Theoretical and Practical Aspects of Mutual Information Based Side Channel Analysis

Emmanuel Prouff[2] and Matthieu Rivain[1,2]

[1] University of Luxembourg
[2] Oberthur Technologies
{e.prouff,m.rivain}@oberthur.com

Abstract. A large variety of side channel analyses performed on embedded devices involve the linear correlation coefficient as wrong-key distinguisher. This coefficient is actually a sound statistical tool to quantify linear dependencies between univariate variables. However, when those dependencies are non-linear, the correlation coefficient stops being pertinent so that another statistical tool must be investigated. Recent works showed that the *Mutual Information* measure is a promising candidate, since it detects any kind of statistical dependency. Substituting it for the correlation coefficient may therefore be considered as a natural extension of the existing attacks. Nevertheless, the first applications published at CHES 2008 have revealed several limitations of the approach and have raised several questions. In this paper, an in-depth analysis of side channel attacks involving the mutual information is conducted. We expose their theoretical foundations and we assess their limitations and assets. Also, we generalize them to higher orders where they seem to be an efficient alternative to the existing attacks. Eventually, we provide simulations and practical experiments that validate our theoretical analyses.

1 Introduction

Side Channel Analysis (SCA) is a cryptanalytic technique that consists in analyzing the physical leakage produced during the execution of a cryptographic algorithm embedded on a physical device. This side channel leakage is indeed statistically dependent on the intermediate variables of the computation which enables key recovery attacks.

Since their introduction in the nineties, several kinds of SCA have been proposed which essentially differ in the involved distinguisher. A first family is composed of SCA based on linear correlation distinguishers. When such an attack is performed, the adversary implicitly assumes that there is a linear dependence between its predictions and the leakage measurements. Actually, the attack effectiveness depends on the accuracy of this assumption. The most well-known examples of such attacks are the *Differential Power Analysis* (DPA) [1] that is based on a Boolean correlation and the *Correlation Power Analysis* (CPA) [2] that involves *Pearson correlation* coefficient. The second important family of SCA is composed of the so-called *Template Attacks* (TA) [3]. They involve

maximum-likelihood distinguishers and can succeed when the DPA or CPA do not. However, TA can only be performed if the attacker owns a profile of the leakage according to the values of some intermediate variables, which is a strong limitation.

Recently a new kind of SCA, called *Mutual Information Analysis* (MIA), has been proposed in [4]. It uses the *Mutual Information* as distinguisher. It is an interesting alternative to the aforementioned attacks since some assumptions about the adversary can be relaxed. In particular it does not require a linear dependency between the leakage and the predicted data (as for CPA) and is actually able to exploit any kind of dependency. Moreover, this gain in generality is obtained without needing to profile the leakage as it is the case for TA.

Despite its advantages, the MIA suffers from several limitations and the preliminary work of Gierlichs *et al.* [4] poses a number of open questions. First of all, the MIA efficiency has not been clearly established and it is not clear whether (and in which contexts) it is better than the other attacks that assume the same adversary capabilities (as *e.g.* the CPA). The first attack experiments presented in [4] suggest that MIA's efficiency is strongly related to the attack context (device, algorithmic target, noise, etc.). However, at this time an in-depth analysis is missing to have a clear idea about this relationship. Secondly, the estimation of the mutual information, which itself requires the estimation of statistical distributions, is a major practical issue that has not been fully investigated in [4]. This problematic has been dealt with in Statistics and Applied Probabilities Theory (see for instance [5] for an overview). Among the existing estimation methods, it is of crucial interest to determine the one that optimizes the MIA. Only such a study will indeed allow us to form an unbiased opinion about its efficiency *versus* the one of attacks involving linear dependence based distinguishers.

2 Preliminaries on Probability and Information Theory

We use the calligraphic letters, like $\mathcal{X}$, to denote sets. The corresponding large letter X is then used to denote a random variable (r.v. for short) over $\mathcal{X}$, while the lowercase letter x - a particular element from $\mathcal{X}$. For every positive integer n, we denote by $\mathbf{X}$ a n-dimensional r.v. $(X_1, \cdots, X_n) \in \mathcal{X}^n$, while the lowercase letter $\mathbf{x}$ - a particular element from $\mathcal{X}^n$. To every discrete r.v. $\mathbf{X}$, one associates a probability mass function $\mathrm{p}_{\mathbf{X}}$ defined by $\mathrm{p}_{\mathbf{X}}(\mathbf{x}) = \mathrm{p}[\mathbf{X} = \mathbf{x}]$. If X is continuous, one associates to $\mathbf{X}$ its *probability density function* (pdf for short), denoted by $g_{\mathbf{X}}$: for every $\mathbf{x} \in \mathcal{X}^n$, we have $\mathrm{p}_{\mathbf{X}}[X_1 \leq x_1, \cdots, X_n \leq x_n] = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} g_{\mathbf{X}}(t_1, \cdots, t_n) dt_1 \cdots dt_n$.

The *Gaussian distribution* is an important family of probability distributions, applicable in many fields. A r.v. $\mathbf{X}$ having such a distribution is said to be *Gaussian* and its pdf $g_{\mu,\Sigma}$ is defined for every $\mathbf{x} \in \mathcal{X}^n$ by:

$$g_{\mu,\Sigma}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right) \quad , \tag{1}$$

where μ and Σ respectively denote the *mean* and the *covariance matrix* of $\mathbf{X}$.

In this paper, we will study r.v. whose pdf is a finite linear combination of Gaussian pdfs. Such a pdf, which is called a *Gaussian mixture* (GM for short), is denoted by g_θ and it satisfies for every $\mathbf{x} \in \mathcal{X}^n$:

$$g_\theta(\mathbf{x}) = \sum_{t=1}^{T} a_t g_{\mu_t, \Sigma_t}(\mathbf{x}) \quad , \tag{2}$$

where $\theta = ((a_t, \mu_t, \Sigma_t))_{1 \leq t \leq T}$ is a $3T$-dimensional vector containing the so-called *mixing probabilities* a_t's (that satisfy $\sum_t a_t = 1$), as well as the means μ_t and the covariance matrices Σ_t of the T Gaussian pdfs in the mixture.

The *entropy* $\mathrm{H}(\mathbf{X})$ of a discrete n-dimensional r.v. $\mathbf{X}$ aims at measuring the amount of information provided by an observation of $\mathbf{X}$. It is defined by $\mathrm{H}(\mathbf{X}) = -\sum_{\mathbf{x} \in \mathcal{X}^n} \mathrm{p}_{\mathbf{X}}(\mathbf{x}) \log_2(\mathrm{p}_{\mathbf{X}}(\mathbf{x}))$. The *differential entropy* extends the notion of entropy to continuous n-dimensional r.v. Contrary to the entropy, the differential entropy can be negative. It is defined by:

$$\mathrm{H}(\mathbf{X}) = -\int_{\mathbf{x} \in \mathcal{X}^n} g_{\mathbf{X}}(\mathbf{x}) \log_2(g_{\mathbf{X}}(\mathbf{x})) d\mathbf{x} \ . \tag{3}$$

If $\mathbf{X}$ is a n-dimensional Gaussian r.v. with pdf $g_{\mu,\Sigma}$, then its entropy satisfies:

$$\mathrm{H}(\mathbf{X}) = \frac{1}{2} \log((2\pi e)^n |\Sigma|) \ . \tag{4}$$

In the general case when $\mathbf{X}$ has a GM pdf mixing more than one Gaussian pdf, there is no analytical expression for its differential entropy. However, upper and lower bounds can be derived. We recall hereafter the lower bound.

Proposition 1. *[6] Let $\mathbf{X} \in \mathcal{X}^n$ be a Gaussian mixture whose pdf g_θ is such that $\theta = ((a_i, \mu_i, \Sigma_i))_{i=1,\cdots,T}$. Then, its differential entropy satisfies:*

$$\frac{1}{2} \log \left((2\pi e)^n \prod_{t=1}^{T} |\Sigma_t|^{a_t} \right) \leq H(\mathbf{X}) \ . \tag{5}$$

To quantify the amount of information that a second r.v. $\mathbf{Y}$ reveals about $\mathbf{X}$, the notion of *mutual information* is usually involved. It is the value $\mathrm{I}(\mathbf{X}, \mathbf{Y})$ defined by $\mathrm{I}(\mathbf{X}, \mathbf{Y}) = \mathrm{H}(\mathbf{X}) - \mathrm{H}(\mathbf{X}|\mathbf{Y})$, where $\mathrm{H}(\mathbf{X}|\mathbf{Y})$ is called the *conditional entropy of* $\mathbf{X}$ *knowing* $\mathbf{Y}$. If $\mathbf{Y}$ is discrete, then it is defined by:

$$\mathrm{H}(\mathbf{X}|\mathbf{Y}) = \sum_{y \in \mathcal{Y}} \mathrm{p}_{\mathbf{Y}}(y) \mathrm{H}(\mathbf{X}|\mathbf{Y} = y) \ , \tag{6}$$

Thanks to the mutual information (or to the conditional entropy), we have a way to decide about the dependency of two multi-variate random variables: $\mathbf{X}$ and $\mathbf{Y}$ are *independent* iff $I(\mathbf{X}, \mathbf{Y})$ equals 0 or equivalently iff $\mathrm{H}(\mathbf{X}|\mathbf{Y}) = \mathrm{H}(\mathbf{X})$.

3 Brief Overview of Side Channel Attacks

Any intermediate variable which is a function $f(X, k^\star)$ of a plaintext X and a guessable secret key $k^\star$ is *sensitive* and its manipulation can be targeted by an SCA. For every key-candidate $k \in \mathcal{K}$, we denote by f_k the function $x \mapsto f(x, k)$ and by $L(k^\star)$ the *leakage variable* that models the leakage produced by the manipulation/computation of $f_{k^\star}(X)$ by the device. The leakage variable can be expressed as:

$$L(k^\star) = \varphi \circ f_{k^\star}(X) + B \ , \tag{7}$$

where φ denotes a deterministic function and B denotes an independent noise.

In (7), the definition of f only depends on the algorithm that is implemented and it is known to the attacker (it can for instance be a S-box function). On the opposite, φ only depends on the device and its exact definition is usually unknown to the attacker who will estimate it according to the device specifications and/or to a leakage profiling phase. Actually, the SCAs essentially differ in the degree of knowledge on φ and B that is required for the attack to succeed.

In a DPA, the attacker only needs to know that the mean of the r.v. $\varphi \circ f_{k^\star}(X)$ depends on a given bit of $f_{k^\star}(X)$. Based on this assumption, each key candidate k is involved to split the measurements into two sets and the candidates are discriminated by computing differences of means between those sets. This essentially amounts to process a Boolean correlation.

In a CPA, the attacker must know a function $\hat{\varphi}$ that is a good linear approximation of φ (*i.e.* such that $\hat{\varphi}$ and φ are linearly correlated). Usually, he chooses the Hamming weight function for $\hat{\varphi}$. Based on this assumption, key candidates k are discriminated by testing the linear correlation between $\hat{\varphi} \circ f_k(x_i)$ and $L(k^\star)$ for a sample of plaintexts $(x_i)_i$. This attack can be more efficient than the single-bit DPA. However, its success highly depends on the correctness of the linear approximation of φ by $\hat{\varphi}$.

In a TA, the attacker must know a good approximation of the pdf of the leakage $L(k)$ for every possible key k. It amounts for the attacker to have a good approximation of φ and of the standard deviation of the noise B (or its covariance matrix in a multivariate model). Wrong key hypotheses are discriminated in a maximum likelihood attack (see [3]). To pre-compute the pdf's of all the variables $L(k)$, the attacker needs to have an open access to a copy of the device under attack. This is a strong constraint since it is often very difficult to have such an open copy in practice.

As noticed in [4, 5], MIA attacks are an alternative to the approaches above. They consist in estimating the mutual information $\mathrm{I}(L(k^\star), \hat{\varphi} \circ f_k(X))$ instead of the correlation coefficient or the difference of means. In an MIA, the attacker is potentially allowed to make weaker assumptions on φ than in the CPA. Indeed, he does not need a good linear approximation of φ but only a function $\hat{\varphi}$ s.t. the mutual information $\mathrm{I}(\hat{\varphi}, \varphi)$ is non-negligible (which may happen even if φ and $\hat{\varphi}$ are not linearly correlated). It for instance allows the attacker to choose the identity function for $\hat{\varphi}$ which is of particular interest since no knowledge about the leakage parameters is required.

The effectiveness of a key-recovery side channel attack is usually characterized by its *success rate*, namely the probability that the attack outputs the correct key as a the most likely key candidate. This notion can be extended to higher orders [7]: an attack is said to be *o-th order successful* if it classifies the correct key among the o most likely key candidates. In the following, we shall investigate the (o-th order) success rate of MIA.

Let us denote by $Z(k)$ the r.v. $\hat{\varphi} \circ f_k(X)$. Moreover, for every function F defined over $\mathcal{K}$, let us denote by argmax-o $_{k\in\mathcal{K}}$ $F(k)$ the set composed of the o key candidates k such that $F(k)$ is among the o highest values in $\{F(k); k \in \mathcal{K}\}$. An MIA succeeds at the o-th order iff the estimations $\hat{\mathrm{I}}(L(k^\star), Z(k))$ of $\mathrm{I}(L(k^\star), Z(k))$ satisfy:

$$k^\star \in \underset{k\in\mathcal{K}}{\text{argmax-}o} \ \hat{\mathrm{I}}(L(k^\star), Z(k)) \ . \tag{8}$$

We therefore deduce two necessary conditions for an MIA to succeed at the o-th order:

- *Theoretical.* The mutual information $(\mathrm{I}(L(k^\star), Z(k)))_{k\in\mathcal{K}}$ must satisfy:

$$k^\star \in \underset{k\in\mathcal{K}}{\text{argmax-}o} \ \mathrm{I}(L(k^\star), Z(k)) \ . \tag{9}$$

- *Practical.* The estimations of $(\mathrm{I}(L(k^\star), Z(k)))_{k\in\mathcal{K}}$ must be good enough to satisfy (8) while (9) is satisfied.

In the next section, we study when Relation (9) is satisfied. This will allow us to characterize (with regards to f, φ, $\hat{\varphi}$) when an MIA is theoretically possible. Then, for 3-tuples $(f, \varphi, \hat{\varphi})$ s.t. (9) is satisfied, we shall study in Sec. 6 the success probability of the MIA according to the estimation method used to compute $\hat{\mathrm{I}}$ and according to the noise variation. This will allow us to characterize when an MIA is practically feasible (*i.e.* when (8) is satisfied) and when it is more efficient than the other SCA attacks.

4 Study of the MIA in the Gaussian Model

In this section we focus on first order MIA and, in a second time, we extend our analysis to the higher order case *i.e.* when the target implementation is protected by masking [8]. Our analyses are done under the three following assumptions which are realistic in a side channel analysis context and make the formalization easier.

Assumption 1 (Uniformity) *The plaintext X has a uniform distribution over $\mathbb{F}_2^n$.*

Assumption 2 (Balancedness) *For every $k \in \mathcal{K}$, the (n, m)-function $f_k : x \mapsto f_k(x)$ is s.t. $\#\{x \in \mathbb{F}_2^n; y = f_k(x)\}$ equals 2^{n-m} for every $y \in \mathbb{F}_2^m$.*

Remark 1. This assumption states that the algorithmic functions targeted by the SCA are balanced which is usually the case in a cryptographic context.

Assumption 3 (Gaussian Noise) *The noise B in the leakage (see (7)) has a Gaussian distribution with zero mean and standard deviation σ.*

Remark 2. This assumption is realistic and is therefore often done in the literature (see for instance [8, 9, 7]). Practical attacks and pdf estimations presented in Sec. 6 provide us with an experimental validation of this assumption.

For clarity reasons, in the next sections we shall denote by L (resp. by Z) the random variable $L(k^\star)$ (resp. $Z(k)$) when there is no ambiguity.

4.1 First Order MIA

The mutual information $\mathrm{I}(L, Z(k))$ equals $\mathrm{H}(L) - \mathrm{H}(L, Z(k))$. Since $\mathrm{H}(L)$ does not depend on the key prediction, $\mathrm{I}(L|Z(k))$ reaches one of its o highest values when k ranges over $\mathcal{K}$ iff the conditional entropy $\mathrm{H}(L|Z(k))$ reaches one of its o smallest values. One deduces that an MIA is theoretically possible iff the 3-tuple $(f, \varphi, \hat{\varphi})$ is s.t.:

$$k^\star \in \underset{k \in \mathcal{K}}{\text{argmin-}o} \ \ \mathrm{H}(L(k^\star)|Z(k)) \ , \tag{10}$$

where argmin-o is defined analogously to argmax-o .

The starting point of our analysis is that studying the MIA effectiveness is equivalent to investigating the minimality of $\mathrm{H}(L|Z(k))$ over $\mathcal{K}$. As a consequence of (6), we have $\mathrm{H}(L|Z(k)) = \sum_{z \in \mathrm{Im}(\hat{\varphi})} \mathrm{p}_Z(z) \mathrm{H}(L|Z(k) = z)$. From (3), one deduces:

$$\mathrm{H}(L|Z(k)) = - \sum_{z \in \mathrm{Im}(\hat{\varphi})} \mathrm{p}_Z(z) \int_\ell g_{L|Z=z}(\ell) \log g_{L|Z=z}(\ell) d\ell \ . \tag{11}$$

To reveal the relationship between $\mathrm{H}(L|Z(k))$ and the key-prediction k, the expression of the pdf $g_{L|Z=z}$ in (11) needs to be developed. Let us denote by $E_k(z)$ the set $[\hat{\varphi} \circ f_k]^{-1}(z)$. Since X has a uniform distribution over $\mathbb{F}_2^n$, for every $\ell \in \mathcal{L}$ and every $z \in \mathrm{Im}(\hat{\varphi} \circ f_k)$ we have:

$$g_{L|Z=z}(\ell) = \frac{1}{\#E_k(z)} \sum_{x \in E_k(z)} g_{\varphi \circ f_{k^\star}(x), \sigma}(\ell) \ . \tag{12}$$

The next proposition directly follows.

Proposition 2. *If X is a r.v. with uniform distribution, then for every pair $(k^\star, k) \in \mathcal{K}^2$ and every $z \in \mathcal{Z}$ the pdf of the r.v. $(L(k^\star) \mid Z(k) = z)$ is a GM g_θ whose parameter θ satisfies $\theta = \left((a_{z,t}, t, \sigma^2)\right)_{t \in Im(\varphi)}$, with $a_{z,t} = \mathrm{p}\left[\varphi \circ f_{k^\star}(X) = t \mid \hat{\varphi} \circ f_k(X) = z\right]$.*

In Proposition 2, the key hypothesis k only plays a part in the definition of the weights $a_{z,t}$ of the GM. In other terms, $g_{L|Z(k)=z}$ is always composed of the same Gaussian pdfs and the key hypothesis k only impacts the way how the Gaussian pdfs are mixed. To go further in the study of the relationship between k and $\mathrm{H}(L(k^\star)|Z(k)=z)$, let us introduce the following diagram where z is an element of $\mathrm{Im}(\hat{\varphi})$, where F', F and T are image sets:

$$z \xrightarrow{\hat{\varphi}^{-1}} F' \xrightarrow{f_k^{-1}} E_k(z) \xrightarrow{f_{k^\star}} F \xrightarrow{\varphi} T \ ,$$

Based on the diagram above, we can make the two following observations:

- If the set T is reduced to a singleton set $\{t_1\}$ (*i.e.* if $\hat{\varphi} \circ f_k$ is constant equal to t_1 on $E_k(z)$), then all the probabilities $a_{z,t}$ s.t. $t \neq t_1$ are null and a_{z,t_1} equals 1. In this case, one deduces from Proposition 2 that the distribution of $(L(k^\star)|Z(k)=z)$ is Gaussian and, due to (4), its conditional entropy satisfies
$$\mathrm{H}(L(k^\star)|Z(k)=z) = \frac{1}{2}\log(2\pi e\sigma^2) \ .$$
- If $\#T > 1$ (*i.e.* if $\#\varphi \circ f_{k^\star}(E_k(z)) > 1$), then there exist at least two probabilities a_{z,t_1} and a_{z,t_2} which are non-null and the distribution of $(L(k^\star)|Z(k)=z)$ is a GM (not Gaussian). Due to (5), its entropy satisfies:
$$\mathrm{H}(L(k^\star)|Z(k)=z) \geq \frac{1}{2}\log(2\pi e\sigma^2) \ .$$

When φ is constant on F' (*e.g.* when $\hat{\varphi}=\varphi$ or $\hat{\varphi}=\mathrm{Id}$), the two observations above provide us with a discriminant property. If $k^\star = k$, then we have $F = F'$ and thus T is a singleton and $\mathrm{H}(L|Z=z)$ equals $\frac{1}{2}\log(2\pi e\sigma^2)$. Otherwise, if $k \neq k^\star$, then $f_{k^\star} \circ f_k$ is likely to behave as a random function[3]. In this case, F is most of the time different from F' and T is therefore likely to have more than one element[4]. This implies that $\#\varphi \circ f_{k^\star}(E_k(z))$ is strictly greater than 1 and thus that $\mathrm{H}(L|Z=z)$ is greater than or equal to $\frac{1}{2}\log(2\pi e\sigma^2)$. Eventually, we get the following proposition in which we exhibit a tight lower bound for the differential entropy $\mathrm{H}(L(k^\star)|Z(k))$.

Proposition 3. *For every $(k^\star, k) \in \mathcal{K}^2$, the conditional entropy of the r.v. $(L(k^\star)|Z(k))$ satisfies:*

$$\frac{1}{2}\log\left(2\pi e\sigma^2\right) \leq H((L(k^\star)|Z(k)) \ . \tag{13}$$

If $\varphi \circ f_{k^\star}$ is constant on $E_k(z)$ for every $z \in \mathcal{Z}$, then the lower bound is tight.

Proof. Relation (13) is a straightforward consequence of (6) and of Propositions 1 and 2. The tightness is a direct consequence of (4) and Proposition 2. ◇

[3] This property, sometimes called *wrong key assumption* [10], is often assumed to be true in a cryptographic context, due to the specific properties of the primitive f.

[4] As detailed later, this is only true if $\hat{\varphi} \circ f_k$ is non-injective.

Remark 3. Intuitively, the entropy $\mathrm{H}(Y)$ is a measure of the diversity or randomness of Y. It is therefore reasonable to think that the more components in the GM pdf of $(L(k^\star)|Z(k))$, the greater its entropy. Relation (13) provides a first validation of this intuition. The entropy is minimal when the pdf is a Gaussian one (*i.e.* when the GM has only one component). In our experiments (partially reported in Sec. 6), we noticed that the entropy of a GM whose components have the same variance, increases with the number of components.

Corollary 1. *If $\hat{\varphi} \circ f_k$ is injective, then $H(L(k^\star)|Z(k))$ equals $\frac{1}{2}\log(2\pi e\sigma^2)$.*

Proof. If $\hat{\varphi} \circ f_k$ is injective, then $E_k(z)$ is a singleton and $\varphi \circ f_{k^\star}$ is thus constant on $E_k(z)$.

If the functions $\hat{\varphi} \circ f_k$'s are all injective, then Corollary 1 implies that the MIA cannot succeed at any order. Indeed, in this case the entropy $\mathrm{H}(L(k^\star)|Z(k))$ stays unchanged when k ranges over $\mathcal{K}$ and thus, $k^\star$ does not satisfy (10). As a consequence, when the f_k's are injective (which is for instance the case when f_k consists in a key addition followed by the AES S-box), then the attacker has to choose $\hat{\varphi}$ to be non-injective (*e.g.* the Hamming weight function). It must be noticed that this is a necessary but not sufficient condition since the function $\hat{\varphi}$ must also be s.t. $\mathrm{I}(\hat{\varphi}, \varphi)$ is non-negligible (otherwise the MIA would clearly failed). In this case, the attacker must have a certain knowledge about the leakage function φ in order to define an appropriate function $\hat{\varphi}$ and hence, the MIA does no longer benefit from one of its main advantages. This drawback can be overcome by exclusively targeting intermediate variables s.t. the f_k's are not injective (in AES, the attacker can for instance target the bitwise addition between two S-box outputs during the *MixColumns* operation).

4.2 Generalization to the Higher Order Case

In this section, we extend the analysis of MIA to higher orders and we assume that the implementation protected by masking. The sensitive variable $f_{k^\star}(X)$ is now masked with $d-1$ independent random variables M_1, ..., M_{d-1} which are uniformly distributed over $\mathrm{Im}(f)$.

The masked data $f_{k^\star}(X) \oplus M_1 \oplus \cdots \oplus M_{d-1}$ and the different masks M_j's are processed at different times. The leakage about $f_{k^\star}(X) \oplus M_1 \oplus \cdots \oplus M_{d-1}$ is denoted by L_0 and the leakages about the M_j's are denoted by L_1, ..., L_{d-1}. Under Assumption 3, the L_j's satisfy:

$$L_j = \begin{cases} \varphi[f_{k^\star}(X) \oplus \bigoplus_{t=1}^{d-1} M_t] + B_0 & \text{if } j = 0, \\ \varphi_j(M_j) + B_j & \text{if } j \neq 0, \end{cases} \tag{14}$$

where the B_j's are independent Gaussian noises with mean 0 and standard deviations σ_j, and where φ, φ_1, $\cdots$, φ_{d-1} are d device dependent functions that are *a priori* unknown to the attacker. The vector $(L_0, \cdots, L_{d-1})$ is denoted by $\mathbf{L}$. The vector of masks $(M_1, \cdots, M_{d-1})$ is denoted by $\mathbf{M}$. We denote by $\Phi_{k^\star}(X, \mathbf{M})$ the vector $(\varphi(f_{k^\star}(X) \oplus \bigoplus_{t=1}^{d-1} M_t), \varphi_1(M_1), \cdots, \varphi_{d-1}(M_{d-1}))$.

To simplify our analysis, we assume that the attacker knows the manipulation times exactly and is therefore able to get a sample for the r.v. $\mathbf{L}$. Under this assumption and for the same reasons as in the univariate case, the higher order MIA essentially consists in looking for the key candidate k which minimizes an estimation of the conditional entropy $\mathrm{H}(\mathbf{L}|Z(k))$. Due to (6), this entropy equals $\sum_{z\in \mathrm{Im}(\hat{\varphi})} \mathrm{p}_{Z(k)}(z)\mathrm{H}(\mathbf{L}|Z(k)=z)$. Since Z equals $\hat{\varphi}\circ f_k(X)$, the probabilities $\mathrm{p}_{Z(k)}(z)$ in this sum can be exactly computed by the attacker. Once this computation has been performed, estimating $\mathrm{H}(\mathbf{L}|Z(k))$ amounts to estimate the entropies $\mathrm{H}(\mathbf{L}|Z(k)=z)$ for all the hypotheses k. These entropies are estimated as for the first order case (see (11)), but the pdfs $g_{\mathbf{L}|Z(k)=z}$ are multivariate. More precisely, after denoting by Σ the matrix $\left(\mathrm{Cov}\left[B_i,B_j\right]\right)_{i,j}$, we get:

$$g_{\mathbf{L}|Z(k)=z}(\ell) = \frac{1}{\#E_k(z)(\#Im(f))^{d-1}} \sum_{\substack{x\in E_k(z)\\ \mathbf{m}\in Im(f)^{d-1}}} g_{\Phi_{k^\star}(x,\mathbf{m}),\Sigma}(\ell)\ . \tag{15}$$

In a similar way than in Sec. 4, the next proposition directly follows.

Proposition 4. *If X is a r.v. with uniform distribution, then for every pair $(k^\star,k)\in\mathcal{K}^2$ and every $z\in\mathcal{Z}$ the pdf of the r.v. $(\mathbf{L}(k^\star)\mid Z(k)=z)$ is a GM g_θ whose parameter θ satisfies $\theta=\left((a_{z,\mathbf{t}},\mathbf{t},\Sigma)\right)_{\mathbf{t}}$, with $\Sigma=\left(Cov\left[B_i,B_j\right]\right)_{i,j}$ and $a_{z,\mathbf{t}}=\mathrm{p}[\Phi_{k^\star}(X,\mathbf{M})=\mathbf{t}\mid\hat{\varphi}\circ f_k(X)=z]$.*

We deduce from Propositions 1 and 4 the following result.

Proposition 5. *If X is a r.v. with uniform distribution over $\mathcal{X}$, then for every $(k^\star,k)\in\mathcal{K}^2$, the entropy of the r.v. $(\mathbf{L}(k^\star)|(Z(k),\mathbf{M}))$ satisfies:*

$$\frac{1}{2}\log\left((2\pi e)^d|\Sigma|\right) \leq H(\mathbf{L}(k^\star)|(Z(k),\mathbf{M}))\ . \tag{16}$$

If $\varphi\circ f_{k^\star}$ is constant on $E_k(z)$ for every $z\in Im(Z)$, then the bound is tight.

We cannot deduce from the proposition above a wrong-key discriminator as we did in the univariate case. Indeed, to compute the entropy in (16) the attacker must know the mask values, which is impossible in our context. However, if the 3-tuple $(f,\varphi,\hat{\varphi})$ satisfies the condition of Proposition 5, then it can be checked that for every z the number of components in the multi-variate GM pdf of $(L(k^\star)|Z(k)=z)$ reaches its minimum for $k=k^\star$. As discussed in Remark 3, this implies that the entropy of $\mathbf{L}(k^\star)|Z(k)$ is likely to be minimum for $k=k^\star$. The simulations and experiments presented in Sec. 6 provides us with an experimental validation of this fact.

In the next sections, we assume that an MIA is theoretically possible. Namely, we assume that $k^\star$ belongs to argmin-o $_k$ $\mathrm{H}(\mathbf{L}(k^\star)|Z(k))$ for a given order o. At first, we study the success probability of an MIA according to the method used to estimate $\mathrm{H}(\mathbf{L}(k^\star)|Z(k))$ and the noise variation. Secondly, we compare the efficiency of an MIA with the one of the CPA in different contexts.

5 Conditional Entropy Estimation

Let $\mathbf{L}$ be a d-dimensional r.v. defined over $\mathcal{L}^d$ (*i.e.* $\mathbf{L}$ is composed of d different instantaneous leakage measurements) and let k be a key-candidate. We assume that the attacker has a sample of N leakage-message pairs $(\mathbf{l}_i, x_i) \in \mathcal{L}^d \times \mathcal{X}$ corresponding to a key $k^\star$, and that he wants to compute $\mathrm{H}(\mathbf{L}|Z(k))$ to discriminate key-candidates k. Due to (6), estimating $\mathrm{H}(\mathbf{L}|Z(k))$ from the sample $((\mathbf{l}_i, x_i))_i$ essentially amounts to estimate the entropy $\mathrm{H}(\mathbf{L}|Z(k) = z)$ for every $z \in \mathcal{Z}$. For such a purpose, a first step is to compute estimations $\hat{g}_{\mathbf{L}|Z=z}$ of the $g_{\mathbf{L}|Z=z}$'s. Then, depending on the estimation method that has been applied, the $\mathrm{H}(\mathbf{L}|Z(k) = z)$'s are either directly computable (Histogram method) or must still be estimated (Kernel and Parametric methods). In the following we present three estimation methods and we discuss their pertinency in our context.

5.1 Histogram method

Description. We choose d *bin widths* h_0, ..., h_{d-1} (one for each coordinate of the leakage vectors) and we partition the leakage space $\mathcal{L}^d$ into regions $(\mathcal{R}_\alpha)_\alpha$ with equal volume $\upsilon = \prod_j h_j$. Let k be a key-candidate and let z be an element of $\mathcal{Z}$. We denote by $\mathcal{S}_z$ the sub-sample $\left(\mathbf{l}_i; x_i \in [\varphi \circ f_k]^{-1}(z)\right)_i \subseteq (\mathbf{l}_i)_i$ and by $\ell_{i,j}$ the jth coordinate of $\mathbf{l}_i$. To estimate the pdf $g_{\mathbf{L}|Z=z}$, we first compute the density vector D_z whose coordinates are defined by:

$$D_z(\alpha) = \frac{\#(\mathcal{S}_z \cap \mathcal{R}_\alpha)}{\#\mathcal{S}_z} \ , \tag{17}$$

where $\mathcal{S}_z \cap \mathcal{R}_\alpha$ denotes the sample of all the $\mathbf{l}_i$'s in $\mathcal{S}_z$ that belong to $\mathcal{R}_\alpha$.

The estimation $\hat{g}_{L|Z=z}$ is then defined for every $\mathbf{l} \in \mathcal{L}^d$ by $\hat{g}_{\mathbf{L}|Z=z}(\mathbf{l}) = \frac{D_z(i_\mathbf{l})}{\upsilon}$, where $i_\mathbf{l}$ is the index of the region $\mathcal{R}_{i_\mathbf{l}}$ that contains $\mathbf{l}$. Integrating the pdf estimation according to formula (3) gives the following estimation for the conditional entropy: $\hat{\mathrm{H}}(\mathbf{L}|Z = z) = -\sum_\alpha D_z(\alpha) \log(D_z(\alpha)/\upsilon)$. We eventually get:

$$\hat{\mathrm{H}}(\mathbf{L}|Z) = -\sum_{z \in \mathcal{Z}} \mathrm{p}_Z(z) \sum_\alpha D_z(\alpha) \log\left((\frac{D_z(\alpha)}{\upsilon}\right) \ . \tag{18}$$

The optimal choice of the bin widths h_j is an issue in Statistical Theory. Actually, there are several rules that aim at providing *ad hoc* formulae for computing the h_j's based on the nature of the samples (see for instance [11, 12]). In our simulations, we chose to follow the Scott Rule. Namely, if $\hat{\sigma}_j$ denotes the estimated standard deviation of the sample $(\ell_{i,j})_i$ of size N_j, then h_j satisfies $h_j = 3.49 \times \hat{\sigma}_j \times N_j^{-\frac{1}{3}}$ (notice that in our context all the N_j's are equal to N).

Simulations. In order to illustrate the Histogram method in the context of an MIA attack, we generated 10000 leakage measurements in the Gaussian model (7) for φ being the Hamming weight function, for f being the first DES S-box parameterized with the key $k^\star = 11$ and for $\sigma = 0.1$. Since the DES S-box is

non-injective, we chose the identity function for $\hat{\varphi}$. Fig. 1 plots the estimations of the pdf $g_{L|Z=1}$ when $k = 11$ and when $k = 5$ (for a number of bins equal to 285). As expected (Proposition 2 and Corollary 1), a Gaussian pdf seems to

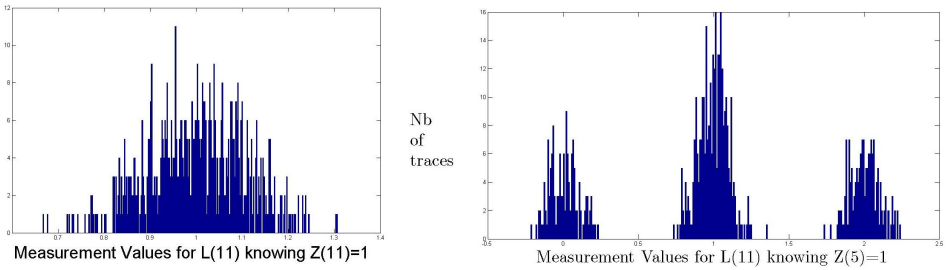


Fig. 1. Histogram Method in the First Order Case.

be estimated when $k = 11$ (good key prediction), whereas a mixture of three Gaussian distributions seems to be estimated when $k = 5$ (wrong key prediction). For the experimentation described in the left-hand figure we obtained $\hat{\mathrm{H}}(L(11)|Z(11) = 1) = -1.31$ (due to (4) we have $\mathrm{H}(L(11)|Z(11) = 1) = -1.27$) and we got $\hat{\mathrm{H}}(L(11)|Z(5) = 1) = -0.0345$ for the one in the right-hand side. Moreover, we validated that the estimated conditional entropy is minimum for the good key hypothesis.

In order to illustrate the Histogram method in the context of a 2nd order MIA attack, we generated 10000 pairs of leakage measurements in the higher order Gaussian model (14) with $d = 2$, with φ and φ_1 being the Hamming Weight function, with f being the first DES S-box parametric with the key $k^\star = 11$ and with $\sigma_0 = \sigma_1 = 0.1$. Fig. 2 plots the estimations of the pdf $g_{\mathbf{L}|Z=1}$ when $k = 11$ and when $k = 5$. As expected, the mixture of Gaussian distributions for $k = 11$

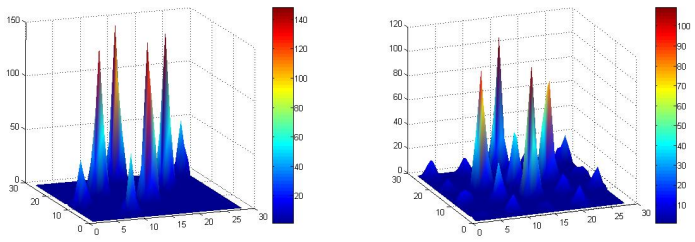


Fig. 2. Histogram Method in the Second Order Case.

have less components than for $k = 5$. For the experimentation in the left-hand figure we obtained $\hat{\mathrm{H}}(\mathbf{L}(11)|Z(11) = 1) = 0.22$ (and $\hat{\mathrm{H}}(\mathbf{L}(11)|Z(11)) = 0.14$), whereas we got 1.12 for $\hat{\mathrm{H}}(\mathbf{L}(11)|Z(5) = 1)$ (and 1.15 for $\hat{\mathrm{H}}(\mathbf{L}(11)|Z(5))$). Here again, the estimated conditional entropy was minimum for the good key hypothesis.

5.2 Kernel Density Method

Description. Although the Histogram method can be made to be asymptotically consistent, other methods can be used that converge at faster rates. For instance, rather than grouping observations together in bins, the so-called *Kernel density estimator* (or *Parzen window* method) can be thought to place small "bumps" at each observation, determined by the Kernel function (see for instance [13]). The estimator consists of a "sum of bumps" and is clearly smoother as a result than the Histogram method.

The Kernel density estimation $\hat{g}_{\mathbf{L}|Z=z}$ based on the sample $\mathcal{S}_z$ is defined for every $\mathbf{l} = (\ell_0, ..., \ell_{d-1}) \in \mathcal{L}^d$ by:

$$\hat{g}_{L|Z=z}(\mathbf{l}) = \frac{1}{\#\mathcal{S}_z} \sum_{\mathbf{l}_i = (\ell_{i,0}, ..., \ell_{i,d-1}) \in \mathcal{S}_z} \frac{1}{\upsilon} \times \prod_{j=0}^{d-1} \mathbf{K}\left(\frac{\ell_j - \ell_{i,j}}{h_j}\right) \ ,$$

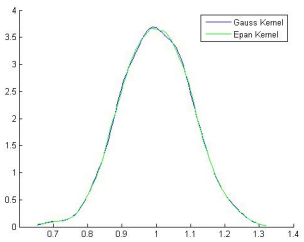
where $\mathbf{K}$ is a *Kernel function* chosen among the classical ones (see for instance [14]), where the h_i's are *Kernel bandwidths* and where υ equals $\prod_j h_j$. As recalled in [15], the following Parzen-windows entropy estimation of $\mathrm{H}(\mathbf{L}|Z = z)$ is sound when the sample size is large enough:

$$\hat{\mathrm{H}}(\mathbf{L}|Z = z) = -\frac{1}{\#\mathcal{S}_z} \sum_{\mathbf{l}_i \in \mathcal{S}_z} \log\left(\frac{1}{\#\mathcal{S}_z} \sum_{\mathbf{l}_r \in \mathcal{S}_z} \frac{1}{\upsilon} \times \prod_{j=0}^{d-1} \mathbf{K}\left(\frac{\ell_{i,j} - \ell_{r,j}}{h_j}\right)\right) \ ,$$

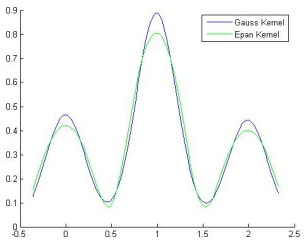
In our attack simulations, we chose the Kernel function to be the Epanechnikov one defined for every u by $\mathbf{K}(u) = \frac{3}{4}(1-u^2)$ if $|u| \leq 1$ and by $\mathbf{K}(u) = 0$ otherwise (another usual choice is the Gaussian Kernel [14]). Our choice was motivated not only by the fact that this Kernel function has a simple form, but also by the fact that its efficiency is asymptotically optimal among all the Kernels [16]. Let $\hat{\sigma}_j$ denotes the estimated standard deviation of the sample $(\ell_{i,j})_i$ of size N_j. To select the Kernel bandwidth h_j, we followed the *normal scale rule* [13]. Namely, we chose the h_j's s.t. $h_j = 1.06 \times \hat{\sigma}_j \times N_j^{-\frac{1}{5}}$.

Simulations. In order to illustrate the effectiveness of the Kernel method, we applied it for the same simulated traces used for our 1st and 2nd order Histogram experiments (Fig. 1 and Fig. 2). We present our results in Fig. 3(a–b) for the first order and in Fig. 3(c–d) for the second order.

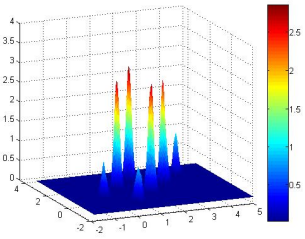
As expected, the pdf estimated in Fig. 3(a) when $k = 11$ seems to be a Gaussian one, whereas the pdf estimated when $k = 5$ seem to be a mixture



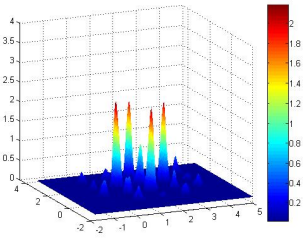
(a) 1st order for $k = 11$.



(b) 1st order for $k = 5$.



(c) 2nd order for $k = 11$.



(d) 2nd order for $k = 5$.

Fig. 3. Kernel Method

of three Gaussian distributions. Moreover, the estimations are smoother than in the case of the Histogram Method and there is no noticeable differences between the estimation with Gaussian Kernel and the estimation with the Epanechnikov one. For the experimentation described in the left-hand figure we obtained $\mathrm{H}(L(11)|Z(11) = 1) = -0.88$ and we got 0.54 for $\mathrm{H}(L(11)|Z(5) = 1)$ (right-hand side).

As expected, in Fig. 3(c) the mixture of Gaussian distributions for $k = 11$ have less components than for $k = 5$. For the experimentation in the left-hand figure we obtained $\mathrm{H}(\mathbf{L}(11)|Z(11)) = 0.17$, whereas we got 0.52 for $\mathrm{H}(\mathbf{L}(11)|Z(5))$. Moreover, we validated that the conditional entropy $\mathrm{H}(\mathbf{L}(11)|Z(k))$ is minimum for $k = k^\star = 11$.

5.3 Parametric Estimation

Description. Under Assumption 3, (12) shows that $g_{L|Z=z}$ is a GM g_θ whose parameter θ satisfies:

$$\theta = \left(\frac{1}{\#E_k(z)}, \varphi \circ f(x, k^\star), \sigma^2 \right)_{x \in E_k(z)} . \tag{19}$$

Based on this relation, an alternative to the methods presented above is to compute an estimation $\hat{\theta}$ of the parameter θ so that we get $\hat{g}_{L|Z=z} = g_{\hat{\theta}}$ and

thus:

$$\hat{\mathrm{H}}(\mathbf{L}|Z=z) = -\int_{\mathbf{l}\in\mathcal{L}^d} g_{\hat{\theta}}(\mathbf{l}) \log_2 g_{\hat{\theta}}(\mathbf{l}) d\mathbf{l} \ .$$

For every x, the mean value $\varphi \circ f(x, k^\star)$ in (19) can be estimated by $\bar{\mathbf{l}}_x = \frac{1}{\#\{i;x_i=x\}} \sum_{i;x_i=x} \mathbf{l}_i$ and the noise variance σ^2 by $\hat{\sigma}^2 = \sum_i \left(\mathbf{l}_i - \bar{\mathbf{l}}_{x_i}\right)^2$. On the whole, this provides us with the following estimation $\hat{\theta}$ of θ:

$$\hat{\theta} = \left(\frac{1}{\#E_k(z)}, \bar{\mathbf{l}}_x, \hat{\sigma}^2\right)_{x\in E_k(z)} \ .$$

For Higher Order MIA, (15) can be rewritten:

$$g_\theta = \frac{1}{\#E_k(z)} \sum_{x\in E_k(z)} g_{\theta_x} \ , \tag{20}$$

where g_{θ_x} denotes the GM pdf of the r.v. $(\mathbf{L}|X=x)$ whose parameter satisfies:

$$\theta_x = \left(\frac{1}{(\#Im(f))^{d-1}}, \Phi_k(x,\mathbf{m}), \Sigma\right)_{\mathbf{m}\in Im(f)^{d-1}} \ .$$

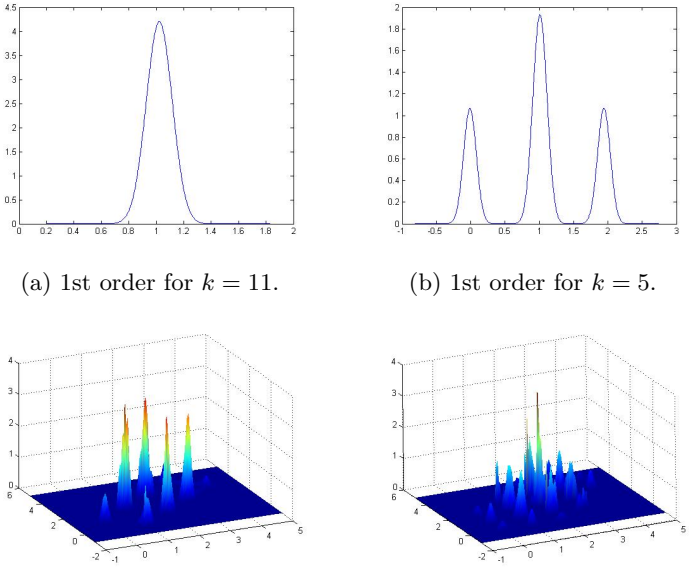
The mean values $\Phi_k(x,\mathbf{m})$ of the different components cannot be directly estimated as in the first order case since the values taken by the masks $\mathbf{m}$ for the different leakage observations $\mathbf{l}_i$ are not assumed to be known. To deal with this issue, a solution is to involve GM estimation methods such as the *Expectation Maximization Algorithm*. By applying it on the sample $(\mathbf{l}_i \ ; \ x_i = x)_i$ we get an estimation $\hat{\theta}_x$ of θ_x for every $x \in \mathcal{X}$. Then, according to (20), we obtain:

$$\hat{\mathrm{H}}(\mathbf{L} \mid Z=z) = -\sum_x \int_{\mathbf{l}\in\mathcal{L}^d} g_{\hat{\theta}_x}(\mathbf{l}) \log g_{\hat{\theta}_x}(\mathbf{l}) d\mathbf{l} \ .$$

Remark 4. As an advantage of the Parametric estimation method, the mean values $\mathbf{l}_x$'s (resp. the estimated parameters $\hat{\theta}_x$'s) are only computed once for every x and are then used to compute $\hat{\mathrm{H}}(\mathbf{L}|Z(k)=z)$ for every pair (k,z).

Simulations. As for the previous estimation methods, we applied the Parametric estimation to the same simulated traces. The resulting estimated pdfs $(\hat{g}_{\mathbf{L}(11)|Z(k)=1})_{k=11,5}$ are plotted in Fig. 4(a–b) for the first order and in Fig. 4(c–d) for the second order.

The results are similar to those of the previous estimation methods. For the first order case, we distinguish a mixture of three Gaussian distributions for the wrong key hypothesis while a single Gaussian pdf is observed for the correct one. For the second order case, the GM obtained for the wrong key hypothesis contains more components than the one for the correct key hypothesis. Once again, the estimated entropy is lower for the correct key hypothesis than for the wrong one. For instance, the entropies of the plotted pdfs equal -0.94 (correct hyp.) and 0.13 (wrong hyp.) for the first order case and 0.24 (correct hyp.) and 0.60 (wrong hyp.) for the second order case.



(a) 1st order for $k = 11$.

(b) 1st order for $k = 5$.

(c) 2nd order for $k = 11$.

(d) 2nd order for $k = 5$.

Fig. 4. Parametric Estimation

6 Experimental Results

6.1 First Order Attack Simulations

To compare the efficiency of the MIA with respect to the estimation method, we simulated leakage measurements in the Gaussian model (7) with φ being the Hamming weight function and f being the first DES S-box (we therefore have $n = 6$ and $m = 4$). For various noise standard deviations σ and for the estimation methods described in previous sections, we estimated the number of messages required to have an attack first order success rate greater than or equal to 90% (this success rate being computed for 1000 attacks). Moreover, we included the first Order CPA in our tests to determine whether and when an MIA is more efficient than a CPA[5]. Each attack was performed with $\hat{\varphi}$ being the identity function in order to test the context in which the attacker has no knowledge about the leakage model. Moreover, each attack was also performed with $\hat{\varphi}$ being the Hamming weight function in order to test the context where the attacker has a good knowledge of the leakage model. The results are given in Table 1 where MIA_H, MIA_K and MIA_P respectively stand for the Histogram, the Kernel and the Parametric MIA.

[5] Attacks have been performed for measurements numbers ranging over 50 different values from 30 to 10^6.

Table 1. Attack on the first DES S-box – Number of measurements required to achieve a success rate of 90% according to the noise standard deviation σ.

Attack \ σ	0.5	1	2	5	10	15	20	50	100
CPA, $\hat{\varphi} = \mathrm{Id}$	30	30	100	1000	3000	7000	15000	70000	260000
MIA_H (Hist), $\hat{\varphi} = \mathrm{Id}$	80	160	600	4000	20000	50000	95000	850000	10^6+
MIA_K (Kernel), $\hat{\varphi} = \mathrm{Id}$	70	140	500	3000	15000	35000	60000	500000	10^6+
MIA_P (Param.), $\hat{\varphi} = \mathrm{Id}$	60	100	300	2000	5000	15000	20000	150000	500000
CPA, $\hat{\varphi} = \mathrm{HW}$	30	30	70	400	2000	4000	7000	45000	170000
MIA_H (Hist), $\hat{\varphi} = \mathrm{HW}$	40	70	300	1500	7000	20000	40000	320000	10^6+
MIA_K (Kernel), $\hat{\varphi} = \mathrm{HW}$	30	60	190	1500	5500	15000	25000	190000	900000
MIA_P (Param.), $\hat{\varphi} = \mathrm{HW}$	70	70	150	1000	3000	7000	15000	65000	300000

It can be checked in Table 1 that the CPA is always better than the MIA when $\hat{\varphi} = \mathrm{HW}$. This is not an astonishing result in our model, since the deterministic part of the leakage corresponds to the Hamming weight of the target variable. More surprisingly, this stays true when $\hat{\varphi}$ is chosen to be the identity function. This can be explained by the strong linear dependency between the identity function and the Hamming weight function over $\mathbb{F}_2^4 = \{0, \ldots, 15\}$. Eventually, both results suggest that the CPA is more suitable than the MIA for attacking a device leaking first order information in a model close to the Hamming weight model with Gaussian noise. When looking at the different MIAs, we can notice that MIA_P becomes much more efficient than MIA_H and MIA_K when the noise standard deviation increases.

6.2 Second Order Attack Simulations

In a CPA, the attacker computes Pearson correlation coefficients which is a function of two univariate samples. Thus, when the CPA is applied against dth order masking (see (14)) a multivariate function must be defined to combine the different leakage signals (corresponding to the masked data and the masks) [9]. This signal processing induces an information loss which strongly impacts the Higher order CPA efficiency when the noise is increasing. Because an Higher Order MIA can operate on multivariate samples, it does not suffer from the aforementioned drawback. We can therefore expect the MIA to become more efficient than the CPA when it is performed against masking. To check this intuition, we simulated power consumption measurements such as in (14) with $d = 2$, with $\varphi = \varphi_1 = \mathrm{Id}$, with $\sigma_1 = \sigma_2 = \sigma$ and with f being the first DES S-box. For various noise standard deviations σ and for the estimation methods described in previous sections, we estimated the number of measurements required to have an attack success rate greater than or equal to 90% (this success rate being computed over 100 attacks). In the following table, we compare second order

MIA with Histogram estimation method (2O-MIA$_H$) with second order CPA (2O-CPA) for two different *combining function*[6].

Table 2. Second Order Attack on DES S-box – Number of measurements required to achieve a success rate of 90% according to the noise standard deviation σ.

Attack σ	0.5	1	2	5	7	10
2O-CPA ($\hat{\varphi}$ = HW, abs. diff. combining))	300	800	5000	200000	10^6+	10^6+
2O-CPA ($\hat{\varphi}$ = HW, norm. product combining)	300	400	3000	70000	300000	10^6+
2O-MIA$_H$ ($\hat{\varphi}$ = Id)	7000	7000	8000	15000	30000	55000

The results presented in Table 2 corroborate our intuition: when the noise standard deviation crosses the threshold 4, second order MIA attacks become much more efficient than second order CPA even for leakage measurements simulated in the Gaussian Model with φ = HW which is favorable to CPA-like attacks.

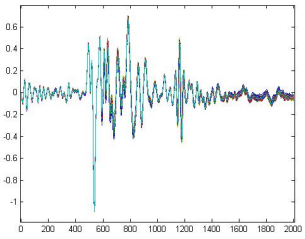
6.3 Practical Attacks

To test the MIA in a real-life context, we performed it against two AES S-box implementations that use a lookup-table (*i.e.* f_k corresponds to the AES S-box). The first one is a hardware implementation on the chip SecMat V3/2 (see [17] for details about the chip and the circuit's layout). The corresponding power consumption measurements are plotted in Fig. 5(a) over the time. It can be noticed that they are not very noisy. The second one is a software implementation running on a 8-bit architecture smart card. As it can be seen in Fig. 6(a), the signal is much more noisy in this case.

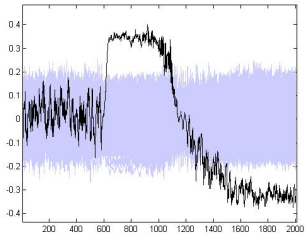
For both set of traces, we performed the CPA and the MIA attacks with the Histogram estimation method and the Parametric estimation method (see Sec. 5). For all of these attacks the prediction function $\hat{\varphi}$ was chosen to be the Hamming weight function (since $\hat{\varphi} \circ f_k$ must be non-injective – see Corollary 1 –). The obtained correlation and mutual information curves are plotted in Fig. 5(b–d) and Fig. 6(b–d) over the time. For each attack the curve corresponding to the correct (resp. wrong) key hypothesis is drawn in black (resp. gray).

In both cases, the attacks succeed with a few number of traces. It can be noticed that the MIA with a Parametric estimation is more discriminating than the MIA with the Histogram estimation. This confirms the simulations performed in Sec. 6.1. However, even when the Parametric estimation method is involved, the CPA is always more discriminating than the MIA. Those results suggest that for the attacked devices the power consumption has in fact a high linear dependency with the Hamming weight of the manipulated data. This implies in particular

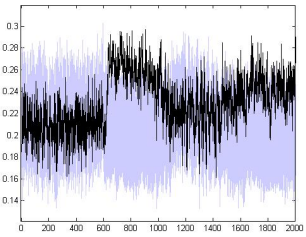
[6] In our simulations we performed 2O-CPAs involving either the absolute difference or the normalized product combining [9].



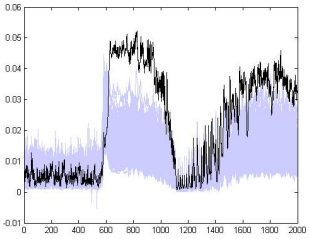
(a) Power Consumption traces.



(b) CPA(HW) attack with 256 traces.



(c) MIA (Hist) attack with 1024 traces.



(d) MIA (param.) attack with 1024 traces.

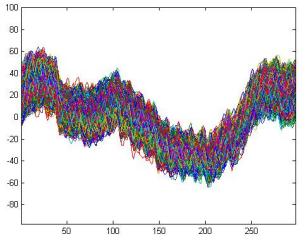
Fig. 5. Practical Attacks on a Hardware AES Implementation

that the Hamming weight Model is sound in this context and that looking for non-linear dependencies is not useful.

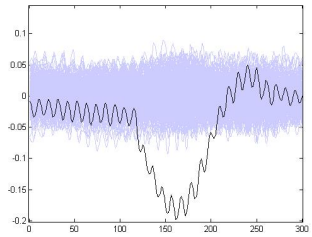
To corroborate that the leakage measured in Fig. 5 and 6 are close to the one simulated in Sec. 5, we plotted in Fig. 7 the estimation of the pdf $g_{\mathbf{L}(0)|Z(k)=1}$ when $k = 0 = k^\star$ and $k = 5 \neq k^\star$ for the hardware implementation. We could verify that actually the conditional pdfs that are estimated look like GM pdfs (a Gaussian pdf when $k^\star$ is correctly guessed and a mixture of two pdfs when it is not).

7 Conclusion

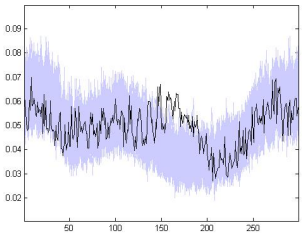
This paper extends the works published in [4] and [5] to expose the theoretical foundations behind the attack and it generalizes it to higher orders. This analysis clarifies assets and limitations of the MIA. In particular, it shows that the MIA is less efficient than the CPA when the deterministic part of the leakage is a linear function of the prediction made by the attacker. This implies that the CPA must be preferred to the MIA when the targeted device leaks a linear function of the Hamming weight of the manipulated data. This paper also argues that the way to estimate the mutual information has an impact on the attack efficiency. A parametric estimation method has been introduced which renders the MIA efficiency close to the one of the CPA when the noise is increasing. When



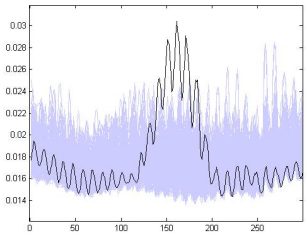
(a) Power Consumption traces.



(b) CPA(HW) attack with 2000 traces.



(c) MIA (Hist) attack with 2000 traces.



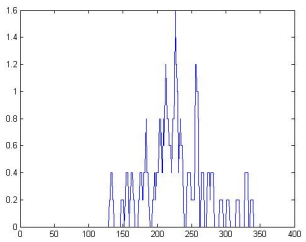
(d) MIA (Param.) attack with 2000 traces.

Fig. 6. Practical Attacks on a Software AES Implementation

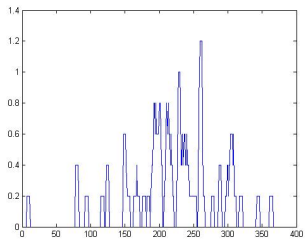
masking is used to protect the implementation, an extension of the MIA has been proposed which is, for our simulations, much more efficient than classical higher order CPA. It actually seems that this is the context in which the MIA offers an efficient alternative to correlation-based attacks.

References

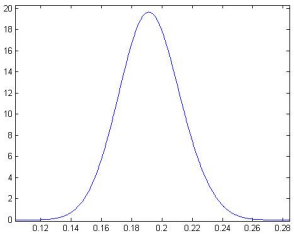
1. Kocher, P., Jaffe, J., Jun, B.: Differential Power Analysis. In CRYPTO '99. Volume 1666 of LNCS. Springer (1999), 388–397.
2. Brier, E., Clavier, C., Olivier, F.: Correlation Power Analysis with a Leakage Model. In CHES 2004. Volume 3156 of LNCS. Springer (2004), 16–29.
3. Chari, S., Rao, J., Rohatgi, P.: Template Attacks. In CHES 2002. Volume 2523 of LNCS. Springer (2002), 13–29.
4. Gierlichs, B., Batina, L., Tuyls, P., Preneel, B.: Mutual information analysis. In CHES 2008. Volume 5154 of LNCS. Springer (2008), 426–442.
5. Aumonier, S.: Generalized Correlation Power Analysis. Published in the Proceedings of the Ecrypt Workshop Tools For Cryptanalysis 2007 (2007).
6. Carreira-Perpinan, M.: Mode-finding for mixtures of Gaussian distributions Carreira-Perpinan. Pattern Analysis and Machine Intelligence, IEEE Transactions **22**(11) (November 2000) 1318–1323.
7. Standaert, F.X., Malkin, T.G., Yung, M.: A Formal Practice-Oriented Model For The Analysis of Side-Channel Attacks. In EUROCRYPT 2009. To appear.



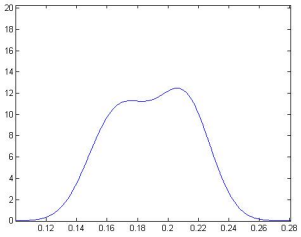
(a) Pdf Estim. by Hist Met. ($k = k^\star$).



(b) Pdf Estim. by Hist Met. ($k \neq k^\star$).



(c) Param. pdf Estim. ($k = k^\star$).



(d) Param. pdf Estim. ($k \neq k^\star$).

Fig. 7. Pdf estimations on power measurements.

8. Chari, S., Jutla, C., Rao, J., Rohatgi, P.: Towards Sound Approaches to Counteract Power-Analysis Attacks. In CRYPTO '99. Volume 1666 of LNCS. Springer (1999), 398–412.
9. Prouff, E., Rivain, M., Bévan, R.: Statistical Analysis of Second Order Differential Power Analysis. IEEE Transactions on Computers **99**(1) (january 2009).
10. Canteaut, A., Trabbia, M.: Improved Fast Correlation Attacks Using Parity-Check Equations of Weight 4 and 5. In EUROCRYPT 2000. Volume 1807 of LNCS. Springer (2000), 573–588.
11. Turlach, B.A.: Bandwidth selection in kernel density estimation: A review. In: CORE and Institut de Statistique. (1993) 23–493.
12. Wand, M.P.: Data-based choice of histogram bin width. The American Statistician **51** (1997) 59–64.
13. Silverman, B.: Density Estimation for Statistics and Data Analysis. Chapman and Hall (1986).
14. Wasserman, L.: All of Statistics: A Concise Course in Statistical Inference. Springer Texts in Statistics (2005).
15. Beirlant, J., Dudewicz, E.J., Györfi, L., Meulen, E.C.: Nonparametric entropy estimation: An overview. International Journal of the Mathematical Statistics Sciences **6** (1997) 17–39.
16. Gray, A.G., Moore, A.W.: Nonparametric density estimation: Toward computational tractability. In SIAM (2003).
17. Guilley, S., Sauvage, L., Hoogvorst, P., Pacalet, R., Bertoni, G.M., Chaudhuri, S.: Security Evaluation of WDDL and Seclib Countermeasures Against Power Attacks. IEEE Transactions on Computers **57**(11) (november 2008) 1482–1497.