treatment of higher-order power analysis. They allow putting forward the strengths and weaknesses of various approaches to second-order DPA and provide a sound explanation for them. Our analysis illustrates that in the context of cryptographic devices protected with masking, it is not sufficient to run a single arbitrary distinguisher to quantify the security of an implementation. Evaluations should hold in two steps. First, an information theoretic analysis determines the actual information leakage (*i.e.* the impact of the countermeasure, independently of the adversary). Second, a security analysis determines the efficiency of various distinguishers in exploiting this leakage. By applying such a methodology to simulations and practical experiments, we consequently obtain a fair and comprehensive evaluation of the security level that a masking scheme can ensure.

While not in contradiction with previous results in the field, these investigations reshape the understanding of certain assumptions and allow refined intuitions. First, theoretical analysis and empirical attacks sometimes show a large gap between the efficiency of profiled attacks that best exploit the information from two or more leakage samples and the one of non-profiled attacks that are most frequently used in practice. This relates to the observation that the statistics in side-channel attacks are only used to discriminate secret data (while their natural objective is to allow a good estimation). Hence, the study of advanced pdf estimation techniques in the context of side-channel attacks is an interesting direction for further research, as initiated with MIA in [5].

Second, the security improvement obtained when increasing the order of a masking scheme beyond one is negligible if it is not combined with a sufficient amount of noise in the leakages. This observation relates to the generally accepted intuition that side-channel resistance requires the combination of several countermeasures in order to be effective. We additionally show in this paper that an information theoretic analysis has very convenient features for evaluating this noise threshold precisely. As a result, the best combination of masking with other countermeasures (*e.g.* dual rail logic styles, time randomization, *etc.*) is a second interesting scope for further research. Finally, the relationship between the mutual information and the success rate of a profiled attack, that is experimentally exhibited in this paper in the context of second- (and higher-) order DPA, could be analyzed in order to obtain a more formal justification of it, *e.g.* under the assumption of Gaussian noise in the leakages.

Acknowledgements. This work has been funded in part by the European Commission's ECRYPT-II NoE (ICT-2007-216676), by the Belgian State's IAP program P6/26 BCRYPT, by the Walloon region's SCEPTIC project, by FWO project G.0300.07, by the K.U. Leuven-BOF (OT/06/40), and by the Austrian Science Fund (FWF) under grant number P22241-N23: "Investigation of Implementation Attacks". E. Oswald has been supported in part by the EPSRC grant number EP/F039638/1. F.-X. Standaert is a Research Associate of the Belgian Fund for Scientific Research (FNRS-F.R.S).

References

1. 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.
2. S. Chari, C.S. Jutla, J.R. Rao, P. Rohatgi, *Towards Sound Approaches to Counteract Power Analysis Attacks*, CRYPTO 1999, LNCS, vol 1666, pp 398-412, Santa Barbara, California, USA, August 1999.
3. S. Chari, J. Rao, P. Rohatgi, *Template Attacks*, CHES 2002, LNCS, vol 2523, pp 13-28, Redwood Shores, California, USA, August 2002.
4. S. Faust, T. Rabin, L. Reyzin, E. Tromer, V. Vaikuntanathan, *Protecting Circuits from Leakage: the Computationally-Bounded and Noisy Cases*, Eurocrypt 2010, LNCS, vol 6110, pp 135-156, Nice, France, May 2010.
5. B. Gierlichs, L. Batina, P. Tuyls, B. Preneel, *Mutual Information Analysis - A Generic Side-Channel Distinguisher*, in the proceedings of CHES 2008, LNCS, vol 5154, pp 426-442, Washington DC,USA, August 2008.
6. B. Gierlichs, L. Batina, B. Preneel, I. Verbauwhede, *Revisiting Higher-Order DPA Attacks: Multivariate Mutual Information Analysis*, in the proceedings of CT-RSA 2010, LNCS, vol 5985, pp 221-234, San Francisco, USA, February 2010.
7. L. Goubin, J. Patarin, *DES and Differential Power Analysis*, CHES 1999, LNCS, vol 1717, pp 158-172, Worcester, MA, USA, August 1999.
8. P.C. Kocher, J. Jaffe, B. Jun, *Differential Power Analysis*, CRYPTO 1999, LNCS, vol 1999, pp 388-397, Santa Barbara, CA, USA, August 1999.
9. S. Mangard, T. Popp, B.M. Gammel, *Side-Channel Leakage of Masked CMOS Gates*, CT-RSA 2005, LNCS, vol 3376, pp 351-365, San Francisco, USA, February 2005.
10. M. Joye, P. Paillier, B. Schoenmakers, *On Second-Order Differential Power Analysis*, CHES 2005, LNCS, vol 3659, pp 293-308, Edinburgh, UK, September 2005.
11. S. Mangard, E. Oswald, T. Popp, *Power Analysis Attacks*, Springer, 2007.
12. S. Mangard, E. Oswald, F.-X. Standaert, *One for All, All for One: Unifying Standard DPA Attacks*, Cryptology ePrint Archive, Report 2009/449.
13. T.S. Messerges, *Using Second-Order Power Analysis to Attack DPA Resistant Software*, CHES 2000, LNCS, vol 1965, pp 238-251, Worcester, MA, August 2001.
14. E. Oswald, S. Mangard, *Template Attacks on Masking - Resistance Is Futile*, CT-RSA 2007, LNCS, vol 4377, pp 243-256, San Francisco, CA, USA, February 2007.
15. G. Piret, F.-X. Standaert, *Security Analysis of Higher-Order Boolean Masking Schemes for Block Ciphers (with Conditions of Perfect Masking)*, in IET Information Security, vol 2, num 1, pp 1-11, March 2008.
16. E. Prouff, M. Rivain, *A Generic Method for Secure S-box Implementation*, WISA 2007, LNCS, vol 4867, pp 227-244, Jeju Island, Korea, August 2007.
17. E. Prouff, M. Rivain, R. Bévan, *Statistical Analysis of Second Order DPA*, in IEEE Transactions on Computers, vol 58, num 6, pp 799-811, June 2009.
18. E. Prouff, M. Rivain, *Theoretical and Practical Aspects of Mutual Information Based Side-Channel Analysis*, in the proccedings of ACNS 2009, LNCS, vol 5536, pp 499-518, Paris, France June 2009.
19. K. Schramm, C. Paar, *Higher Order Masking of the AES*, CT-RSA 2006, LNCS, vol 3860, pp 208-225, San Jose, California, USA, February 2006.
20. F.-X. Standaert, T.G. Malkin, M. Yung, *A Unified Framework for the Analysis of Side-Channel Key Recovery Attacks*, in the proccedings of Eurocrypt 2009, LNCS, vol 5479, pp 443-461, Cologne, Germany, April 2009, extended version available on the Cryptology ePrint Archive, Report 2006/139, http://eprint.iacr.org/2006/139.

21. F.-X. Standaert, E. Peeters, C. Archambeau, J.-J. Quisquater, *Towards Security Limits in Side-Channel Attacks*, in the proceedings of CHES 2006, LNCS, vol 4249, pp. 30–45, Yokohama, Japan, October 2006, latest version available on the Cryptology ePrint Archive, Report 2007/222, http://eprint.iacr.org/2007/222.
22. 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, September 2009.
23. J. Waddle, D. Wagner, *Towards Efficient Second-Order DPA*, CHES 2004, LNCS, vol 3156, pp 1-15, Cambridge, MA, August 2004.

A Additional figures

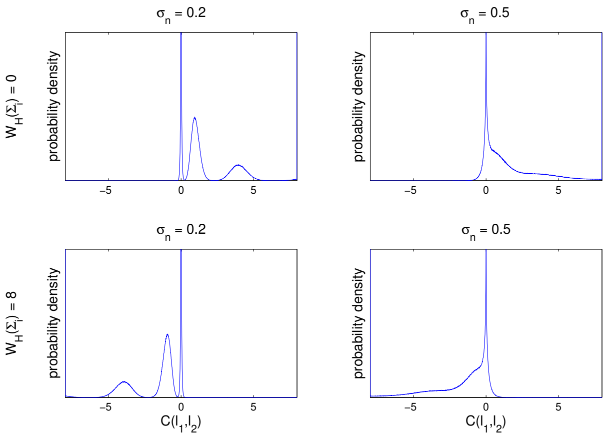

Fig. 9. Leakage probability distributions for the product combining function.

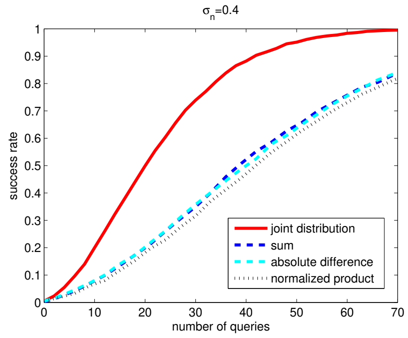

Fig. 10. Success rate of (simulated) profiled attacks against a masked AES S-box.

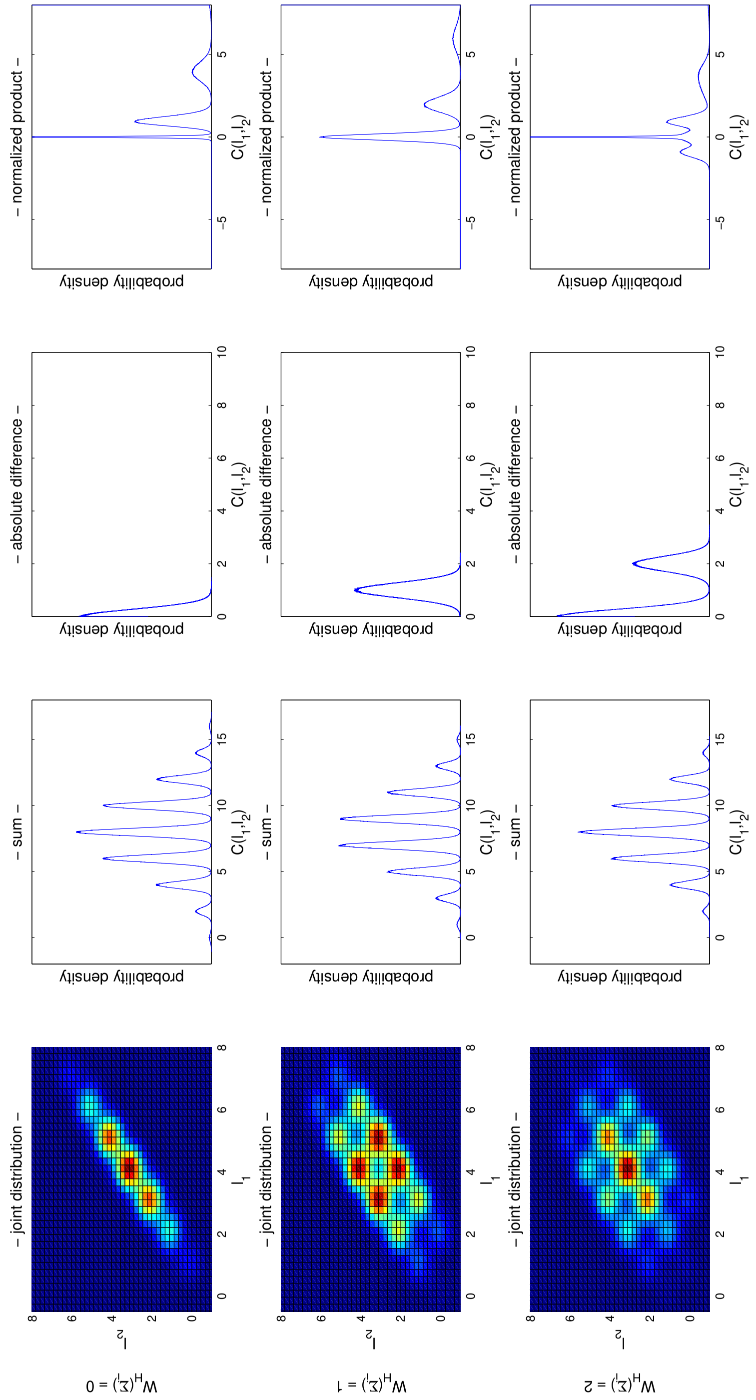

Fig. 11. Leakage probability distributions for three secret states ($\mathrm{W_H}(\Sigma_i) = 0, 1, 2$).

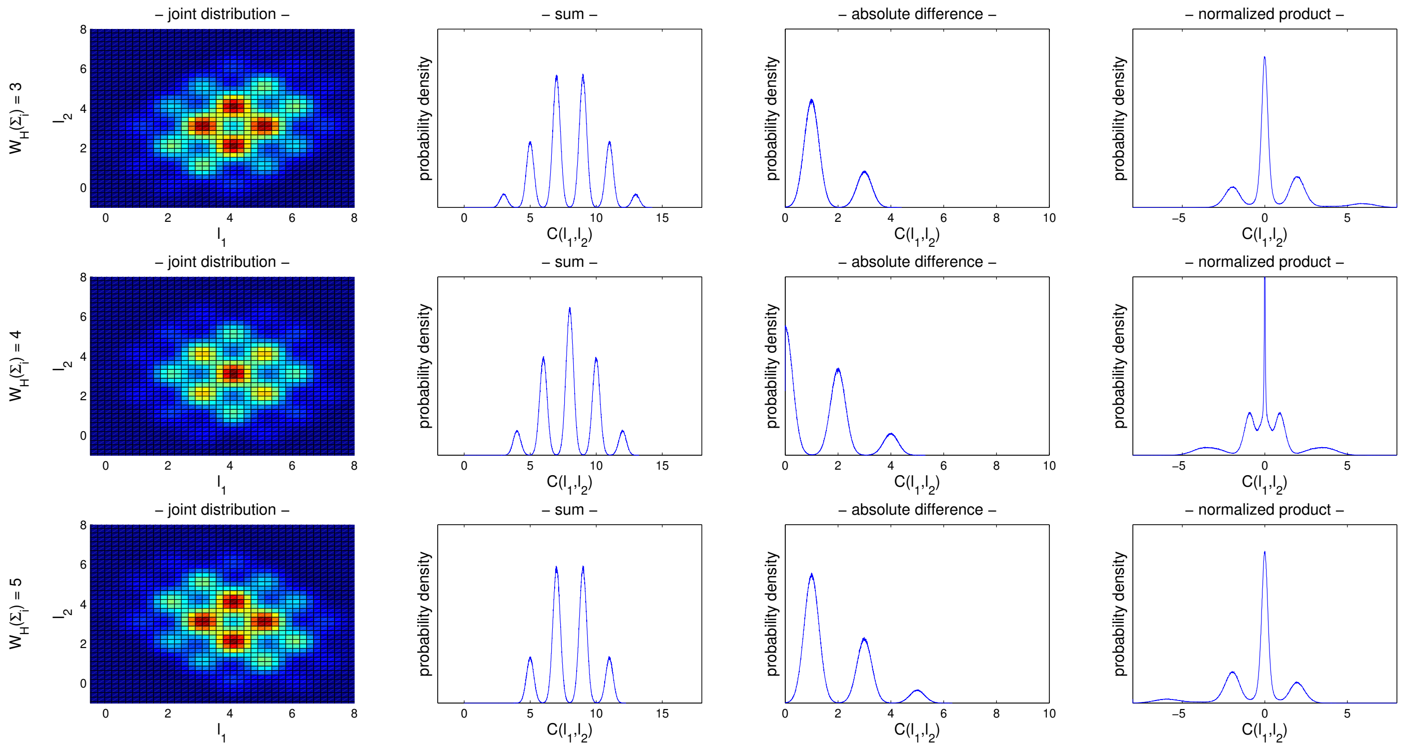

Fig. 12. Leakage probability distributions for three secret states ($W_H(\Sigma_i) = 3, 4, 5$).

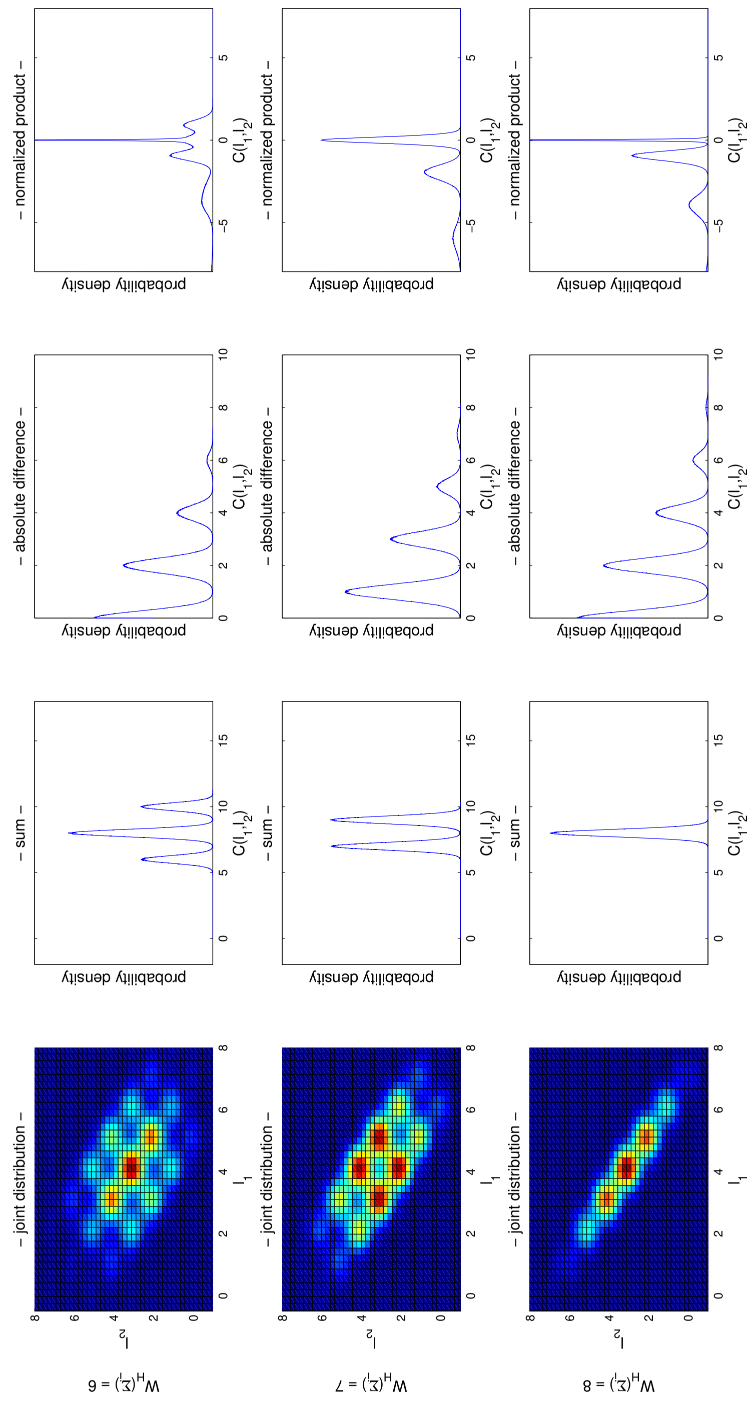

Fig. 13. Leakage probability distributions for three secret states ($\mathrm{W_H}(\Sigma_i) = 6, 7, 8$).