On the Optimality and Practicability of Mutual Information Analysis in Some Scenarios

Eloi de Chérisey[1], Sylvain Guilley[1,2], Annelie Heuser[1] and Olivier Rioul[1]

[1]Télécom ParisTech, LTCI, CNRS, Université Paris-Saclay, 75 013 Paris, France.
[2] Secure-IC, 15 rue Claude Chappe, Bâtiment B, 35 510 Cesson-Sévigné, France.

Abstract. The best possible side-channel attack maximizes the success rate and would correspond to a maximum likelihood distinguisher if the leakage probabilities were totally known or accurately estimated in a profiling phase. When profiling is unavailable, however, it is not clear whether Mutual Information Analysis (MIA), Correlation Power Analysis (CPA), or Linear Regression Analysis (LRA) would be the most successful in a given scenario. In this paper, we show that MIA coincides with the maximum likelihood expression when leakage probabilities are replaced by online estimated probabilities.
We then exhibit two case-studies where MIA outperforms CPA. One case is when the leakage model is known but the noise is not Gaussian. The second case is when the leakage model is partially unknown and the noise is Gaussian. In the latter scenario MIA is more efficient than LRA of any order.

Keywords: Side-channel analysis, unprofiled distinguishers, MIA, CPA, LRA, maximum likelihood.

1 Introduction

Many embedded systems implement cryptographic algorithms, which use secret keys that must be protected against extraction. Side-channel analysis (SCA) is one effective threat: physical quantities, such as instant power or radiated electromagnetic field, leak outside the embedded system boundary and reveal information about internal data. SCA consists in exploiting the link between the *leakage* signal and key-dependent internal data called *sensitive variables*.

The cryptographic algorithm is generally public information, whereas the implementation details are kept secret. For high-end security products, the confidentiality of the design is mandated by certification schemes, such as the Common Criteria [6]. For instance, to comply with `ALC_DVS` (Life-Cycle support – Development Security) requirement, the developer must provide a documentation that describes "*all the physical, procedural, personnel, and other security measures that are necessary to protect the confidentiality and integrity of the TOE (target of evaluation) design and implementation in its development environment*" [6, clause 2.1 C at page 141]. In particular, an attacker does not have enough information to precisely model the

leakage of the device. On commercial products, the attacker cannot set specific secret key values hence cannot profile the leakage. Therefore, many side-channel attacks can only be performed online using some distinguisher.

Correlation Power Analysis (CPA) [2] is one common side-channel distinguisher. It is known [11, Theorem 5] that its *optimality* holds only for a specific noise model (Gaussian) and for a specific knowledge of the deterministic part of the leakage—namely it should be perfectly known up to an unknown scaling factor and an unknown offset.

Linear Regression Analysis (LRA) [7] has been proposed in the context where the leakage model is drifting apart from a Hamming weight model. Its parametric structure and ability to include several basis functions makes it a very powerful tool, that can adjust to a broad range of leakage models when the additive noise is Gaussian. Incidentally, CPA may be seen as a 2-dimensional LRA [11].

When both model and noise are partially or fully unknown, generic distinguishers have been proposed, such as Mutual Information Analysis (MIA) [9], Kolmogorov-Smirnov test [20, 23] or Cramér-von-Mises test [20, Sec. 3.3.]. Thorough investigations have been carried out (e.g., [3, 13, 21]) to identify strengths and weaknesses of various distinguishers in various scenarios. In keeping with these results, we aim at showing some mathematical justification regarding MIA versus CPA and LRA.

Contributions. In this article, we derive MIA anew as the distinguisher which maximizes the success rate when the exact probabilities are replaced by online estimations. In order to assess the practicability of this mathematical result, we show two scenarios where MIA can outperform its competitors CPA and LRA, which themselves do not estimate probabilities. In these scenarios, we challenge the two hypotheses needed for CPA to be optimal: additive Gaussian noise and perfect knowledge of the model up to an affine transformation. This is illustrated in Fig. 1.

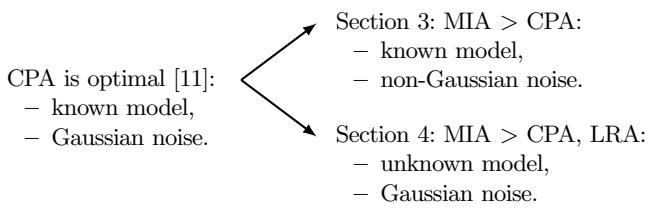


Fig. 1: Illustration of two practical situations where MIA can defeat CPA

Organization. The remainder of this paper is organized as follows. Section 2 provides notations, assumptions, and the rigorous mathematical derivation that MIA reduces to a maximum likelihood distinguisher, where exact leakage probabilities are replaced by online probabilities. Section 3 studies two examples where the attacker knows the one-bit model under non-Gaussian algorithmic noise, and for which MIA is shown to outperform CPA. Section 4 provides a scenario in which the leakage model is partially unknown under additive Gaussian noise, and where MIA outperforms CPA and LRA. Section 5 concludes.

2 Optimality of Mutual Information Analysis

2.1 Notations and Assumptions

We assume that the attacker has at his disposal $\tilde{q}$ independent *online* leakage measurements[1] $\tilde{\mathbf{x}}=(\tilde{x}_1,...,\tilde{x}_{\tilde{q}})$[2] for some sequence of independent and uniformly distributed text n-bit words $\tilde{\mathbf{t}}=(\tilde{t}_1,...,\tilde{t}_{\tilde{q}})$ (random but known). The n-bit secret key k^* is fixed but unknown.

We do not make any precise assumption on the leakage model—in particular the attacker is *not* able to estimate the actual probability density in a profiling phase. Instead we choose an algorithmic-specific function f and a device-specific function φ to compute, *for each key hypothesis* k, sensitive values $\tilde{\mathbf{y}}=(\tilde{y}_1,...,\tilde{y}_{\tilde{q}})$ by the formula

$$\tilde{\mathbf{y}}=\varphi(f(k,\tilde{\mathbf{t}})), \tag{1}$$

that is, $\tilde{y}_i=\varphi(f(k,\tilde{t}_i))$ for all $i=1,...,\tilde{q}$. In practice, a suitable choice of φ should be optimized depending on some leakage model but in what follows, f and φ can be taken arbitrarily such that they fulfill the following *Markov condition*.

Assumption 1 (Markov condition) *The leakage* $\tilde{\mathbf{x}}$ *depends on the actual secret key* k^* *only through the computed model* $\tilde{\mathbf{y}}=\varphi(f(k^*,\tilde{\mathbf{t}}))$.

Thus, while the conditional distribution $\mathbb{P}_k(\tilde{\mathbf{x}}|\tilde{\mathbf{t}})$ depends on the value k of the secret key, the expression $\mathbb{P}(\tilde{\mathbf{x}}|\tilde{\mathbf{y}})$ depends on k only through $\tilde{\mathbf{y}}=\varphi(f(k,\tilde{\mathbf{t}}))$. If we let $\mathbb{P}_k(\tilde{\mathbf{x}},\tilde{\mathbf{t}})$ be the joint probability distribution of $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{t}}$ when $k^*=k$, one has the Fisher factorization [4]

$$\mathbb{P}_k(\tilde{\mathbf{x}},\tilde{\mathbf{t}})=\mathbb{P}(\tilde{\mathbf{t}})\mathbb{P}_k(\tilde{\mathbf{x}}|\tilde{\mathbf{t}})=\mathbb{P}(\tilde{\mathbf{t}})\mathbb{P}(\tilde{\mathbf{x}}|\tilde{\mathbf{y}}) \qquad \text{where } \tilde{\mathbf{y}}=\varphi(f(k,\tilde{\mathbf{t}})). \tag{2}$$

In the latter expression we have $\mathbb{P}(\tilde{\mathbf{t}})=2^{-\tilde{q}n}$ since all text n-bit words are assumed independent and identically distributed (i.i.d.) and uniformly distributed.

In the case of an additive noise model, we simply have $\tilde{\mathbf{x}}=\tilde{\mathbf{y}}+\tilde{\mathbf{n}}$ where $\tilde{\mathbf{n}}$ is the noise vector, and the Markov condition is obviously satisfied. In general, in order to fulfill the Markov condition the attacker needs some knowledge on the actual leakage model. We give two examples regarding the Markov condition:

Example 1. If leakage x_i is linked to t_i and k^* through the relationship $x_i = w_H(k^* \oplus t_i) + n_i$ for all $i = 1,\ldots,\tilde{q}$, where w_H is the Hamming weight and n_i is the noise (independent of t_i), then both models $y_i=k\oplus t_i$ and $y_i=w_H(k\oplus t_i)$ satisfy the Markov condition[3]. Sections 3 and 4 give other, more sophisticated, examples that satisfy the Markov condition.

[1] We comply with the usual notations of [8] where offline quantities are indicated with a hat, whereas online quantities are indicated with a tilde. In this paper, there is no profiling phase hence no offline quantities.

[2] We use bold letters to indicate vectors while scalars are presented using small italic letters.

[3] In order to uniquely distinguish the correct key, some conditions on the expressions of y are required. Specifically, let us denote by y_k the function $t\mapsto y_k(t)=y(k,t)$, and let

Example 2. In the same scenario as in Example 1, consider the bit-dropping strategy (called `7LSB` in [9] and used in [18, 22]). Then e.g., $y_i = (k \oplus t_i)[1:7]$ (the first seven bit components) does *not* satisfy the Markov condition. Note that the leakage model in this exemple intentionally discards some information, hence may not be satisfactory [18].

Let $\tilde{k}$ be the key estimate that maximizes a distinguisher $\mathcal{D}$ given $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{t}}$, i.e.,

$$\tilde{k} = \arg\max_{k \in \mathcal{K}} \mathcal{D}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \tag{5}$$

where $\mathcal{K}$ is the key space.

We also assume that leakage values are quantized[4] in a suitable finite set $\mathcal{X}$. Letting $\mathcal{Y}$ denote the discrete sensitive value space, we have $\tilde{\mathbf{x}} \in \mathcal{X}^{\tilde{q}}$ and $\tilde{\mathbf{y}} \in \mathcal{Y}^{\tilde{q}}$. The actual probability densities being unknown, the attacker estimates them online, during the attack, from the available data in the sequences $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ (*via* $\tilde{\mathbf{t}}$), by counting all instances of possible values of $x \in \mathcal{X}$ and $y \in \mathcal{Y}$:

$$\tilde{\mathbb{P}}(x) = \frac{1}{\tilde{q}} \sum_{i=1}^{\tilde{q}} \mathbb{1}_{\tilde{x}_i = x}, \tag{6}$$

$$\tilde{\mathbb{P}}(y) = \frac{1}{\tilde{q}} \sum_{i=1}^{\tilde{q}} \mathbb{1}_{\tilde{y}_i = y}, \tag{7}$$

$$\tilde{\mathbb{P}}(x,y) = \frac{1}{\tilde{q}} \sum_{i=1}^{\tilde{q}} \mathbb{1}_{\tilde{x}_i = x, \tilde{y}_i = y}, \tag{8}$$

$$\tilde{\mathbb{P}}(x|y) = \frac{\sum_{i=1}^{\tilde{q}} \mathbb{1}_{\tilde{x}_i = x, \tilde{y}_i = y}}{\sum_{i=1}^{\tilde{q}} \mathbb{1}_{\tilde{y}_i = \tilde{y}}} = \frac{\tilde{\mathbb{P}}(x,y)}{\tilde{\mathbb{P}}(y)}, \tag{9}$$

where $\mathbb{1}_A$ denotes the indicator function of A: $\mathbb{1}_A = 1$ if A is true and $= 0$ otherwise.

$\mathscr{B}$ the set of bijections on the leakage space $\mathcal{X}$. We have:

$$\text{if } \forall k, \exists k' \neq k, \quad \exists \beta \in \mathscr{B} \text{ s.t. } y_{k'} = \beta \circ y_k, \quad \text{then the distinguisher features a } \textit{tie}, \tag{3}$$

$$\text{if } \forall k, \forall k' \neq k, \quad \exists \beta \in \mathscr{B} \text{ s.t. } y_{k'} = \beta \circ y_k, \quad \text{then the distinguisher is } \textit{not sound}. \tag{4}$$

Indeed, in Eq. (3), there is no way for the distinguisher to tell k^* from k', and in Eq. (4), the distinguisher yields the same value for all the key guesses.

We refer the interested reader to the work done in [24, Sec. 3]. We note that $y_i = k \oplus t_i$ does not lead to a sound distinguisher, as for all k', $x \mapsto x \oplus k'$ is bijective, and maps y_k to $y_{k \oplus k'}$. On the contrary, there is no bijection β such that for all t, $w_H(k \oplus t) = \beta(w_H(k \oplus k' \oplus t))$. So the choice $y_i = w_H(k \oplus t_i)$ is sound.

[4] Some side-channels are discrete by nature, such as the timing measurements (measured in units of clock period). In addition, oscilloscopes or data acquisition appliances rely on ADCs (Analog to Digital Converters), which usually sample a continuous signal into a sequence of integers, most of the time represented on 8 bits (hence $\mathcal{X} = \mathbb{F}_2^8$).

Definition 1 (Empirical Mutual Information). *The empirical mutual information is defined as*

$$\tilde{I}(\tilde{\mathbf{x}},\tilde{\mathbf{y}}) = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \tilde{\mathbb{P}}(x,y)\ \log_2 \frac{\tilde{\mathbb{P}}(x,y)}{\tilde{\mathbb{P}}(x)\tilde{\mathbb{P}}(y)}, \tag{10}$$

which can also be written as

$$\tilde{I}(\tilde{\mathbf{x}},\tilde{\mathbf{y}}) = \tilde{H}(\tilde{\mathbf{x}}) - \tilde{H}(\tilde{\mathbf{x}}|\tilde{\mathbf{y}}), \tag{11}$$

where the empirical entropies are defined as

$$\tilde{H}(\tilde{\mathbf{x}}) = \sum_{x\in\mathcal{X}} \tilde{\mathbb{P}}(x)\ \log_2 \frac{1}{\tilde{\mathbb{P}}(x)} \tag{12}$$

and

$$\tilde{H}(\tilde{\mathbf{x}}|\tilde{\mathbf{y}}) = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \tilde{\mathbb{P}}(x,y)\ \log_2 \frac{1}{\tilde{\mathbb{P}}(x|y)}. \tag{13}$$

These quantities are functions of the sequences $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ since $\tilde{\mathbb{P}}(x,y)$ is a function of $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$. They also depend on the key guessed value k, *via* the expression of $\tilde{\mathbf{y}}$.

2.2 Mathematical Derivation

In this subsection, we show that MIA coincides with the maximum likelihood expression where leakage probabilities $\mathbb{P}$ are replaced by online estimated probabilities $\tilde{\mathbb{P}}$.

Definition 2 (Success Rate [19, Sec. 3.1]). *The* success rate *(averaged over all possible secret key values) is defined as:*

$$SR = \frac{1}{2^n} \sum_{k=0}^{2^n-1} \mathbb{P}_k(\tilde{k}=k). \tag{14}$$

Here we follow a frequentist approach. An equivalent alternative Bayesian approach would be to assume a uniform prior key distribution [11].

Theorem 3 (Maximum Likelihood [5]). *Let* $\tilde{\mathbf{y}} = \varphi(f(k,\tilde{\mathbf{t}}))$. *The* optimal key estimate *that maximizes the success rate* (14) *is:*

$$\tilde{k} = \arg\max_k\ \mathbb{P}(\tilde{\mathbf{x}}|\tilde{\mathbf{y}}). \tag{15}$$

Proof. We give here a formal proof, which nicely relates to Definition 2. Straightforward computation yields:

$$\mathrm{SR} = \frac{1}{2^n}\sum_{k=1}^{2^n}\sum_{\tilde{\mathbf{x}},\tilde{\mathbf{t}}} \mathbb{P}_k(\tilde{\mathbf{x}},\tilde{\mathbf{t}})\ \mathbb{1}_{k=\tilde{k}} \tag{16}$$

$$= \frac{1}{2^n}\sum_{k=1}^{2^n}\sum_{\tilde{\mathbf{x}},\tilde{\mathbf{t}}} \mathbb{P}(\tilde{\mathbf{x}}|\tilde{\mathbf{y}}=\varphi(f(k,\tilde{\mathbf{t}})))\ \mathbb{P}(\tilde{\mathbf{t}})\ \mathbb{1}_{k=\tilde{k}} \quad \text{(by (2) \& Assumption 1)} \tag{17}$$

$$= \frac{1}{2^{n(\tilde{q}+1)}}\sum_{k=1}^{2^n}\sum_{\tilde{\mathbf{x}},\tilde{\mathbf{t}}} \mathbb{P}(\tilde{\mathbf{x}}|\tilde{\mathbf{y}}=\varphi(f(k,\tilde{\mathbf{t}})))\ \mathbb{1}_{k=\tilde{k}} \tag{18}$$

$$= \frac{1}{2^{n(\tilde{q}+1)}}\sum_{\tilde{\mathbf{x}},\tilde{\mathbf{t}}} \mathbb{P}(\tilde{\mathbf{x}}|\tilde{\mathbf{y}}=\varphi(f(k=\tilde{k},\tilde{\mathbf{t}}))). \tag{19}$$

Thus, for each given sequences $\tilde{\mathbf{x}},\tilde{\mathbf{t}}$ maximizing the success rate amounts to choosing $k=\tilde{k}$ so as to maximize $\mathbb{P}(\tilde{\mathbf{x}}|\tilde{\mathbf{y}})=\mathbb{P}(\tilde{\mathbf{x}}|\tilde{\mathbf{y}}=\varphi(f(k=\tilde{k},\tilde{\mathbf{t}})))$:

$$\tilde{k} = \arg\max_k \mathbb{P}(\tilde{\mathbf{x}}|\tilde{\mathbf{y}}). \tag{20}$$

□

When no profiling is possible the conditional distribution

$$\mathbb{P}(\tilde{\mathbf{x}}|\tilde{\mathbf{y}}) = \prod_{i=1}^{\tilde{q}} \mathbb{P}(\tilde{x}_i|\tilde{y}_i) \tag{21}$$

is unknown to the attacker. Therefore, Theorem 3 is no longer practical and we require a *universal*[5] version of it.

Definition 3 (Universal Maximum Likelihood). *Let $\tilde{\mathbf{y}} = \varphi(f(k,\tilde{\mathbf{t}}))$. The universal maximum likelihood (UML) key estimate is defined by*

$$\tilde{k} = \arg\max_k \tilde{\mathbb{P}}(\tilde{\mathbf{x}}|\tilde{\mathbf{y}}), \tag{22}$$

where

$$\tilde{\mathbb{P}}(\tilde{\mathbf{x}}|\tilde{\mathbf{y}}) = \prod_{i=1}^{\tilde{q}} \tilde{\mathbb{P}}(\tilde{x}_i|\tilde{y}_i). \tag{23}$$

Here $\tilde{\mathbb{P}}$, defined in Equations (9), (8), (7) and (6), is estimated directly from the available data, that is, from the actual values in the sequences $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$.

Theorem 4 (UML is MIA). *The universal maximum likelihood key estimate is equivalent to the mutual information analysis (MIA) [9]:*

$$\tilde{k} = \arg\max_k \tilde{\mathbb{P}}(\tilde{\mathbf{x}}|\tilde{\mathbf{y}}) = \arg\max_k \tilde{I}(\tilde{\mathbf{x}},\tilde{\mathbf{y}}), \tag{24}$$

where $\tilde{I}(\tilde{\mathbf{x}},\tilde{\mathbf{y}})$ is the universal mutual information (definition 1).

[5] *Universal*, in the information theoretic sense of the word, means: computed from the available data without prior information.

Proof. Rearrange the likelihood product according to values taken by the $\tilde{x}_i$ and $\tilde{y}_i$'s:

$$\tilde{\mathbb{P}}(\tilde{\mathbf{x}}|\tilde{\mathbf{y}}) = \prod_{i=1}^{\tilde{q}} \tilde{\mathbb{P}}(\tilde{x}_i|\tilde{y}_i) = \prod_{x\in\mathcal{X},y\in\mathcal{Y}} \tilde{\mathbb{P}}(x|y)^{\tilde{n}_{x,y}} \tag{25}$$

where $\tilde{n}_{x,y}$ is the number of components $(\tilde{x}_i,\tilde{y}_i)$ equal to (x,y), i.e.,

$$\tilde{n}_{x,y} = \sum_{i=1}^{\tilde{q}} \mathbb{1}_{\tilde{x}_i=x,\tilde{y}_i=y} = \tilde{q}\tilde{\mathbb{P}}(x,y). \tag{26}$$

The second inequality in Eqn. (25) is based on a counting argument: some events collide, i.e., we have $(x_i,y_i) = (x_{i'},y_{i'})$ for $i \neq i'$. The exponent $\tilde{n}_{x,y}$ is meant to enumerate all such possible collisions. This gives

$$\tilde{\mathbb{P}}(\tilde{\mathbf{x}}|\tilde{\mathbf{y}}) = \prod_{x\in\mathcal{X},y\in\mathcal{Y}} \tilde{\mathbb{P}}(x|y)^{\tilde{q}\tilde{\mathbb{P}}(x,y)} = 2^{-\tilde{q}\tilde{H}(\tilde{\mathbf{x}}|\tilde{\mathbf{y}})}, \tag{27}$$

(see Definition 1). Therefore, maximizing $\tilde{\mathbb{P}}(\tilde{\mathbf{x}}|\tilde{\mathbf{y}})$ amounts to minimizing the empirical conditional entropy $\tilde{H}(\tilde{\mathbf{x}}|\tilde{\mathbf{y}})$. Since $\tilde{H}(\tilde{\mathbf{x}})$ is key-independent, this in turn amounts to maximizing the empirical mutual information $\tilde{I}(\tilde{\mathbf{x}},\tilde{\mathbf{y}}) = \tilde{H}(\tilde{\mathbf{x}}) - \tilde{H}(\tilde{\mathbf{x}}|\tilde{\mathbf{y}})$. □

From Theorem 4 we can conclude that MIA is "optimal" as a universal maximum likelihood estimation. This constitutes a rigorous proof that mutual information is a relevant tool for key recovery when the leakage is unknown (in the case where the model satisfies the Markov condition) as was already hinted in [9, 16, 18, 21].

Corollary 5. *MIA coincides with the ML distinguisher as* $\tilde{q} \to \infty$.

Proof. By the law of large numbers, the online probability $\tilde{\mathbb{P}}$ converges almost surely to the exact probability of the leakage as $\tilde{q} \to \infty$. For any fixed values of $\tilde{x} \in \mathcal{X}, \tilde{y} \in \mathcal{Y}$,

$$\tilde{\mathbb{P}}(\tilde{x}|\tilde{y}) \underset{\tilde{q}\to\infty}{\longrightarrow} \mathbb{P}(\tilde{x}|\tilde{y}) \qquad a.s.$$

Thus in the limit, MIA coincides with the maximum likelihood rule. □

Remark 1. It is well known [16] that if the mapping $\tilde{t} \mapsto \tilde{y} = \varphi(f(k,\tilde{t}))$ is one-to-one (for all values of k), then MIA cannot distinguish the correct key. This is also clear from Eq. (4) in footnote 3: given two different keys k,k', there is a bijection between y_k and $y_{k'}$, which is simply $\beta = y_{k'} \circ y_k^{-1}$. In our present setting this is easily seen by noting that when $\tilde{y} = \varphi(f(k,\tilde{t}))$,

$$\tilde{\mathbb{P}}(x|y) = \frac{\sum_{i=1}^{\tilde{q}} \mathbb{1}_{\tilde{x}_i=x,\tilde{y}_i=y}}{\sum_{i=1}^{\tilde{q}} \mathbb{1}_{\tilde{y}_i=y}} = \frac{\sum_{i=1}^{\tilde{q}} \mathbb{1}_{\tilde{x}_i=x,\tilde{t}_i=t}}{\sum_{i=1}^{\tilde{q}} \mathbb{1}_{\tilde{t}_i=t}} \tag{28}$$

is independent of the value k. Note that this is true for any fixed number of measurements $\tilde{q}$ during the attack.

3 Non-Gaussian Noise Challenge

In this section, we show two examples where MIA outperforms CPA due to non-Gaussian noise. The first example presented in subsection 3.1 is an academic (albeit artificial) example built in order to have the success rate of CPA collapse. The second example in subsection 3.2 is more practical.

3.1 Pedagogical Case-study

We consider a setup where the variables are $X=Y+N$, with $Y=\varphi(f(k^*,T))$, where $Y\in\{\pm 1\}$, and $N\sim\mathcal{U}(\{\pm\sigma\})$ (meaning that N takes values $-\sigma$ and $+\sigma$ randomly, with probabilities $\frac{1}{2}$ and $\frac{1}{2}$), where σ is an integer. Specifically, we assume that k^*, $t\in\mathbb{F}_2^n$, with $n=4$, and that $f:\mathbb{F}_2^n\times\mathbb{F}_2^n\to\mathbb{F}_2^m$ is a (truncated version) of the SERPENT Sbox[6] fed by the XOR of the two inputs (key and plaintext nibbles) and $\varphi=w_H$ is the Hamming weight (which reduces to the identity $\mathbb{F}_2\to\mathbb{F}_2$ if $m=1$ bit).

The optimal distinguisher (Theorem 3) in this scenario has the following closed-form expression:

$$\mathcal{D}(\tilde{\mathbf{x}},\tilde{\mathbf{t}})=\arg\max_k \mathbb{P}(\tilde{\mathbf{x}}|\tilde{\mathbf{t}},k)=\arg\max_k \frac{1}{2^{\tilde{q}}}\prod_{i=1}^{\tilde{q}}\delta(\tilde{x}_i,\tilde{t}_i,k), \tag{29}$$

where $\delta:\mathbb{F}_2^m\times\mathbb{F}_2^n\times\mathbb{F}_2^n\to\{0,1\}$ is defined as:

$$\delta(x,t,k)=\begin{cases}1 & \text{if } x-\varphi(f(k,t))=-\sigma,\\ 1 & \text{if } x-\varphi(f(k,t))=\sigma,\\ 0 & \text{otherwise.}\end{cases}$$

The evaluation of this quantity requires the knowledge of σ, which by definition is an unknown quantity related to the noise. Our simulations have been carried out as follows.

1. Generate two large uniformly distributed random vectors $\tilde{\mathbf{t}}$ and $\tilde{\mathbf{n}}$ of length $\tilde{q}$;
2. Deliver the pair of vectors $(\tilde{\mathbf{t}},\tilde{\mathbf{x}}=\varphi(f(k^*,\tilde{\mathbf{t}}))+\tilde{\mathbf{n}})$ to the attacker;
3. Estimate averages and PMFs (probability mass functions) of this data for $\tilde{q}_{\mathsf{step}}$ $(=1)$, then for $2\tilde{q}_{\mathsf{step}}$, $3\tilde{q}_{\mathsf{step}}$ and so on;
4. At each multiple of $\tilde{q}_{\mathsf{step}}$, carry out CPA and MIA.

The attacks are reproduced 100 times to allow for narrow error bars on the estimated success rate.

Remark 2. We do not consider *linear regression analysis* because the model is not parametric. The only unknown parameter is related to the noisy part of the leakage, not its deterministic part.

[6] The least significant bit S_0 of the PRESENT Sbox S is not suitable because one has $\forall z\in\mathbb{F}_2^4$, $S_0(z)=S_0(z\oplus\mathtt{0x9})=\neg S_0(z\oplus\mathtt{0x1})=\neg S_0(z\oplus\mathtt{0x8})$. As in Eq. (3) of footnote 3, *ties* occur: it is not possible to distinguish k^*, $k^*\oplus\mathtt{0x9}$, $k^*\oplus\mathtt{0x1}$, $k^*\oplus\mathtt{0x8}$ (the corresponding bijections are respectively $x\mapsto x$ and $x\mapsto 1-x$). Therefore, we consider component 1 instead of 0, which does not satisfy such relationships.

Simulation results are given in Fig. 2 for $\sigma = 2$ and $\sigma = 4$. The success rate of the "optimal" distinguisher (the maximum likelihood distinguisher of Theorem 3 – see Eqn. (29)) is drawn in order to visualize the limit between feasible (below) and unfeasible (above) attacks. It can be seen that MIA is almost as successful as the maximum likelihood distinguisher, despite the knowledge of the value of σ is not required for the MIA. In addition, one can see that the CPA performs worse, and all the worst as σ increases. In this case, the CPA is not the optimal distinguisher (as e.g., underlined in [11, Theorem 5]) since the noise is not Gaussian (but discrete).

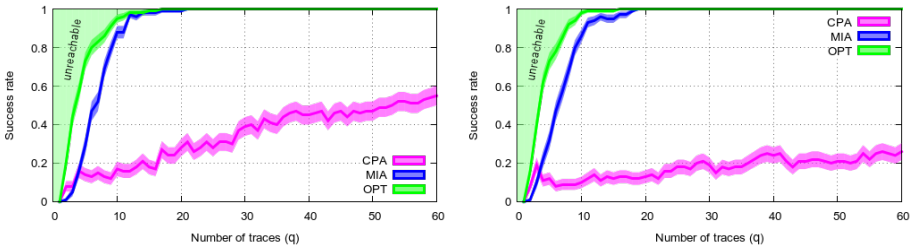


Fig. 2: Success rate for $\sigma = 2$ (left) and $\sigma = 4$ (right), when $Y \sim \mathcal{U}(\{\pm 1\})$ and $N \sim \mathcal{U}(\{\pm\sigma\})$

Remark 3. Another attack strategy for the leakage model presented in this subsection would simply be to filter out the noise. One could for instance dispose of all traces where the leakage is negative. The remaining traces (half of them) contain a constant noise $N = +\sigma > 1$, hence the signal Y can be read out without noise. Such attack, known as the *subset attack* [14, Sec. 5.2], is not far from the optimal one (Eqn. (29)). It actually does coincide with the optimal attack if the attacker recovers Y from both subsets $\{i/X_i > 0\}$ and $\{i/X_i < 0\}$. Still it can noted that MIA is very close to being optimal for this scenario.

Asymptotics. We can estimate the theoretical quantities for CPA and MIA as follows. We have $\mathsf{Var}(Y) = 1$ and $\mathsf{Var}(N) = \sigma^2$, hence a signal to noise ratio $\mathrm{SNR} = 1/\sigma^2$. In addition, X can only take four values: $\pm 1 \pm \sigma$. Since $\mathbb{E}(XY) = \mathbb{E}(X^2) + \mathbb{E}(YN) = Var(X) + \mathbb{E}(Y)\mathbb{E}(N) = 1 + 0 \times 0 = 1$, the correlation is simply $\rho(X,Y) = 1/\sigma$, which vanishes as σ increases.

However, for $\sigma > 1$, the mutual information $\mathsf{I}(X,Y) = 1$ bit. Indeed, $\mathsf{H}(X) = -\sum_{x \in \{\pm 1 \pm \sigma\}} \mathbb{P}(X = x) \log_2 \mathbb{P}(X = x) = -\sum_{x \in \{\pm 1 \pm \sigma\}} \frac{1}{4} \log_2 \frac{1}{4} = \log_2 4 = 2$ bit, $\mathsf{H}(X|y = \pm 1) = \log_2 2 = 1$ bit, so $\mathsf{I}(X,Y) = \log_2 4 - \sum_{y \in \{\pm 1\}} \mathbb{P}(X = x) \log_2 2 = \log_2 4 - \log_2 2 = 1$ bit, irrespective of $\sigma \in \mathbb{N}$.

The important fact is that the mutual information does not depend on the value of σ. Accordingly, it can be seen from Fig. 2 that the success rate of the MIA is not affected by the noise variance. This explains why MIA will outperform the CPA for large enough σ.

3.2 Application to Bitslice PRESENT

Bitslicing algorithms is a common practice. This holds both for standard [17] (e.g., AES) and lightweight [12] (PRESENT, Piccolo) block ciphers. Here the distinguishers must be single-bit: $Y \in \{\pm 1\}$. However, compared to the case of Sec. 3.1, the noise takes now more than two values: On an 8-bit computer, the 7 other bits will leak independently. They are, however, not concerned by the attack, and constitute *algorithmic noise* N which follows a binomial law $\alpha \times \mathcal{B}(7, \frac{1}{2})$, where α is a scaling factor.

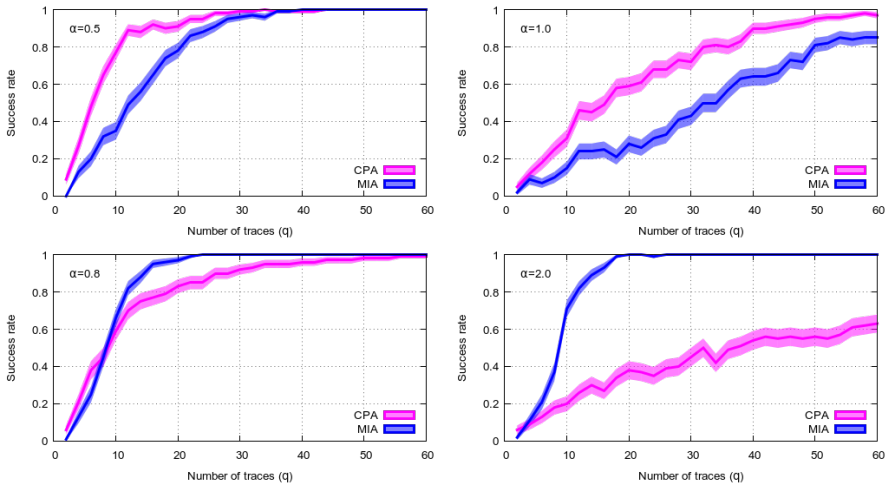


Fig. 3: Success rate for the attack of a bitsliced algorithm on an 8-bit processor, where 7 bits make up algorithmic noise, and have weight 0.5, 1.0 (top) and 0.8 and 2.0 (bottom).

Simulation results for various values of α are in Fig. 3. Interestingly, MIA is efficient for the cases where the leakage $Y \sim \mathcal{U}(\{\pm 1\})$ is not altered by the addition of noise: For $\alpha = 0.8$ and $\alpha = 2.0$, it is still possible to tell unambiguously from X what is the value of Y. On the contrary, when $\alpha = 0.5$ or $\alpha = 1.0$, the function $(Y,N) \mapsto X = Y + N$ is not one-to-one. For instance, in the case $\alpha = 1.0$, the value $X = 2$ can result as well from $Y = -1$ and $N = 3$, or $Y = +1$ and $N = 2$. (see Fig. 4).

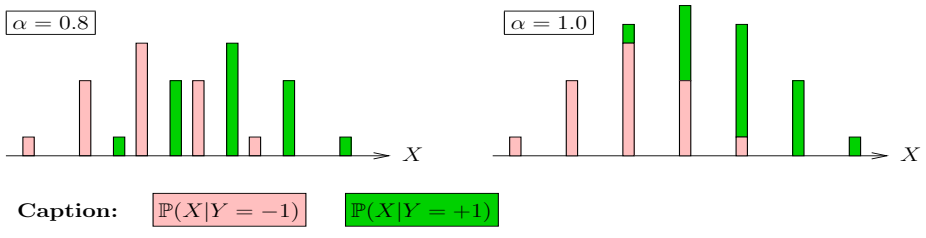


Fig. 4: Illustration bijectivity (left) vs. non-injectivity (right) of the leakage function.

4 Partially Unknown Model Challenge

Veyrat-Charvillon and Standaert [20, section 4] have already noticed that MIA can outperform CPA if the model is drifted too far away from the real leakage. However, LRA is able to make up for the model drift of [20] (which considered unevenly weighted bits). In this section, we challenge CPA and LRA with a partially unknown model. We show that, in our example, MIA has a much better success rate than both CPA and LRA.

For our empirical study we used the following setup:

$$X = \psi(Y(k^*)) + N, \qquad Y(k^*) = w_H(\mathrm{Sbox}(k^* \oplus T)),$$

where Sbox is the AES substitution box, ψ is the non-linear function given by:

x	0	1	2	3	4	5	6	7	8
$\psi(x)$	$+1$	$+2$	$+3$	$+4$	0	-4	-3	-2	-1

which is unknown to the attacker, and N is a centered Gaussian noise with unknown standard deviation σ. The non-linearity of ψ is motivated by [13], where it is discussed that a linear model favors CPA over MIA.

The leakage is continuous due to the Gaussian noise. In order to discretize the leakage to obtain discrete probabilities, we used the *binning* method. We conducted MIA with several different binning sizes:

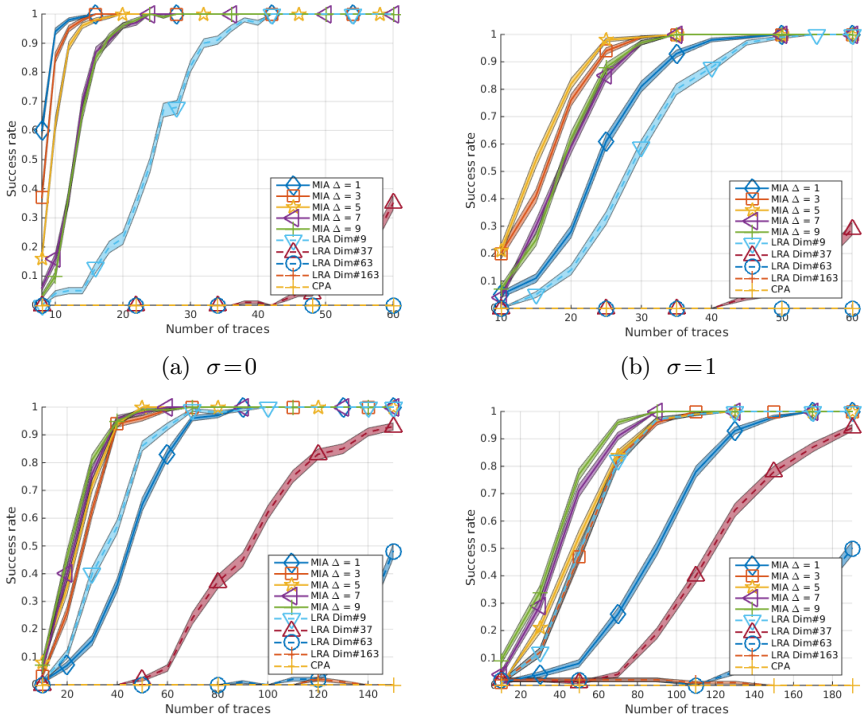
$$B = \{[(i-1) \times \Delta x, i \times \Delta x[, i \in \mathbb{Z}\} \qquad \text{for } \Delta x = \{1,3,5,7,9\}. \tag{30}$$

In this paper, we do not try to establish any specific result about binning, but content ourselves to present empirical results obtained with different bin sizes.

We have carried out LRA for the standard basis in dimension $d = 9$ and higher dimensions $d = \{37, 93, 163\}$. More precisely, for $d = 9$ we have $\tilde{\mathbf{y}}'(k) = (\mathbf{1}, \tilde{\mathbf{y}}_1(k), \tilde{\mathbf{y}}_2(k), \ldots, \tilde{\mathbf{y}}_8(k))$ with $\tilde{\mathbf{y}}_j(k) = [\mathrm{Sbox}(k \oplus T)]_j$ where $[\cdot]_j : \mathbb{F}_2^n \to \mathbb{F}_2$ is the projection mapping onto the j^{th} bit. For $d=37$ the attacker additionally takes into consideration the products between all possible $\tilde{\mathbf{y}}_j$ $(1 \le j \le 8)$, i.e., $\tilde{\mathbf{y}}_1 \cdot \tilde{\mathbf{y}}_2, \tilde{\mathbf{y}}_1 \cdot \tilde{\mathbf{y}}_3, \tilde{\mathbf{y}}_1 \cdot \tilde{\mathbf{y}}_4$ and so on. Consequently, $d=93$ considers additionally the product between 3 $\tilde{\mathbf{y}}$'s and $d=163$ includes also all possible product combinations with 4 columns. See [10] for a detailed description on the selection of basis functions.

Fig. 5 shows the success rate using 100 independent experiments. Perhaps surprisingly, MIA turns out to be more efficient than LRA. Quite naturally, MIA and LRA become closer as the the variance of the independent measurement noise N increases. It can be seen that LRA using higher dimension requires a sufficient number of traces for estimation (for $d=37$ around 100, $d=93$ around 150, and $d=137$ failed below 200 traces). Consequently, in this scenario using high dimensions is not appropriate, even if the high dimension in question might fit the unknown function ψ.

One reason why MIA outperforms CPA and LRA in this scenario is that the function ψ was chosen to have a null covariance. Moreover, one can observe that the most efficient binning size depends on the noise variance and thus on the scattering of the leakage. As σ grows larger values of Δ should be chosen. This is contrary to the suggestions made in [9], which proposes to estimate the probability distributions as good as possible



(a) $\sigma=0$ (b) $\sigma=1$

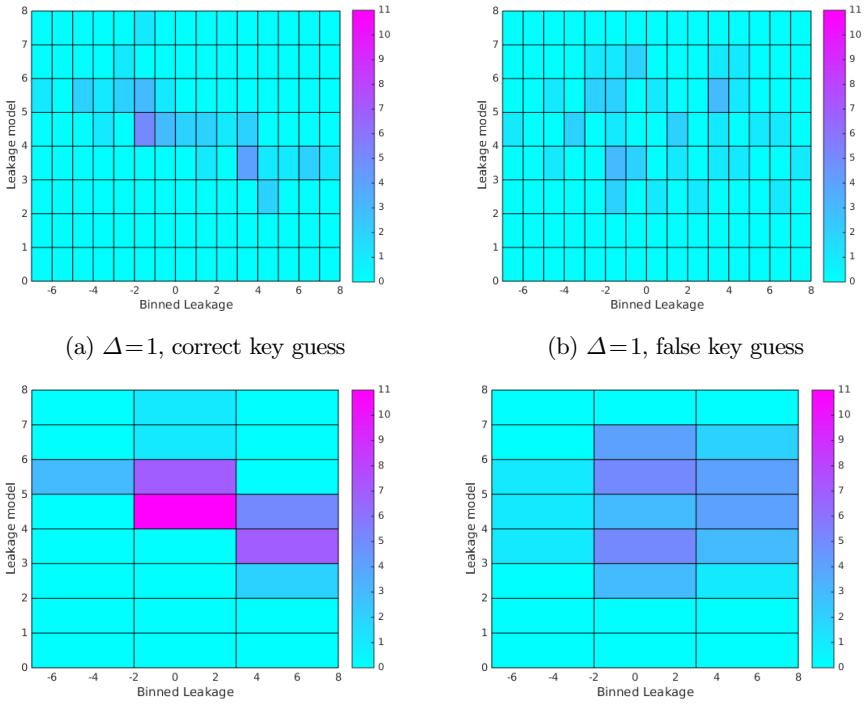
(c) $\sigma=2$ (d) $\sigma=3$

Fig. 5: Success rate for $\sigma\in\{0,1,2,3\}$ when the model is unknown

and thus to consider as many bins as there are distinct values in the traces. In our experiments, when noise is absent ($\sigma=0$) the optimal binning size is $\Delta=1$ which is equivalent to the step size of Y, while for $\sigma=2$ the optimal binning is $\Delta=5$ (see Fig. 5(c)).

It can be seen that using 40 traces the success rate of MIA with $\Delta=5$ reaches 90%, whereas using $\Delta=1$ it is only about 30%. To understand this phenomenon, Fig. 6 displays the estimated $\tilde{\mathbb{P}}(x|y)$ in a 3D histogram for the correct key and one false key hypothesis, such that MIA is able to reveal the correct key using $\Delta=5$ but fails for $\Delta=1$. Clearly, the distinguishability between the correct and false key is much higher in case of $\Delta=5$ than for $\Delta=1$.

More precisely, as the leakage is dispersed by the noise the population of bins of the false key becomes similar to the the ones of the correct key when considering smaller binning size (compare Fig. 6a and 6b). In contrast, the difference is more visible when the leakage is quantified into larger bins (compare Fig. 6c and 6d). Therefore, even if the estimation of $\tilde{I}(\tilde{\mathbf{x}},\tilde{\mathbf{y}})$ using $\tilde{\mathbb{P}}(x|y)$ for larger Δ is more coarse and thus looses some information, the distinguishing ability to reveal the correct key is enhanced.



(a) $\Delta = 1$, correct key guess

(b) $\Delta = 1$, false key guess

(c) $\Delta = 5$, correct key guess

(d) $\Delta = 5$, false key guess

Fig. 6: Estimated $\tilde{\mathbb{P}}(X|Y)$ using 40 traces for $\sigma = 2$ (see Fig. 5(c))

5 Conclusion

We derived MIA anew as the distinguisher which maximizes the success rate when the exact probabilities are replaced by online estimations. This suggests that MIA is an interesting alternative when the attacker is not able to exactly determine the link between the measured leakage and the leakage model. This situation can either result from an unknown deterministic part or from an unknown noise distribution.

We have presented two practical case-studies in which MIA can indeed be more efficient than CPA or LRA. The first scenario is for non-Gaussian noise but known deterministic leakage model. The second scenario is for Gaussian noise with unknown deterministic leakage model, where one leverages a challenging leakage function which results in failure for CPA, and in harsh regression using LRA. Incidentally, this example is in line with the work carried out by Whitnall and Oswald [21] where a notion of relative margin is used to compare attacks. Our findings go in the same direction using the success rate as a figure of merit to compare attacks.

We note that all our results are φ-dependent. It seems obvious that the closer we are from the actual leakage, the better the success rate will be. An open question is

to find an analytic way to determine the function model that will provide the highest success rate.

References

1. Lejla Batina and Matthew Robshaw, editors. *Cryptographic Hardware and Embedded Systems - CHES 2014 - 16th International Workshop, Busan, South Korea, September 23-26, 2014. Proceedings*, volume 8731 of *Lecture Notes in Computer Science*. Springer, 2014.
2. Éric Brier, Christophe Clavier, and Francis Olivier. Correlation Power Analysis with a Leakage Model. In *CHES*, volume 3156 of *LNCS*, pages 16–29. Springer, August 11–13 2004. Cambridge, MA, USA.
3. Mathieu Carbone, Sébastien Tiran, Sébastien Ordas, Michel Agoyan, Yannick Teglia, Gilles R. Ducharme, and Philippe Maurine. On adaptive bandwidth selection for efficient MIA. In Prouff [15], pages 82–97.
4. George Casella and Roger L. Berger. *Statistical Inference*. Duxbury Press, 2002. Second edition. ISBN-10: 0534243126 – ISBN-13: 978-0534243128.
5. Suresh Chari, Josyula R. Rao, and Pankaj Rohatgi. Template Attacks. In *CHES*, volume 2523 of *LNCS*, pages 13–28. Springer, August 2002. San Francisco Bay (Redwood City), USA.
6. Common Criteria Consortium. Common Criteria (*aka* CC) for Information Technology Security Evaluation (ISO/IEC 15408), 2013.
   Website: `http://www.commoncriteriaportal.org/`.
7. Julien Doget, Emmanuel Prouff, Matthieu Rivain, and François-Xavier Standaert. Univariate side channel attacks and leakage modeling. *J. Cryptographic Engineering*, 1(2):123–144, 2011.
8. François Durvaux, François-Xavier Standaert, and Nicolas Veyrat-Charvillon. How to Certify the Leakage of a Chip? In Phong Q. Nguyen and Elisabeth Oswald, editors, *Advances in Cryptology - EUROCRYPT 2014 - 33rd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Copenhagen, Denmark, May 11-15, 2014. Proceedings*, volume 8441 of *Lecture Notes in Computer Science*, pages 459–476. Springer, 2014.
9. Benedikt Gierlichs, Lejla Batina, Pim Tuyls, and Bart Preneel. Mutual information analysis. In *CHES, 10th International Workshop*, volume 5154 of *Lecture Notes in Computer Science*, pages 426–442. Springer, August 10-13 2008. Washington, D.C., USA.
10. Annelie Heuser, Michael Kasper, Werner Schindler, and Marc Stöttinger. A New Difference Method for Side-Channel Analysis with High-Dimensional Leakage Models. In Orr Dunkelman, editor, *CT-RSA*, volume 7178 of *Lecture Notes in Computer Science*, pages 365–382. Springer, 2012.
11. Annelie Heuser, Olivier Rioul, and Sylvain Guilley. Good Is Not Good Enough - Deriving Optimal Distinguishers from Communication Theory. In Batina and Robshaw [1], pages 55–74.
12. Seiichi Matsuda and Shiho Moriai. Lightweight cryptography for the cloud: Exploit the power of bitslice implementation. In Emmanuel Prouff and Patrick Schaumont, editors, *Cryptographic Hardware and Embedded Systems - CHES 2012 - 14th International Workshop, Leuven, Belgium, September 9-12, 2012. Proceedings*, volume 7428 of *LNCS*, pages 408–425. Springer, 2012.
13. Amir Moradi, Nima Mousavi, Christof Paar, and Mahmoud Salmasizadeh. A Comparative Study of Mutual Information Analysis under a Gaussian Assumption. In *WISA (Information Security Applications, 10th International Workshop)*, volume 5932 of *Lecture Notes in Computer Science*, pages 193–205. Springer, August 25-27 2009. Busan, Korea.

14. Elke De Mulder, Benedikt Gierlichs, Bart Preneel, and Ingrid Verbauwhede. Practical DPA attacks on MDPL. In *First IEEE International Workshop on Information Forensics and Security, WIFS 2009, London, UK, December 6-9, 2009*, pages 191–195. IEEE, 2009.
15. Emmanuel Prouff, editor. *Constructive Side-Channel Analysis and Secure Design - 5th International Workshop, COSADE 2014, Paris, France, April 13-15, 2014. Revised Selected Papers*, volume 8622 of *Lecture Notes in Computer Science*. Springer, 2014.
16. Emmanuel Prouff and Matthieu Rivain. Theoretical and practical aspects of mutual information-based side channel analysis. *International Journal of Applied Cryptography (IJACT)*, 2(2):121–138, 2010.
17. Chester Rebeiro, A. David Selvakumar, and A. S. L. Devi. Bitslice implementation of AES. In David Pointcheval, Yi Mu, and Kefei Chen, editors, *Cryptology and Network Security, 5th International Conference, CANS 2006, Suzhou, China, December 8-10, 2006, Proceedings*, volume 4301 of *Lecture Notes in Computer Science*, pages 203–212. Springer, 2006.
18. Oscar Reparaz, Benedikt Gierlichs, and Ingrid Verbauwhede. Generic DPA Attacks: Curse or Blessing? In Prouff [15], pages 98–111.
19. François-Xavier Standaert, Tal Malkin, and Moti Yung. A Unified Framework for the Analysis of Side-Channel Key Recovery Attacks. In *EUROCRYPT*, volume 5479 of *LNCS*, pages 443–461. Springer, April 26-30 2009. Cologne, Germany.
20. Nicolas Veyrat-Charvillon and François-Xavier Standaert. Mutual Information Analysis: How, When and Why? In Christophe Clavier and Kris Gaj, editors, *Cryptographic Hardware and Embedded Systems - CHES 2009, 11th International Workshop, Lausanne, Switzerland, September 6-9, 2009, Proceedings*, volume 5747 of *Lecture Notes in Computer Science*, pages 429–443. Springer, 2009.
21. Carolyn Whitnall and Elisabeth Oswald. A Comprehensive Evaluation of Mutual Information Analysis Using a Fair Evaluation Framework. In Phillip Rogaway, editor, *CRYPTO*, volume 6841 of *Lecture Notes in Computer Science*, pages 316–334. Springer, 2011.
22. Carolyn Whitnall and Elisabeth Oswald. A Fair Evaluation Framework for Comparing Side-Channel Distinguishers. *J. Cryptographic Engineering*, 1(2):145–160, 2011.
23. Carolyn Whitnall, Elisabeth Oswald, and Luke Mather. An Exploration of the Kolmogorov-Smirnov Test as a Competitor to Mutual Information Analysis. In Emmanuel Prouff, editor, *CARDIS*, volume 7079 of *Lecture Notes in Computer Science*, pages 234–251. Springer, 2011.
24. Carolyn Whitnall, Elisabeth Oswald, and François-Xavier Standaert. The Myth of Generic DPA ... and the Magic of Learning. In Josh Benaloh, editor, *CT-RSA*, volume 8366 of *Lecture Notes in Computer Science*, pages 183–205. Springer, 2014.