Theoretical and practical aspects of mutual information-based side channel analysis

E. Prouff*

Oberthur Technologies,
71-73, rue des Hautes Pâtures, 92726 Nanterre Cedex, France
E-mail: e.prouff@oberthur.com
*Corresponding author

M. Rivain

CryptoExperts,
37 Cours de Vincennes, F-75020 Paris, France
E-mail: matthieu.rivain@cryptoexperts.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. At CHES 2008, Gierlichs et al. proposed to use the mutual information measure as an alternative to the correlation coefficient since it detects any kind of statistical dependency. Substituting it for the correlation coefficient may indeed be considered as a natural extension of the existing attacks. Nevertheless, the first published applications have raised several open issues. In this paper, we conduct a theoretical analysis of MIA in the Gaussian leakage model to explore the reasons why and when it is a sound key recovery attack. Also, we generalise MIA to higher-orders (i.e., against masked implementations). Secondly, we address the main practical issue of MIA: the mutual information estimation which itself relies on the estimation of statistical distributions. We describe three classical estimation methods and we apply them in the context of MIA. Eventually, we present various attack simulations and practical attack experiments that allow us to check the efficiency of MIA in practice and to compare it to classical correlation-based attacks.

Keywords: applied cryptography; embedded security; side channel analysis; SCA; mutual information analysis; MIA; density estimation.

Reference to this paper should be made as follows: Prouff, E. and Rivain, M. (2010) 'Theoretical and practical aspects of mutual information-based side channel analysis', *Int. J. Applied Cryptography*, Vol. 2, No. 2, pp.121–138.

Biographical notes: Emmanuel Prouff received his PhD in Computer Science and Applied Mathematics from the University of Caen, France, in 2004. He has been a University Lecturer in Computer Science at the University of Orsay (Paris XI) and at the French Engineering School of Bourges. He is currently managing the cryptography and security research activities for Oberthur Technologies. His specific research interests include implementation of cryptographic algorithms in embedded devices and theoretical aspects of symmetric cryptology.

Matthieu Rivain received his MS from the Grande Ecole of Computer Science and Applied Mathematics of Grenoble (ENSIMAG) in 2006, and PhD in Computer Sciences from the University of Luxembourg in 2009. During his PhD studies, he was a member of the cryptography team at Oberthur Technologies. He is currently working as a Security Expert within CryptoExperts. His research interests include cryptographic implementations, physical attacks (side channel and fault analysis), elliptic curve cryptography and pairing-based cryptography.

1 Introduction

Side channel analysis (SCA) is a cryptanalytic technique that consists in analysing 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 '90s, 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) (Kocher et al., 1999) that is based on a Boolean correlation and the *correlation power analysis* (CPA) (Brier et al., 2004) that involves *Pearson's correlation coefficient*. The second important family of SCA is composed of the so-called *template attacks* (TA) (Chari et al., 2002). 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.

In Gierlichs et al. (2008) have introduced a new kind of SCA called *mutual information analysis* (MIA). It is an interesting alternative to the aforementioned attacks since some assumptions about the adversary can be relaxed. In particular, since it involves the *mutual information* as distinguisher, 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 the advantages of MIA, the preliminary work of Gierlichs et al. (2008) 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., CPA). This questioning gives rise to a more fundamental issue which concerns with the relationship between a good statistical dependency estimator and a good key-distinguisher. The first attack experiments presented in Gierlichs et al. (2008) suggest that MIAs' 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 Gierlichs et al. (2008). This problematic has been dealt with in statistics and applied probability [see i.e., Aumonier (2007) for an overview]. Among the existing estimation methods, it is of crucial interest to determine the one that optimises MIA. Only such a study will indeed allow us to form an unbiased opinion about its efficiency *versus* that 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. $(X_1,\dots,X_n)\in\mathcal{X}^n$, 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 $\mathbf{X}$ is continuous, one associates it with its *probability density function* (pdf for short), denoted by $g_{\mathbf{X}}$: for every $\mathbf{x}\in\mathcal{X}^n$, we have

$$\mathrm{p}_{\mathbf{X}}\left[\mathbf{X}_1\le x_1,\dots,X_n\le x_n\right]=\int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_n}g_{\mathbf{X}}\left(t_1,\dots,t_n\right)dt_1\dots dt_n.$$

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:

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

where μ and Σ respectively denote the *mean* and the *covariance matrix* of $\mathbf{X}$. When X is unidimensional, its covariance matrix is composed of a single element that is the *variance* of X. It is usually denoted by σ^2, where σ is the *standard deviation* of 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*, is denoted by g_θ and it is defined for every $\mathbf{x}\in\mathcal{X}^n$ by:

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

where $\theta=\left(\left(a_t,\mu_t,\Sigma_t\right)\right)_{1\le t\le T}$ 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* $\mathbf{H}(\mathbf{X})$ 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

$$\mathbf{H}(\mathbf{X})=-\sum\nolimits_{\mathbf{x}\in\mathcal{X}^n}\mathrm{p}_{\mathbf{X}}(\mathbf{x})\log_2\left(\mathrm{p}_{\mathbf{X}}(\mathbf{x})\right).$$

The *differential entropy* extends the notion of entropy to continuous n-dimensional r.v. It is defined by:

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

Note that contrary to the entropy, the differential entropy can be negative.

If $\mathbf{X}$ is a n-dimensional Gaussian r.v. with covariance matrix Σ, then its entropy satisfies:

$$\mathbf{H}(\mathbf{X})=\frac{1}{2}\log\left(\left(2\pi e\right)^n\,|\,\Sigma\,|\right).\quad(4)$$

In the general case, there is no analytical expression for the differential entropy of a r.v. $\mathbf{X}$ whose pdf mixes more than one Gaussian pdf. However, upper and lower bounds can be derived. We recall hereafter the lower bound.

Proposition 2.1: (Carreira-Perpinan, 2000) Let $\mathbf{X} \in \mathcal{X}^n$ be a Gaussian mixture whose pdf g_θ is defined by $\theta = \left(\left(a_i, \mu_i, \Sigma_i\right)\right)_{i=1,\ldots,T}$. Then, its differential entropy satisfies:

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

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 $\mathbf{I}(\mathbf{X};\mathbf{Y})$ defined by $\mathbf{I}(\mathbf{X};\mathbf{Y}) = \mathbf{H}(\mathbf{X}) - \mathbf{H}(\mathbf{X} \mid \mathbf{Y})$, where $\mathbf{H}(\mathbf{X} \mid \mathbf{Y})$ is called the conditional entropy of $\mathbf{X}$ knowing $\mathbf{Y}$. If $\mathbf{Y}$ is discrete, then $\mathbf{H}(\mathbf{X} \mid \mathbf{Y})$ is defined by:

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

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

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:

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

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 i.e., be an 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 (Prouff, 2005).

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 $\hat{\varphi} \circ f_k(X)$ and $L\left(k^\star\right)$. 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. Assuming a Gaussian noise, this 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 Chari et al., 2002). To pre-compute the pdfs of all the variables $L(k)$, the attacker needs to have an access to a copy of the device under attack for which he can set (or at least know) the secret key. This is a strong requirement which is rarely fulfilled in practice.

As noticed in Gierlichs et al. (2008) and Aumonier (2007), MIA attacks are an alternative to the approaches above. They consist in estimating the mutual information $\mathrm{I}\left(L\left(k^\star\right); \hat{\varphi} \circ f_k(X)\right)$ 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 a 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 i.e., allows the attacker to choose the identity function for $\hat{\varphi}$ which is of particular interest since no knowledge about the leakage parameter is required.

The effectiveness of a key-recovery side channel attack is usually characterised by its *success rate*, namely the probability that the attack outputs the correct key as the most likely key candidate. This notion can be extended to higher-orders (Standaert et al., 2009): an attack is said to be oth *order successful* if it classifies the correct key among the o most likely key candidates. In the following, we shall investigate the (oth order) success rate of MIA.

Notations: In what follows, the r.v. $\hat{\varphi} \circ f_k(X)$ shall be denoted by $Z(k)$ and the set $\left[\hat{\varphi} \circ f_k\right]^{-1}(z)$ shall be denoted by $E_k(z)$. For clarity reasons, we shall further denote by L (resp. by Z) the r.v. $L(k^\star)$ (resp. $Z(k)$) when there is no ambiguity.

For every function F defined over $\mathcal{K}$, let us denote by $\arg\max - o_{k\in\mathcal{K}} F(k)$ 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 oth order if the estimations $\hat{\mathrm{I}}(L;Z(k))$ of $\mathrm{I}(L;Z(k))$ satisfy:

$$k^\star \in \arg\max_{k\in\mathcal{K}} - o \ \hat{\mathrm{I}}(L;Z(k)). \tag{8}$$

We therefore deduce two necessary conditions for an MIA to succeed at the oth order:

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

  $$k^\star \in \arg\max_{k\in\mathcal{K}} - o \ \mathrm{I}(L;Z(k)). \tag{9}$$

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

When the attacker is assumed to make a perfect guess on the deterministic function of the leakage (i.e., when he chooses $\hat{\varphi} = \varphi$), then $\mathrm{I}\left(L;Z\left(k^\star\right)\right) \geq \mathrm{I}(L;Z(k))$ holds for every $k\in\mathcal{K}$ [which is a necessary but not sufficient condition to (9)]. This has been argued in Moradi et al. (2009) by pointing out the existence of the following Markov chain: $Z(k)\to Z(k^\star)\to L$. In the general case, when $\hat{\varphi}$ may differ from φ, this Markov chain does not necessarily exist and no general statement about (9) is possible. In order to go deeper in the analysis, we conduct in the next section a theoretical study of MIA in the Gaussian model. This will allow us to characterise (with regards to $f,\varphi,\hat{\varphi}$) to what extent (9) may be satisfied. In a second time, we address the practical issue of MIA: the mutual information estimation. We describe in Section 5 several classical estimation methods and we apply them in the context of MIA. For 3-tuples $(f,\varphi,\hat{\varphi})$ s.t. (9) is satisfied, we eventually investigate in Section 6 the success probability of MIA according to the estimation method and according to the noise variation. Those analyses will allow us to characterise when an MIA is practically feasible (i.e., when (8) is satisfied) and to compare its efficiency with that of other SCA attacks.

4 Study of 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 (Chari et al., 1999). Our analyses are done under the three following assumptions which are realistic in a SCA context and make the formalisation 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. $\#\left\{x\in\mathbb{F}_2^n; y = f_k(x)\right\}$ equals 2^{n-m} for every $y\in\mathbb{F}_2^m$.

Remark 4.1: 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 4.2: This assumption is realistic and is therefore often done in the literature (see i.e., Chari et al., 1999; Prouff et al., 2009; Standaert et al., 2009). Practical attacks and pdf estimations presented in Section 6 provide us with an experimental validation of this assumption.

4.1 First-order MIA

The mutual information $\mathrm{I}(L;Z(k))$ equals

$$\mathrm{H}(L) - \mathrm{H}(L \mid Z(k)).$$

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 $\mathcal{K}$ if the conditional entropy $\mathrm{H}(L\mid Z(k))$ reaches one of its o smallest values. One deduces that an MIA is theoretically possible if the 3-tuple $(f,\varphi,\hat{\varphi})$ is s.t.:

$$k^\star \in \arg\min_{k\in\mathcal{K}} - o \ \mathrm{H}(L\mid Z(k)), \tag{10}$$

where $\arg\min - o$ is defined analogously to $\arg\max - o$.

The starting point of our analysis is that studying the MIA effectiveness is equivalent to investigating the minimality of $\mathrm{H}(L\mid Z(k))$ over $\mathcal{K}$. As a consequence of (6), we have

$$\mathrm{H}(L\mid Z(k)) = \sum\nolimits_{z\in\mathrm{Im}(\hat{\varphi})} \mathrm{p}_{Z(k)}(z)\mathrm{H}(L\mid Z(k) = z).$$

Since Z equals $\hat{\varphi}\circ f_k(X)$, the probabilities $\mathrm{p}_{Z(k)}(z)$ in this sum can be easily computed by the attacker and the main difficulty therefore essentially lies in the computation of the $\mathrm{H}(L\mid Z(k) = z)$'s. From (3), one deduces:

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

To reveal the relationship between $\mathrm{H}(L\mid Z(k))$ and the key-prediction k, the expression of the pdf $g_{L|Z=z}$ in (11) needs to be developed. 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:

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

where we recall that $E_k(z)$ denotes $[\hat{\varphi} \circ f_k]^{-1}(z)$. The next proposition directly follows.

Proposition 4.1: For every pair $(k^\star, k) \in \mathcal{K}^2$ and every $z \in \mathcal{Z}$ the pdf of the r.v. $\left(L\left(k^\star\right) \mid Z(k) = z\right)$ is a Gaussian mixture g_θ whose parameter θ satisfies:

$$\theta = \left(\left(a_{z,t}, t, \sigma^2\right)\right)_{t \in \mathrm{Im}(\varphi)},$$

with

$$a_{z,t} = \frac{\#\left\{x \in E_k(z); \varphi \circ f_{k^\star}(x) = t\right\}}{\# E_k(z)}.$$

In Proposition 4.1, the key hypothesis k only plays a part in the definition of the weights $a_{z,t}$ of the Gaussian mixture. 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 $\mathrm{H}(L \mid Z(k) = z)$, let us introduce the following diagram where z is an element of $\mathrm{Im}(\hat{\varphi})$, where F', F and T are image sets:

$$\begin{tikzcd} z \arrow[r, "\hat{\varphi}^{-1}"] & F' \arrow[r, "f_k^{-1}"] & E_k(z) \arrow[r, "f_{k^\star}"] & F \arrow[r, "\varphi"] & T, \end{tikzcd}$$

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 4.1 that the distribution of $(L \mid Z(k) = z)$ is Gaussian and, due to (4), its conditional entropy satisfies:

  $$\mathrm{H}(L \mid Z(k) = z) = \frac{1}{2}\log\left(2\pi e\sigma^2\right).$$

- If $\#T > 1$ $\left(\text{i.e., if } \#\varphi \circ f_{k^\star}(E_k(z)) > 1\right)$, then there exist at least two probabilities a_{z,t_1} and a_{z,t_2} which non-null and the distribution of $(L \mid Z(k) = z)$ is a Gaussian mixture (not Gaussian). Due to (5), its entropy satisfies:

  $$\mathrm{H}(L \mid Z(k) = z) \geq \frac{1}{2}\log\left(2\pi e\sigma^2\right).$$

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 $\mathrm{H}(L \mid Z(k) = z)$ equals $\frac{1}{2}\log\left(2\pi e\sigma^2\right)$. Otherwise, if $k \neq k^\star$, then $f_{k^\star} \circ f_k$ is likely to behave as a random function[1]. In this case, F is most of the time different from F' and T is therefore likely to have more than one element[2]. This implies that $\#\varphi \circ f_{k^\star}(E_k(z))$ is strictly greater than 1 and thus, that $\mathrm{H}(L \mid Z(k) = z)$ is greater than or equal to $\frac{1}{2}\log\left(2\pi e\sigma^2\right)$. Eventually, we get the following proposition in which we exhibit a tight lower bound for the differential entropy $\mathrm{H}(L \mid Z(k))$.

Proposition 4.2: For every $(k^\star, k) \in \mathcal{K}^2$, the conditional entropy of the r.v. $\left(L\left(k^\star\right) | Z(k)\right)$ satisfies:

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

If $\mathcal{K} \circ f_{k^\star}$ 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 2.1 and 4.1. The tightness is a direct consequence of (4) and Proposition 4.1.

Remark 4.3: Intuitively, the entropy is a measure of the diversity or randomness of a random variable. It is therefore reasonable to think that the more components in the Gaussian mixture pdf of $(L \mid Z(k) = z)$, 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 Gaussian mixture has only one component). In our experiments (partially reported in Section 6), we noticed that the entropy of a Gaussian mixture whose components have the same variance, increases with the number of components.

Corollary 1: If $\hat{\varphi} \circ f_k$ is injective, then $\mathrm{H}(L \mid Z(k))$ equals $\frac{1}{2}\log\left(2\pi e\sigma^2\right)$.

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 MIA cannot succeed at any order. Indeed, in this case the entropy $\mathrm{H}(L \mid Z(k))$ 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 i.e., 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 fail). 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 i.e., target the bitwise addition between two S-box outputs during the *MixColumns* operation. When bits are assumed to leak independently, another way suggested in Gierlichs et al. (2008) is to target a restrictive number of the output bits.

4.2 Generalisation to the higher-orders

In this section, we extend the analysis of MIA to higher-orders i.e., we assume that the target implementation is 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 from $f_{k^\star}(X)\oplus M_1\oplus\cdots\oplus M_{d-1}$ is denoted by L_0 and the leakages from the M_j's are denoted by $L_1,...,L_{d-1}$. Under Assumption 3, the L_j's satisfy:

$$L_j = \begin{cases} \varphi_0\left[f_{k^\star}(X)\oplus\bigoplus_{t=1}^{d-1}\right]+B_0 & \text{if } j=0, \\ \varphi_j\left(M_j\right)+B_j & \text{if } j\neq 0, \end{cases} \tag{14}$$

where the B_j's are independent Gaussian noises with mean 0 and standard deviation σ_j, and where $\varphi_0,\varphi_1,\cdots,\varphi_{d-1}$ are d device dependent functions that are a priori unknown to the attacker.

Notations: The leakage vector $(L_0,\cdots,L_{d-1})$ is denoted by $\mathbf{L}\left(k^*\right)$ (or simply $\mathbf{L}$ when there is no ambiguity) and the vector of masks $(M_1,\cdots,M_{d-1})$ is denoted by M. We further denote by $\Phi_{k^*}(X,\mathrm{M})$ the vector

$$\left(\varphi_0\left(f_{k^\star}(X)\oplus\bigoplus_{t=1}^{d-1}M_t\right),\varphi_1\left(M_1\right),\cdots,\varphi_{d-1}\left(M_{d-1}\right)\right).$$

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, an higher-order MIA essentially consists in looking for the key candidate k which minimises an estimation of the conditional entropy $\mathrm{H}(\mathbf{L}\mid Z(k))$. This entropy is 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 $\left(Cov\left[B_i,B_j\right]\right)_{i,j}$, we get:

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

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

Proposition 4.3: For every pair $\left(k^\star,k\right)\in\mathcal{K}^2$ and every $z\in\mathcal{Z}$ the pdf of the r.v. $\left(\mathbf{L}\left(k^\star\right)\middle|Z(k)=z\right)$ is a Gaussian mixture g_θ whose parameter θ satisfies:

$$\theta=\left(\left(a_{z,\mathbf{t}},\mathbf{t},\Sigma\right)\right)_{\mathbf{t}},$$

with

$$\Sigma=\left(Cov\left[B_i,B_j\right]\right)_{i,j},$$

and

$$a_{z,\mathbf{t}}=\frac{\#\left\{(x,\mathbf{m});\Phi_{k^\star}(x,\mathbf{m})=\mathbf{t}\right\}}{\#E_k(z)\left(\#\mathrm{Im}(f)\right)^{d-1}}.$$

We deduce from Propositions 2.1 and 4.3 the following result.

Proposition 4.4: For every $\left(k^\star,k\right)\in\mathcal{K}^2$, the entropy of the r.v. $\left(\mathbf{L}\left(k^\star\right)\middle|Z(k),\mathbf{M}\right)$ satisfies:

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

If $\Phi_{k^\star}(\cdot,\mathbf{m})$ is constant on $E_k(z)$ for every $z\in\mathrm{Im}(\hat{\varphi})$ and for every $\mathbf{m}\in\mathrm{Im}(f)^{d-1}$, 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 not allowed in our context. However, if the 3-tuple $(f,\Phi,\hat{\varphi})$ satisfies the condition of Proposition 4.4, then it can be checked that for every z the number of components in the multivariate Gaussian mixture pdf of $(\mathbf{L}\mid Z(k)=z)$ reaches its minimum for $k=k^\star$. As discussed in Remark 4.3, this implies that the entropy $\mathrm{H}(\mathbf{L}\mid Z(k))$ is likely to be minimum for $k=k^\star$. The simulations and experiments presented in Section 6 provides us with an experimental validation of this fact.

Remark 4.4: As mentioned in Section 3, for the first-order case and assuming $\hat{\varphi}=\varphi$, the existence of the Markov chain $Z(k)\to Z\left(k^\star\right)\to L$ implies $\mathrm{I}\left(L;Z\left(k^\star\right)\right)\geq\mathrm{I}\left(L;Z(k)\right)$ for every $k\in\mathcal{K}$. We do not have such a straightforward statement for the higher-order case where several functions $\varphi_0, \varphi_1, \ldots, \varphi_{d-1}$ are involved in the leakage (14). Nevertheless, using the identity function as $\hat{\varphi}$ yields the following Markov chain

$$Z(k) = f_k(X) \to Z\left(k^\star\right) = f_{k^\star}(X) \to \Phi_{k^\star}(X, \mathbf{M}) \to \mathbf{L}$$

which implies $\mathrm{I}\left(\mathbf{L}; Z\left(k^\star\right)\right) \geq \mathrm{I}\left(\mathbf{L}; Z\left(k\right)\right)$ for every $k \in \mathcal{K}$. This shows the soundness of using the identity function as $\hat{\varphi}$ when the f_k's are non-injective (otherwise the mutual information is constant with respect to k as argued for the first-order case).

In the next sections, we assume that an MIA is theoretically successful at the first-order. Namely, we assume that we have $k^\star = \arg\min_k \mathrm{H}\left(\mathbf{L} \mid Z(k)\right)$. At first, we study the success probability of an MIA according to the method used to estimate $\mathrm{H}\left(\mathbf{L} \mid Z(k)\right)$ and the noise variation. Secondly, we compare the efficiency of MIA with the one of 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 $\left(\mathrm{l}_i, x_i\right) \in \mathcal{L}^d \times \mathcal{X}$ corresponding to a key k^*, and that he wants to compute $\mathrm{H}\left(\mathbf{L} \mid Z(k)\right)$ to discriminate key-candidates k. Due to (6), estimating $\mathrm{H}\left(\mathbf{L} \mid Z(k)\right)$ from the sample $\left(\left(\mathrm{l}_i, x_i\right)\right)_i$ essentially amounts to estimate the entropy $\mathrm{H}\left(\mathbf{L} \mid Z(k) = z\right)$ for every $z \in \mathcal{Z}$. For such a purpose, a first step is to compute estimations $\hat{g}_{\mathbf{L}|Z(k)=z}$ of the pdfs $g_{\mathbf{L}|Z(k)=z}$. Then, depending on the estimation method that has been applied, the entropies $\mathrm{H}\left(\mathbf{L} \mid Z(k) = z\right)$ 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

The density estimation by histogram has been first applied in the original paper of Gierlichs et al. (2008) to experimentally validate the soundness of first-order MIA. An experimental study has been conducted in the paper of Moradi et al. (2009) but still in the context of first-order attacks. In the following, we detail the histogram method in the general case of multivariate density estimations.

5.1.1 Description

We choose d *bin widths* $h_0, \ldots, h_{d-1}$ (one for each coordinate of the leakage vectors) and we partition the leakage space $\mathcal{L}^d$ into regions $\left(\mathcal{R}_\alpha\right)_\alpha$ with equal volume $v = \prod_j h_j$. Let k be a key-candidate and let z be an element of $\mathcal{Z}$. We denote by $\mathcal{S}_z$ the subsample

$$\left(\mathrm{l}_i ; x_i \in \left[\varphi \circ f_k\right]^{-1}(z)\right)_i \subseteq \left(\mathrm{l}_i\right)_i$$

and by $\ell_{i,j}$ the jth coordinate of 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 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 $\mathrm{l} \in \mathcal{L}^d$ by $g_{\mathbf{L}|Z=z}(\mathrm{l}) = \frac{D_z(i_\mathrm{l})}{v}$, where i_l is the index of the region $\mathcal{R}_{i_\mathrm{l}}$ that contains l. Integrating the pdf estimation according to formula (3) gives the following estimation for the conditional entropy:

$$\hat{\mathrm{H}}\left(\mathbf{L} \mid Z = z\right) = -\sum_\alpha D_z(\alpha) \log\left(D_z(\alpha) / v\right).$$

The optimal choice of the bin widths h_j's is an issue in statistics theory. Several rules however exist that start from the nature of the samples to deduce the h_j's or, equivalently, the number of bins (see i.e., Turlach, 1993; Scott, 1992; Wand, 1997). For univariate density estimation, Sturge's rule (the number of bins is chosen equal to $1 + \log_2(N)$) and Scott's rule are often preferred and several works have studied their asymptotical soundness. In (Scott, 1992), a generalisation of Scott's rule, called *normal reference rule*, is given for multivariate density estimation: it consists in defining each bin width h_j s.t.

$$h_j = 3.49 \times \hat{\sigma}_j \times N_j^{-\frac{1}{2+d}}, \tag{18}$$

where $\hat{\sigma}_j$ denotes the estimated standard deviation of the sample $\left(\ell_{i,j}\right)_i$ of size N_j. In our simulations, we chose to apply the normal reference rule because of its simplicity. However, other methods exist and the optimality of some of them is even formally analysed. The interested reader may turn to the paper of Birgé and Rozenholc (2006) where a survey and a comparative simulation are provided.

5.1.2 Simulations

In order to illustrate the histogram method in the context of an MIA attack, we generated 10,000 leakage measurements in the Gaussian model (7) for φ being the Hamming weight function, for f being the first DES S-box parameterised 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 $\hat{\varphi}$. Figure 1 plots the estimations of the pdf $g_{L|Z=1}$ when $k = 11$ and when $k = 5$.

As expected (Proposition 4.1 and Corollary 1), a Gaussian pdf seems to 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 $\hat{\mathrm{H}}\left(L(11) \mid Z(11) = 1\right) = -1.31$ (due to (4) we have $\mathrm{H}\left(L(11) \mid Z(11) = 1\right) = -1.27$) and we got $\hat{\mathrm{H}}\left(L(11) \mid Z(5) = 1\right) = -0.0345$ for the one in the right-hand side. Moreover, we validated that the estimated conditional entropy is minimum for the good key hypothesis.

Figure 1 Histogram estimation method in the first-order case (see online version for colours)

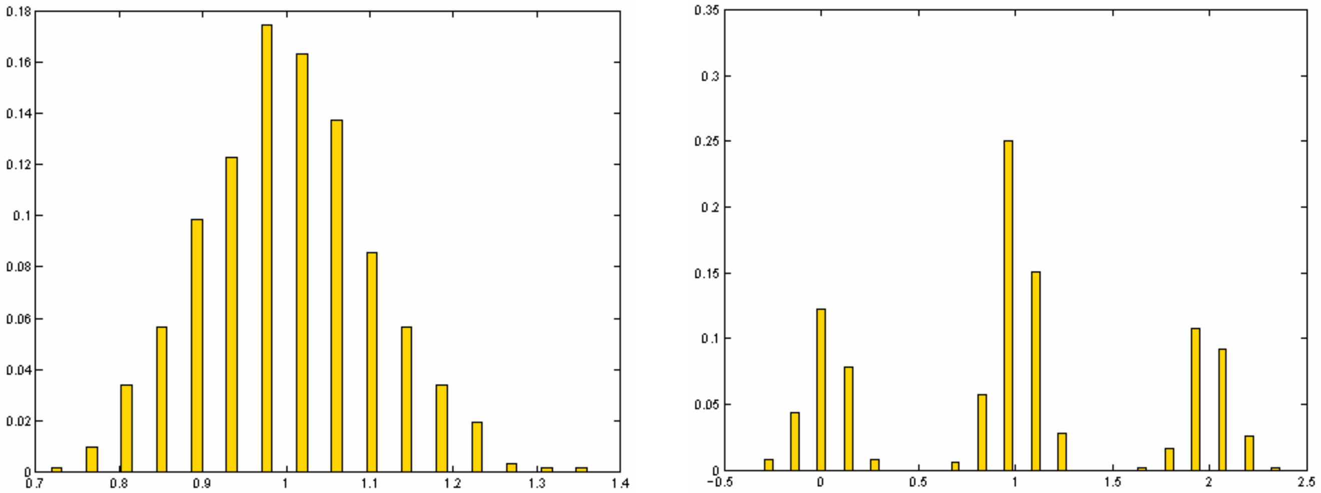
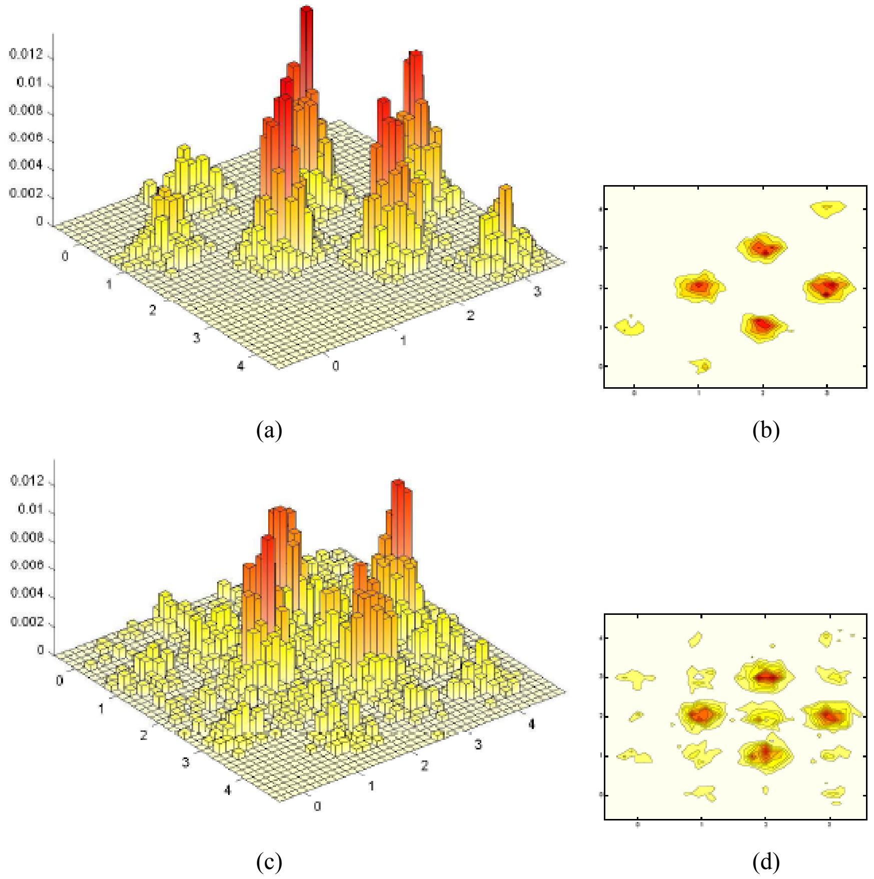


Figure 2 Histogram estimation method in the second-order case, (a) 2nd-order for $k = 11$ (b) 2nd-order for $k = 11$ (c) 2nd-order for $k = 5$ (d) 2nd-order for $k = 5$ (see online version for colours)



(a)

(b)

(c)

(d)

In order to illustrate the histogram method in the context of a 2nd-order MIA attack, we generated 10,000 pairs of leakage measurements in the higher-order Gaussian model (14) with $d = 2$, with φ_0 and φ_1 being the Hamming weight function, with f being the first DES S-box parameterised with the key $k^\star = 11$ and with $\sigma_0 = \sigma_1 = 0.1$. We chose the identity function for $\hat{\varphi}$. Figure 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 $k = 11$ have less components than for $k = 5$. For the experimentation in the left-hand figure we obtained $\hat{\mathrm{H}}(\mathbf{L}(11) \mid Z(11) = 1) = 0.22$ (and $\hat{\mathrm{H}}(\mathbf{L}(11) \mid Z(11)) = 0.14$), whereas we got 1.12 for $\hat{\mathrm{H}}(\mathbf{L}(11) \mid Z(5) = 1)$ (and 1.15 for $\hat{\mathrm{H}}(\mathbf{L}(11) \mid Z(5))$). Here again, the estimated conditional entropy was minimum for the good key hypothesis.

5.2 *Kernel density method*

The density estimation by kernel method has been studied in the context of MIA attacks in Prouff and Rivain (2009) and Veyrat-Charvillon and Standaert (2009). The analysis conducted by Veyrat-Charvillon and Standaert (2009) focuses on univariate data and compares the efficiency of several first-order MIA for different kernel functions. In the following, we detail the kernel method in the general case of multivariate data (i.e., in the context of higher-order MIA).

5.2.1 *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 i.e., Silverman, 1986). The estimator consists of a 'sum of bumps' and is clearly smoother as a result than the histogram method.

The (multivariate) *kernel density estimator* $\hat{g}_{\mathbf{L}|Z=z}$ of the function $g_{\mathbf{L}|Z=z}$ based on the sample $\mathcal{S}_z$ is defined for every $\mathbf{l} \in \mathcal{L}^d$ by:

$$\hat{g}_{L|Z=z}(\mathrm{l}) = \frac{1}{\#\mathcal{S}_z \times h^d} \sum_{\mathrm{l}_i \in \mathcal{S}_z} \mathrm{K}\left(\frac{\mathrm{l} - \mathrm{l}_i}{h}\right), \tag{19}$$

where h is the *kernel bandwidth* (also called *windows width*) and where $\mathbf{K}$ is a (multivariate) *kernel function* defined from $\mathbb{R}^d$ to $\mathbb{R}$ and satisfying:

$$\int_{\mathbb{R}^d} \mathbf{K}(\mathrm{x}) d\mathrm{x} = 1.$$

As explained in Silverman (1986), $\mathbf{K}$ is usually chosen as a radially symmetric unimodal pdf. A classical choice is the *standard (multivariate) normal density* function

$$\mathbf{x} \in \mathbb{R}^d \mapsto \mathbf{K}(\mathbf{x}) = (2\pi)^{-d/2} \exp\left(-\tfrac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{x}\right)$$

or the *(multivariate) Epanechnikov function* defined for every $\mathbf{x} \in \mathbb{R}^d$ by $\mathbf{K}(\mathbf{x}) = 1/2c_d^{-1}(d+2)\left(1 - \mathbf{x}^{\mathrm{T}}\mathbf{x}\right)$ if $\mathbf{x}^{\mathrm{T}}\mathbf{x} < 1$ and by $\mathbf{K}(\mathbf{x}) = 0$ otherwise, with c_d being the volume of the unit d-dimensional sphere ($c_1 = 2, c_3 = \pi$, etc.).

Remark 5.1: The use of a single smoothing parameter h in (19) implies that the version of the kernel placed on each data point is scaled equally in all directions. In certain circumstances, when the distributions of the different leakage points in $\mathbf{L}$ are very different, it may be more appropriate to use a vector of smoothing parameter (i.e., one h_i for each coordinate of $\mathbf{L}$) or even a matrix of *shrinking coefficients* (Silverman, 1986; Wasserman, 2005). In our context, using such generalisations is not of great interest. Indeed, as recalled in Section 4.2, the leakage coordinates are often assumed to be of same nature (thus, the bandwidths must be equal) and to be pairwisely independent (so the shrinking matrix is diagonal).

As recalled in Beirlant et al. (1997), the following Parzen-windows entropy estimation of $\mathrm{H}(\mathbf{L} \mid Z = z)$ is sound when the sample size is large enough:

$$\hat{\mathrm{H}}(\mathbf{L} \mid Z = z) = -\frac{1}{\#\mathcal{S}_z} \sum_{\mathrm{l}_r \in \mathcal{S}_z} \log\left(\hat{g}_{L|Z=z}(\mathrm{l}_r)\right),$$

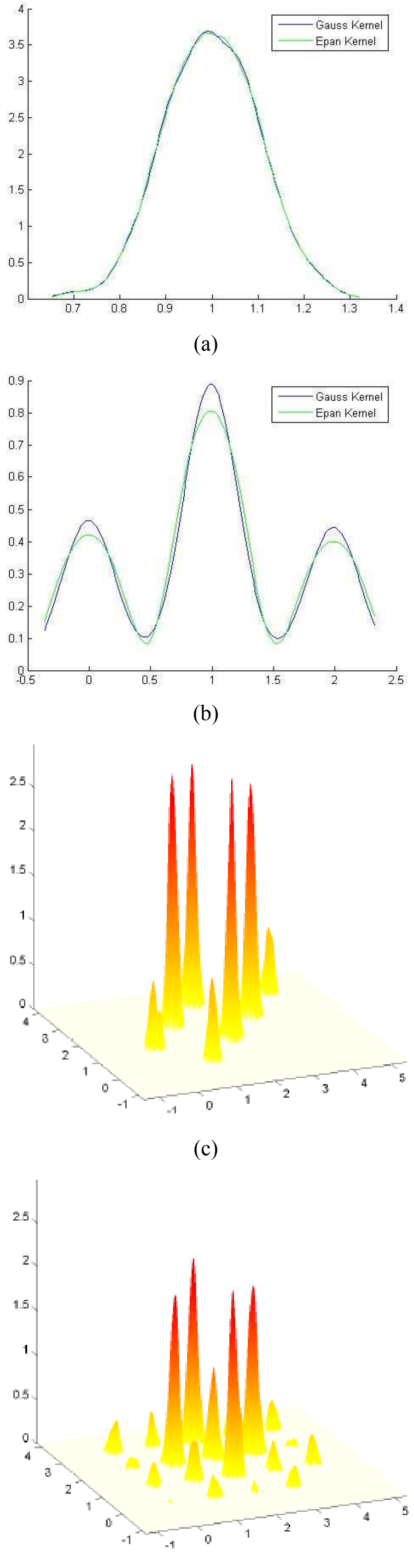
where $\hat{g}_{L|Z=z}(\mathrm{l}_r)$ satisfies (19). In our attack simulations, we chose the kernel function to be the Epanechnikov one. Our choice was motivated not only by the fact that this kernel function has a simple form, but also by the fact that its efficiency is asymptotically optimal among all the kernels (Gray and Moore, 2003). We also assumed that all the coordinates of the leakage vector have the same standard deviation σ and to select the common kernel bandwidth h, we followed two different rules depending on the dimension d of $\mathbf{L}$. For univariate $\mathbf{L}$ (i.e., $d = 1$) we followed the *normal scale rule* (Silverman, 1986): namely, we chose h s.t. $h = 1.06 \times \sigma \times N^{-\frac{1}{5}}$ where N is the sample size and $\hat{\sigma}$ is the sample estimator of σ. For multivariate $\mathbf{L}$ (i.e., $d > 1$) we chose the optimal bandwidth selection that minimises the *mean integrated square error* (see i.e., Scott, 1992; Wasserman, 2005): namely, we chose h s.t.

$$h = \hat{\sigma} \times \left(\frac{4}{(d+2)N}\right)^{1/(d+4)}. \tag{20}$$

5.2.2 *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 (Figure 1 and Figure 2). We present our results in Figure 3(a) to Figure 3(b) for the first-order and in Figure 3(c) to Figure 3(d) for the second-order.

Figure 3 Kernel estimation method, (a) 1st-order for $k = 11$ (b) 1st-order for $k = 5$ (c) 2nd-order for $k = 11$ (d) 2nd-order for $k = 5$ (see online version for colours)



(a)

(b)

(c)

(d)

As expected, the pdf estimated in Figure 3(a) when $k = 11$ seems to be a Gaussian one, whereas the pdf estimated when $k = 5$ seem to be a mixture of three Gaussian distributions. Moreover, the estimations are smoother than in the case of the histogram method and there are 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 $\mathrm{H}(L(11) \mid Z(11) = 1) = -0.88$ and we got 0.54 for $\mathrm{H}(L(11) \mid Z(5) = 1)$ (right-hand side).

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

5.3 *Parametric estimation*

In the following, we present a third pdf estimation method that takes full advantage of the Gaussian noise assumption. As shown in Section 6, this approach yields to a first-order MIA which is more efficient than those based on the histogram and kernel methods.

5.3.1 *Description*

Under the Gaussian noise assumption, the analysis of Section 4 shows that $g_{\mathbf{L}|Z=z}$ is a Gaussian mixture g_θ. An alternative to the methods presented above is therefore to compute an estimation $\hat{\theta}$ of the parameter θ so that we get $\hat{g}_{\mathbf{L}|Z=z} = g_{\hat{\theta}}$ and thus:

$$\hat{\mathrm{H}}(\mathbf{L} \mid Z = z) = -\int_{\mathrm{l}\in\mathcal{L}^d} g_{\hat{\theta}}(\mathrm{l}) \log_2 g_{\hat{\theta}}(\mathrm{l}) d\mathrm{l}. \tag{21}$$

First-order case

According to (12), $g_{L|Z=z}$ is a Gaussian mixture whose parameter θ satisfies:

$$\theta = \left(\frac{1}{\#E_k(z)}, \varphi \circ f\left(x, k^\star\right), \sigma^2 \right)_{x\in E_k(z)}. \tag{22}$$

Since we have $\varphi \circ f\left(x, k^\star\right) = \mathrm{E}[L \mid X = x]$, for every x, the expectation $\varphi \circ f\left(x, k^\star\right)$ can be estimated by:

$$\hat{m}_x = \frac{1}{\#\{i; x_i = x\}} \sum_{i;x_i=x} l_i.$$

And since we have $\sigma^2 = Var[B] = \mathrm{E}\left[\left(L - \varphi \circ f\left(X, k^\star\right)\right)^2\right]$, the variance σ^2 can be estimated by:

$$\hat{\sigma}^2 = \sum_i \left(\mathrm{l}_i - \hat{m}_{x_i}\right)^2 .$$

On the whole, this provides us with the following estimation $\hat{\theta}$ of θ:

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

Higher-order case

For higher-order MIA, (15) can be rewritten:

$$g_\theta = \frac{1}{\# E_k(z)} \sum_{x \in E_k(z)} g_{\theta x}, \qquad (23)$$

where $g_{\theta x}$ denotes the Gaussian mixture pdf of the r.v. $(\mathbf{L} \mid X = x)$ whose parameter satisfies:

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

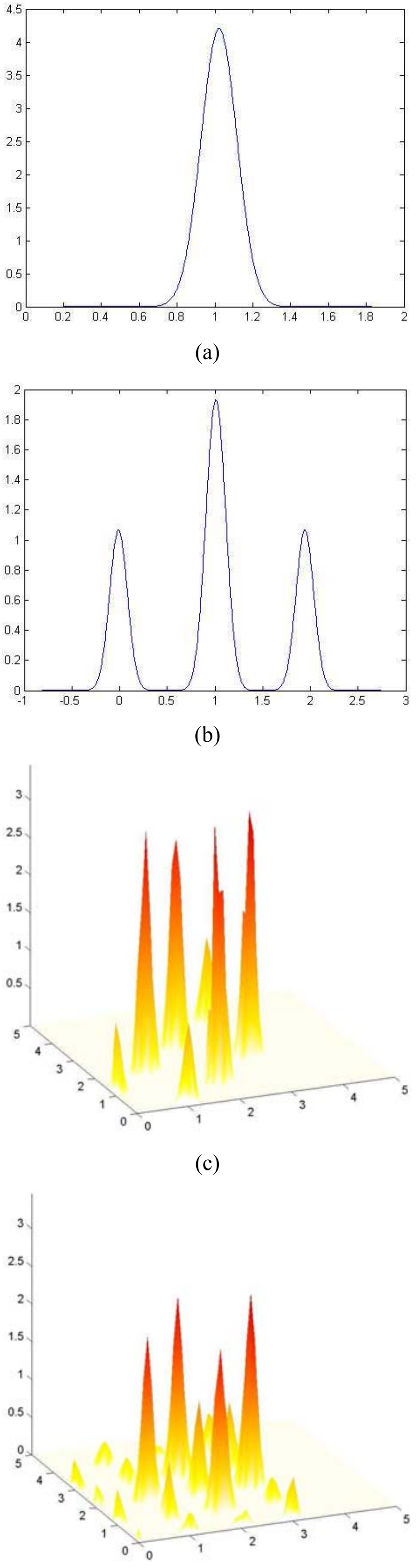
The mean values $\Phi_{k^\star}(x, \mathbf{m}) = \mathrm{E}[\mathbf{L} \mid X = x, \mathrm{M} = \mathbf{m}]$ cannot be directly estimated as in the first-order case since the values $\mathbf{m}_i$ taken by the masks for the different leakage observations $\mathbf{l}_i$ are not assumed to be known. To deal with this issue, a solution is to involve Gaussian mixture estimation methods such as the *expectation maximisation algorithm* (Bishop, 2007). 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, the overall estimation $\hat{\theta}$ directly results from the $\hat{\theta}_x$ following (23), and the conditional entropy can be estimated according to (21).

5.3.2 Simulations

As for the previous estimation methods, we applied the parametric estimation to the same simulated traces. The resulting estimated pdfs $\left(\hat{g}_{\mathbf{L}(11)|Z(11)=1}\right)_{k \in \{5,11\}}$ are plotted in Figure 4(a) to Figure 4(b) for the first-order and in Figure 4(c) to Figure 4(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 Gaussian mixture 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.

Figure 4 Parametric estimation method, (a) 1st-order for $k = 11$ (b) 1st-order for $k = 5$ (c) 2nd-order for $k = 11$ (d) 2nd-order for $k = 5$ (see online version for colours)



(a)

(b)

(c)

(d)

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, $\hat{\varphi} = \text{Id}$	30	30	100	1,000	3,000	7,000	15,000	70,000	260,000
MIA$_\text{H}$ (hist.), $\hat{\varphi} = \text{Id}$	80	160	600	4,000	20,000	50,000	95,000	850,000	10^6+
MIA$_\text{K}$ (kernel), $\hat{\varphi} = \text{Id}$	70	140	500	3,000	15,000	35,000	60,000	500,000	10^6+
MIA$_\text{P}$ (param.), $\hat{\varphi} = \text{Id}$	60	100	300	2,000	5,000	15,000	20,000	150,000	500,000
CPA, $\hat{\varphi} = \text{HW}$	30	30	70	400	2,000	4,000	7,000	45,000	170,000
MIA$_\text{H}$ (hist.), $\hat{\varphi} = \text{HW}$	40	70	300	1,500	7,000	20,000	40,000	320,000	10^6+
MIA$_\text{K}$ (kernel), $\hat{\varphi} = \text{HW}$	30	60	190	1,500	5,500	15,000	25,000	190,000	900,000
MIA$_\text{P}$ (param.), $\hat{\varphi} = \text{HW}$	70	70	150	1,000	3,000	7,000	15,000	65,000	300,000

6 Experimental results

6.1 First-order attack simulations

To compare the efficiency of 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 1,000 attacks). Moreover, we included first-order CPA in our tests to determine whether and when an MIA is more efficient than a CPA.[3] 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$_\text{H}$, MIA$_\text{K}$ and MIA$_\text{P}$ respectively stand for the histogram, the kernel and the parametric MIA.

It can be checked in Table 1 that CPA is always better than MIA when $\hat{\varphi} = \text{HW}$. 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 $\mathbb{F}_2^4 = \{0,...,15\}$. Eventually, both results suggest that CPA is more suitable than 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$_\text{P}$ becomes much more efficient than MIA$_\text{H}$ and MIA$_\text{K}$ when the noise standard deviation increases.

6.2 Second-order attack simulations

In a CPA, the attacker computes Pearson's correlation coefficients which is a function of two univariate samples. Thus, when 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) (Prouff et al., 2009). This signal processing induces an information loss which strongly impacts the higher-order CPA efficiency when the noise increases. Because an higher-order MIA can operate on multivariate samples, it does not suffer from the aforementioned drawback. We could therefore expect MIA to become more efficient than CPA when it is performed against masking. To compare higher-order CPA and higher-order MIA, we simulated power consumption measurements such as in (14) with $d = 2$, with $\varphi_0 = \varphi_1 = \text{HW}$, with $\sigma_0 = \sigma_1 = \sigma$ 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). Table 2 reports the results that we obtained[4] for second-order CPA (2O-CPA) and for second-order MIA with histogram estimation method (2O-MIA$_\text{H}$) and with kernel estimation method (2O-MIA$_\text{K}$). We performed second-order CPA with Hamming weight prediction function and for two different *combining functions*: the absolute difference and the normalised product (Prouff et al., 2009). For MIA, we tried both Hamming weight and identity prediction functions. Moreover, for MIA with histogram estimation, we tried two rules for the choice of the bin-width: Scott's rule (see Section 5.1) and the rule proposed in Gierlichs et al. (2008).

Remark 6.1: We experimented that second-order MIA with parametric estimations using the expectation maximisation algorithm is inefficient. In fact, estimating a Gaussian mixture using the EM algorithm requires a great number of samples, especially when the number of components in the mixture is not small. In our context, the number of components equals the number of possible mask values[5], that is 16 when attacking a DES S-box. To lower the number of components, one could focus on a restricted number of bits (considering the remaining ones as an algorithmic noise). Such an approach has been followed by Lemke-Rust and Paar (2007) in the context of higher-order profiled attacks. Another approach could be to look for other estimation methods dedicated to Gaussian mixtures and, possibly, to adapt them for masked implementations. We let such investigations for future research.

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 ($\hat{\varphi} = \text{HW}$, abs. difference)	300	800	5,000	200,000	10^6+	10^6+
2O-CPA ($\hat{\varphi} = \text{HW}$, norm. product)	300	400	3,000	70,000	300,000	10^6+
2O-MIA$_{\text{H}}$ ($\hat{\varphi} = \text{Id}$, Scott's rule)	1,200	7,000	75,000	10^6+	10^6+	10^6+
2O-MIA$_{\text{H}}$ ($\hat{\varphi} = \text{Id}$, rule in Gierlichs et al. (2008)	1,800	7,000	40,000	1,000,000	10^6+	10^6+
2O-MIA$_{\text{K}}$ ($\hat{\varphi} = \text{Id}$)	600	2,500	25,000	600,000	10^6+	10^6+
2O-MIA$_{\text{H}}$ ($\hat{\varphi} = \text{HW}$, Scott's rule)	600	2,700	34,000	10^6+	10^6+	10^6+
2O-MIA$_{\text{H}}$ ($\hat{\varphi} = \text{HW}$, rule in Gierlichs et al. (2008)	350	1,300	9,000	350,000	10^6+	10^6+
2O-MIA$_{\text{K}}$ ($\hat{\varphi} = \text{HW}$)	300	1,300	9,000	n.a.	n.a.	n.a.

Table 2 shows that, contrary to what we could have expected, second-order CPA is always better than second-order MIA. As for the first-order case, we deduce that CPA is more suitable to attack masked implementations that leak the Hamming weight of the processed data with Gaussian noise. However, we also note that the efficiency of MIA is strongly impacted by the estimation methods and the related parameters (e.g., the choice of the bin-width for histograms). Determining the estimation method/parameters that optimise (or at least improve) the attack efficiency is hence a relevant open issue. Results reported in Table 2 also show that in the considered context, kernels perform better than histograms and that a Hamming weight prediction is better than an identity prediction. These observations are quite natural since, on the one hand, kernels are known to give tighter pdf estimations than histograms and, on the other hand, a Hamming weight prediction enables better discrimination of the wrong key guesses than an identity prediction in the presence of a Hamming weight leakage function.[6] Another observation is that, the bin-width selection rule proposed in Gierlichs et al. (2008) for histogram-based MIA leads to a more efficient attack than Scott's rule. More generally, we experimented that increasing the bin-width improve the attack efficiency until reaching a small number of bins. The analysis of the underlying reasons for this phenomenon and the study of the bin-width choice optimising the MIA efficiency are open issues that deserve more investigations.

6.3 *Practical attacks*

To experimentally validate our theoretical analysis and the simulations reported in Section 5, Section 6.1 and Section 6.2, we experimented MIA with real-life leakage traces measured for different kinds of implementations. We first performed univariate MIA attacks against hardware and software implementations of the AES S-box. Then, we applied second-order MIA attacks against a masked software implementation of the first DES S-box. In both contexts, we also performed a CPA attack to compare its efficiency with that of MIA.

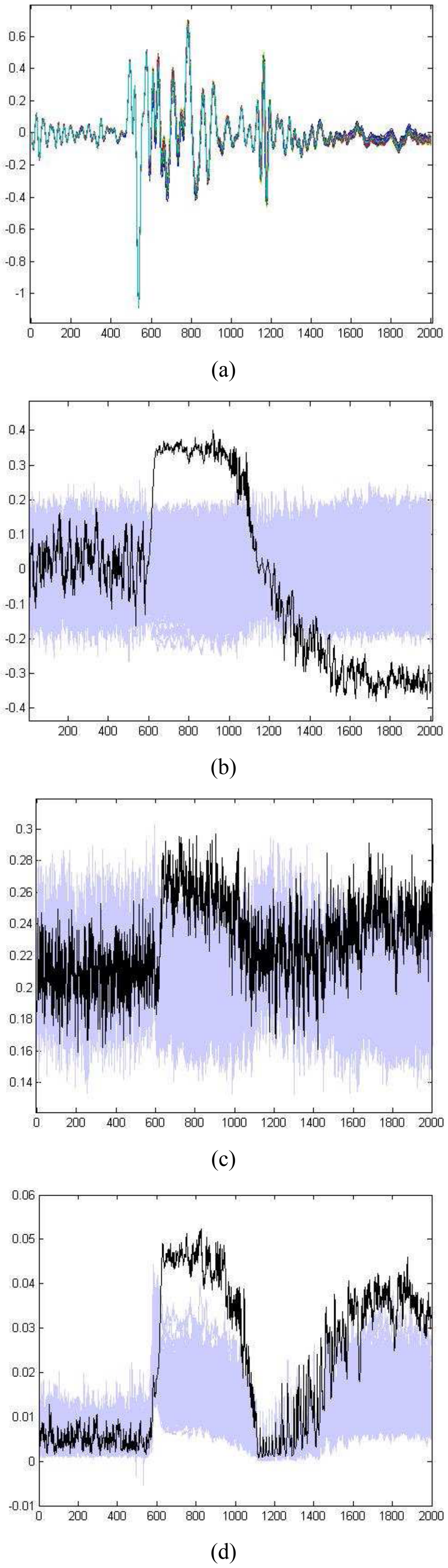
6.3.1 *First-order attacks*

We performed the attacks 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 Guilley et al., 2008) for details about the chip and the circuit layout). The corresponding power consumption measurements are plotted in Figure 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 smart card with 8-bit architecture. As it can be seen in Figure 6(a), the signal is much more noisy in this case.

For both set of traces, we performed CPA and MIA attacks with the histogram estimation method and the parametric estimation method (see Section 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 Figure 5(b) to Figure 5(d) and Figure 6(b) to Figure 6(d) over the time. For each attack the curve corresponding to the correct (resp. wrong) key hypothesis is drawn in black (resp. grey).

In both cases, the attacks succeed with a few number of traces. It can be noticed that MIA with a parametric estimation is more discriminating than MIA with the histogram estimation. This confirms the simulations performed in Section 6.1. However, even when the parametric estimation method is involved, CPA is always more discriminating than MIA. Those results suggest that the power consumption of the attacked devices has in fact a high linear dependency with the Hamming weight of the manipulated data. This implies in particular that the Hamming weight model is sound in this context and that looking for non-linear dependencies is not useful.

Figure 5 Practical attacks on a hardware AES implementation, (a) power consumption traces (b) CPA with 256 traces (c) histogram-based MIA with 1,024 traces (d) parametric MIA with 1,024 traces (see online version for colours)



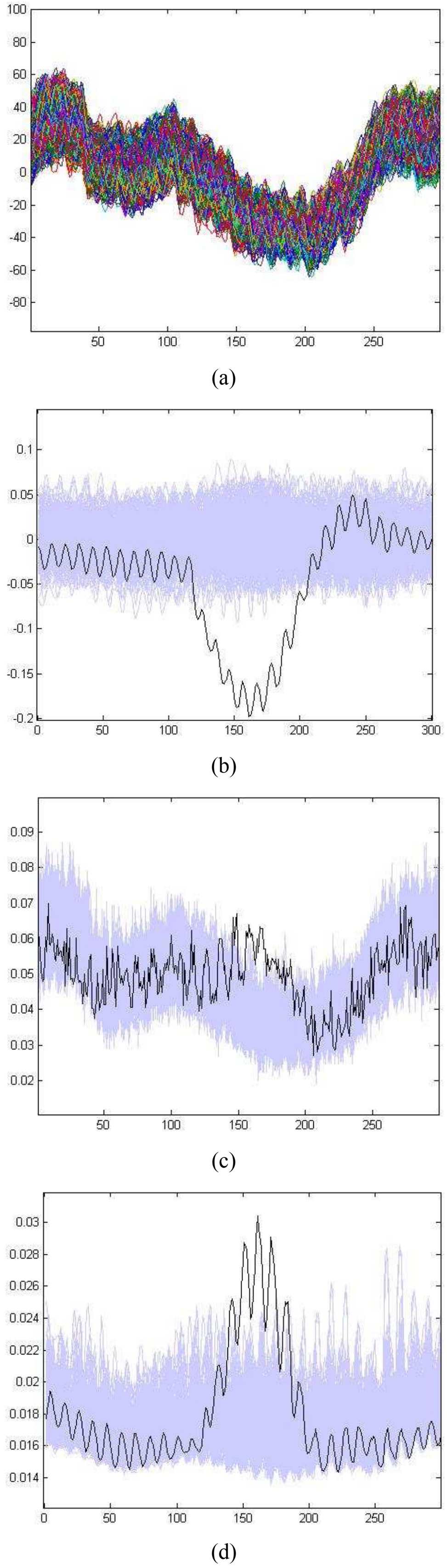
(a)

(b)

(c)

(d)

Figure 6 Practical attacks on a software AES implementation, (a) power consumption traces (b) CPA with 2,000 traces (c) histogram-based MIA with 2,000 traces (d) parametric MIA with 2,000 traces (see online version for colours)



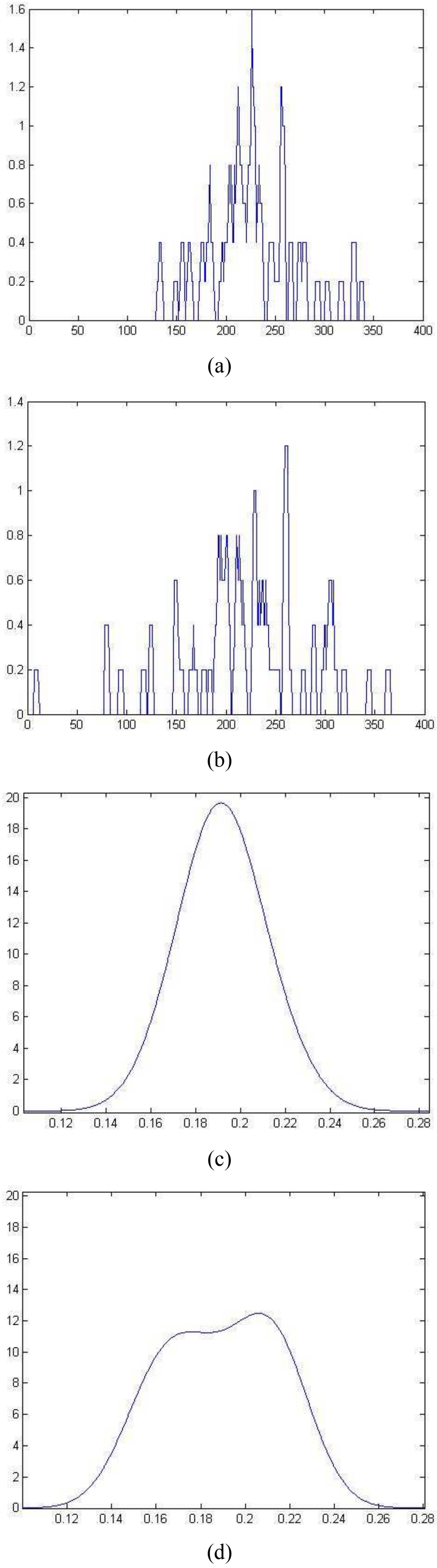
(a)

(b)

(c)

(d)

Figure 7 Pdf estimations on power measurements, (a) histogram estimation $(k = k^\star)$ (b) histogram estimation $(k \neq k^\star)$ (c) parametric estimation $(k = k^\star)$ (d) parametric estimation $(k \neq k^\star)$ (see online version for colours)



(a)

(b)

(c)

(d)

To corroborate the soundness of the analysis given in Section 4 and Section5, we plotted in Figure 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 Gaussian mixture pdfs (a Gaussian pdf when $k^\star$ is correctly guessed and a mixture of two pdfs when it is not).

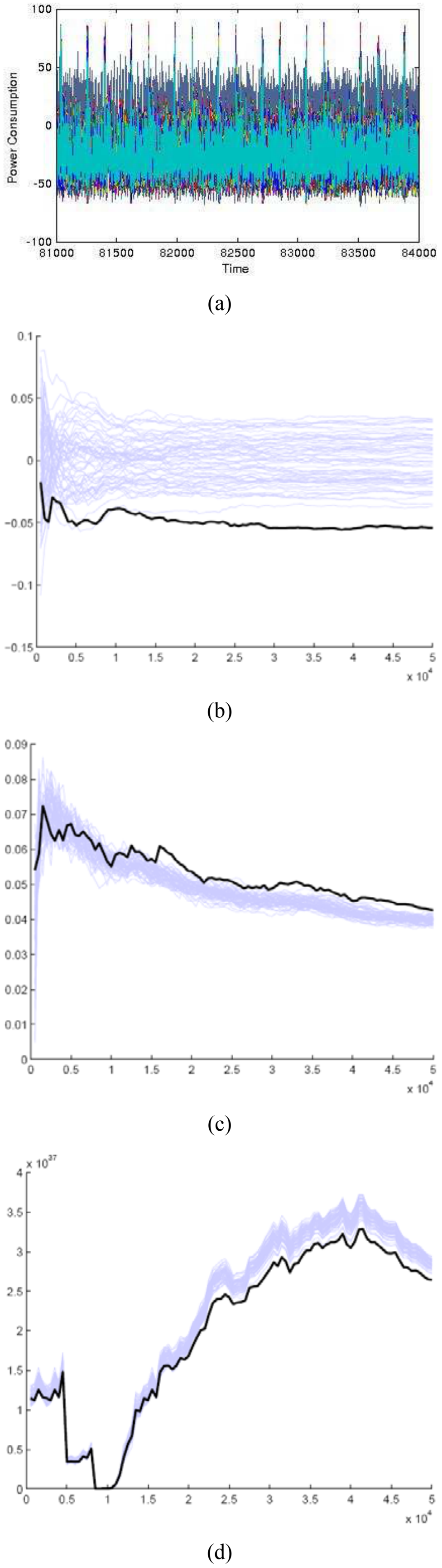
6.3.2 *Second-order attacks*

We performed second-order MIA attacks against a DES S-box implementation that uses a lookup-table and is protected by first-order masking. Namely, the targeted variable $f_{k^\star}(X)$ corresponds to the DES S-box output and the leakage measurement consists in two points L_0 and L_1 satisfying (14) for $d = 2$. The power consumption traces have been measured for a software implementation running on a smart card with 8-bit architecture. They are plotted in Figure 8(a). The traces are composed of 3,000 points and we identified that the masked value $f_{k^\star}(X) \oplus M$ is manipulated at time $t_0 = 81{,}789$ whereas the mask M is manipulated at time $t_1 = 83{,}238$. We also performed a second-order CPA attack involving the normalised product combining with $\hat{\varphi} = \mathrm{HW}$. For the MIA attacks, we choose to define the prediction function $\hat{\varphi}$ either as the identity function or as the Hamming weight function. For the histogram-based MIA, we used the Scott's rule for the bin-width. For the kernel-based MIA, we applied the normal scale rule recalled in (20) to select the kernels bandwidth.

Our attacks results for $\hat{\varphi}$ being the identity function are plotted in Figure 8(c) to Figure 8(d). For each key-candidate, the mutual information/correlation values are plotted over the number of leakage measurements. The curve corresponding to the correct (resp. wrong) key hypothesis is drawn in black (resp. grey). In Figure 9, we plotted for each attack the rank of the good key hypothesis according to the number of traces exploited by the attack. The dot-line corresponds to the second-order CPA attack. Black curves refer to 2O-MIA$_{\mathrm{K}}$ attacks whereas grey curves refer to 2O-MIA$_{\mathrm{H}}$ attacks. In both cases, plain-lines correspond to attacks with $\hat{\varphi} = \mathrm{Id}$ and dashed-lines correspond to $\hat{\varphi} = \mathrm{HW}$.

As we can see from Figure 9, the obtained results validate our simulations. In particular, we see that second-order CPA is clearly more efficient than second-order MIA and that kernel-based MIA is better than histogram-based MIA. We further observe that compared to our simulations where the Hamming weight prediction leads to better efficiency than the identity prediction, both predictions lead to similar results for our practical attacks. This suggests that the power consumption of the attacked device does not fully depend on the Hamming weight of the processed data but rather on some leakage function between Hamming weight and identity (e.g., each bit of the data has a different weight in the power consumption).

Figure 8 Practical second-order attacks on a software DES implementation, (a) power consumption traces (b) 2O-CPA, $\hat{\varphi} = \text{HW}$ (c) 2O-MIAH, $\hat{\varphi} = \text{Id}$ (d) 2O-MIAK, $\hat{\varphi} = \text{Id}$ (see online version for colours)



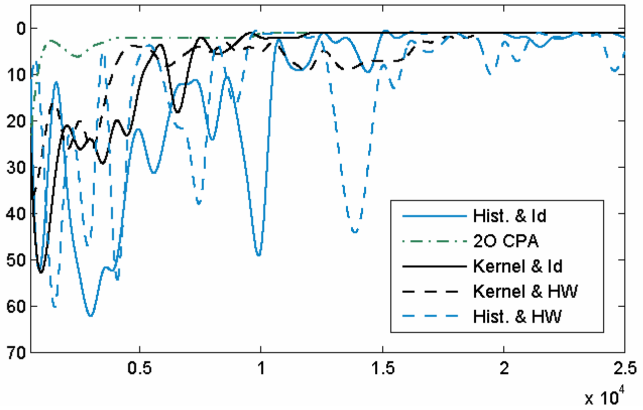
(a)

(b)

(c)

(d)

Figure 9 Rank of the good key hypothesis according to the attack type (see online version for colours)



7 Concluding remarks

This paper extends the works published in Gierlichs et al. (2008) and Aumonier (2007) to expose the theoretical foundations behind MIA and to generalise it to higher-orders. In the first-order context, we have shown that MIA is less efficient than CPA when the deterministic part of the leakage is a linear function of the prediction made by the attacker. This implies that CPA should be preferred to MIA when the targeted device leaks a linear function of the Hamming weight of the manipulated data. This paper also argues that the efficiency of MIA attacks (of first or higher-order) greatly depends on the way how some (joint) pdfs are estimated. In particular, we introduced a parametric estimation method that renders the first-order MIA efficiency close to that of CPA when noise increases.

It is interesting to notice that once good estimations of the joint pdfs are got, other statistical tools than the mutual information can be used for key-guess discrimination. This is to our mind an avenue for further research on this topic. A first attempt towards this direction has been reported by Veyrat-Charvillon and Standaert (2009) who studied the *Kullback-Leibler divergence*, the *Jensen-Shannon divergence* and the *Kolmogorov-Smirnov divergence*. The first two divergence measures rely on Shannon entropy (as the mutual information), whereas the third divergence measure relies on the *Kolmogorov complexity*. The attack experiments reported in Veyrat-Charvillon and Standaert (2009) shows that this approach does not enable to break the first DES S-box of the DPA contest (Guilley et al., 2008) in a more efficient way than MIA does. However, since the measurements proposed in the DPA-contest are almost noise-free, more investigations are still required to have a clear idea about the efficiency of the approach in the general case.

When masking is used to protect the target implementation, an extension of MIA to higher-orders has been proposed in this paper. In the same line of research, Gierlichs et al. proposed an alternative approach using the *multivariate mutual information* (Gierlichs et al., 2010). For further research one could also involve the *absolute* *mutual information* from Kolmogorov complexity. Another perspective is the investigation of Gaussian mixture estimation methods dedicated to masked implementations leakage. A first step in this direction is the work by Lemke-Rust and Paar (2007) who applied the EM algorithm in such a context. Otherwise, the question of optimal choice for the bin-width in the histogram estimation method is also a relevant open issue. We think that these questions definitely require more investigations.

Acknowledgements

We thank Sylvain Guilley and Christophe Giraud for providing us with the power consumption traces involved in our attack experiments. We also thank Benedikt Gierlichs and the anonymous referees of *IJACT* for their fruitful comments.

References

Aumonier, S. (2007) 'Generalized correlation power analysis', *Published in the Proceedings of the Ecrypt Workshop Tools for Cryptanalysis 2007.*.

Beirlant, J., Dudewicz, E.J., Györfi, L. and Meulen, E.C. (1997) 'Nonparametric entropy estimation: an overview', *International Journal of the Mathematical Statistics Sciences*, Vol. 6, pp.17–39.

Birgé, L. and Rozenholc, Y. (2006) 'How many bins should be put in a regular histogram', *ESAIM: P&S*, February, Vol. 10, pp.24–45.

Bishop, C.M. (2007) *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1st ed., Springer.

Brier, E., Clavier, C. and Olivier, F. (2004) 'Correlation power analysis with a leakage model', in Joye, M. and Quisquater, J.J. (Eds.): *Cryptographic Hardware and Embedded Systems – CHES 2004*, Lecture Notes in Computer Science, Springer, Vol. 3156, pp.16–29.

Carreira-Perpinan, M. (2000) 'Mode-finding for mixtures of Gaussian distributions Carreira-Perpinan', *Pattern Analysis and Machine Intelligence, IEEE Transactions*, November, Vol. 22, No. 11, pp.1318–1323.

Chari, S., Jutla, C., Rao, J. and Rohatgi, P. (1999) *Towards Sound Approaches to Counteract Power-Analysis Attacks*, pp.398–412.

Chari, S., Rao, J. and Rohatgi, P. (2002) 'Template attacks', in Kaliski, B., Jr., Koç, Ç. and Paar, C. (Eds.): *Cryptographic Hardware and Embedded Systems – CHES 2002*, Lecture Notes in Computer Science, Springer, Vol. 2523, pp.13–29.

Gierlichs, B., Batina, L., Preneel, B. and Verbauwhede, I. (2010) 'Revisiting higher-order DPA attacks', in Pieprzyk, J. (Ed.): *CT-RSA*, Lecture Notes in Computer Science, Springer, Vol. 5985, pp.221–234.

Gierlichs, B., Batina, L., Tuyls, P. and Preneel, B. (2008) 'Mutual information analysis', in Oswald, E. and Rohatgi, P. (Eds.): *CHES*, Lecture Notes in Computer Science, Springer, Vol. 5154, pp.426–442.

Gray, A.G. and Moore, A.W. (2003) 'Nonparametric density estimation: toward computational tractability', in *Proceedings of the Third SIAM International Conference on Data Mining, SIAM*.

Guilley, S., Sauvage, L., Hoogvorst, P., Pacalet, R., Bertoni, G.M. and Chaudhuri, S. (2008) 'Security evaluation of wddl and seclib countermeasures against power attacks', *IEEE Transactions on Computers*, November, Vol. 57, No. 11, pp.1482–1497.

Kocher, P., Jaffe, J. and Jun, B. (1999) 'Differential power analysis', in Wiener, M. (Ed.): *Advances in Cryptology – CRYPTO '99*, Lecture Notes in Computer Science, Springer, Vol. 1666, pp.388–397.

Lemke-Rust, K. and Paar, C. (2007) 'Gaussian mixture models for higher-order side channel analysis', in Paillier, P. and Verbauwhede, I. (Eds.): *Cryptographic Hardware and Embedded Systems – CHES 2007*, Lecture Notes in Computer Science, Springer, Vol. 4727, pp.14–27.

Moradi, A., Mousavi, N., Paar, C. and Salmasizdeh, M. (2009) 'A comparative study of mutual information analysis under a Gaussian assumption', in Youm, H.Y. and Yung, M. (Eds.): *WISA*, Lecture Notes in Computer Science, Springer, Vol. 5932, pp.193–205.

Prouff, E. (2005) 'DPA attacks and S-boxes', in Handschuh, H. and Gilbert, H. (Eds.): *Fast Software Encryption – FSE 2005*, Lecture Notes in Computer Science, Springer, Vol. 3557, pp.424–442.

Prouff, E. and Rivain, M. (2009) 'Theoretical and practical aspects of mutual information based side channel analysis', in Abdalla, M., Pointcheval, D., Fouque, P.A. and Vergnaud, D. (Eds.): *Applied Cryptography and Network Security – ANCS 2009*, Lecture Notes in Computer Science, Springer, Vol. 5536, pp.499–518.

Prouff, E., Rivain, M. and Bevan, R. (2009) 'Statistical analysis of second order differential power analysis', *IEEE Trans. Computers*, Vol. 58, No. 6, pp.799–811.

Scott, D.W. (1992) *Multivariate Density Estimation: Theory, Practice, and Visualization (Wiley Series in Probability and Statistics)*, Wiley-Interscience, September.

Silverman, B. (1986) *Density Estimation for Statistics and Data Analysis*, Chapman and Hall.

Standaert, F.X., Malkin, T. and Yung, M. (2009) 'A unified framework for the analysis of side-channel key recovery attacks', in Joux, A. (Ed.): *EUROCRYPT*, Lecture Notes in Computer Science, Springer, Vol. 5479, pp.443–461.

Turlach, B.A. (1993) 'Bandwidth selection in kernel density estimation: a review', in *CORE and Institut de Statistique*, pp.23–493.

Veyrat-Charvillon, N. and Standaert, F.X. (2009) 'Mutual information analysis: how, when and why?', in Clavier, C. and Gaj, K. (Eds.): *CHES*, Lecture Notes in Computer Science, Springer, Vol. 5747, pp.429–443.

Wand, M.P. (1997) 'Data-based choice of histogram bin width', *The American Statistician*, Vol. 51, pp.59–64.

Wasserman, L. (2005) *All of Statistics: A Concise Course in Statistical Inference*, Springer Texts in Statistics.

Notes

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

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

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

4 The results given in the conference version of the paper of Prouff and Rivain (2009) for second-order MIA simulations are erroneous. Table 2 provides the corrected results.

5 It may be less in a particular leakage model (e.g., hamming weight model) but the attacker does not *a priori* have such information.