Mutual Information Analysis: How, When and Why?

Nicolas Veyrat-Charvillon*, François-Xavier Standaert**

UCL Crypto Group, Université catholique de Louvain, B-1348 Louvain-la-Neuve.
e-mails: `nicolas.veyrat,fstandae@uclouvain.be`

Abstract. The Mutual Information Analysis (MIA) is a generic side-channel distinguisher that has been introduced at CHES 2008. This paper brings three contributions with respect to its applicability to practice. First, we emphasize that the MIA principle can be seen as a toolbox in which different (more or less effective) statistical methods can be plugged in. Doing this, we introduce interesting alternatives to the original proposal. Second, we discuss the contexts in which the MIA can lead to successful key recoveries with lower data complexity than classical attacks such as, *e.g.* using Pearson's correlation coefficient. We show that such contexts exist in practically meaningful situations and analyze them statistically. Finally, we study the connections and differences between the MIA and a framework for the analysis of side-channel key recovery published at Eurocrypt 2009. We show that the MIA can be used to compare two leaking devices only if the discrete models used by an adversary to mount an attack perfectly correspond to the physical leakages.

1 Introduction

The most classical solutions used in non profiled side-channel attacks are Kocher's original DPA [14] and correlation attacks using Pearson's correlation coefficient, introduced by Brier *et al.* [5]. In 2008, another interesting side-channel distinguisher has been proposed, denoted as Mutual Information Analysis (MIA) [12]. MIA aims at genericity in the sense that it is expected to lead to successful key recoveries with as little assumptions as possible about the leaking devices it targets. In this paper, we confirm and extend the ideas of Gierlichs *et al.* and tackle three important questions with respect to this new distinguisher.

1. How to use MIA? In general, MIA can be viewed as the combination of two subproblems. In a first stage of the attack, an adversary has to *estimate* the leakage *probability density functions* for different key-dependent models. In a second stage of the attack, this adversary has to *test the dependence* of these models with actual measurements. In the original description of [12], the MIA is using histograms for the first stage and a Kullback-Leibler divergence for the second stage. In this paper, we argue that in fact, the MIA can be seen as a

* Work supported in part by the Walloon Region research project SCEPTIC.

** Associate researcher of the Belgian Fund for Scientific Research (FNRS - F.R.S.).

toolbox in which different probability density estimation techniques and notions of divergence can be used. We show that these different solutions (some of them being introduced in [3, 19]) yield different results for the attack effectiveness. We also introduce an alternative test that is at least as generic as the original MIA but does not require an explicit estimation of the leakage probability densities.

2. When to use MIA? In a second part of this paper, we analyze the contexts in which MIA can be necessary (*i.e.* when other side-channel attacks would not succeed). In [19], it is argued that MIA is particularly convenient in higher-order side-channel attacks because of its simple extension to multi-dimensional scenarios. In this paper, we show that MIA can also be useful in univariate side-channel attacks, if the models used by an adversary to mount an attack are not sufficiently precise. Hence, we complement the original experiment of [12] against a dual-rail pre-charged implementation. In order to further validate this intuition, we analyze an arbitrary degradation of the leakage models and show that after a certain threshold, MIA leads to a more effective key recovery than the corresponding correlation attack using Pearson's coefficient. We also discuss the effect of incorrect models theoretically and intuitively.

3. Why to use MIA? Eventually, in a third part of the paper, we investigate the relations between the MIA and the information theoretic *vs.* security model of [22]. We exhibit that although having similar foundations, MIA and this model have significantly different goals and are not equivalent in general. We also show that in certain idealized contexts (namely, when adversaries can exploit leakage predictions that perfectly correspond to the actual measurements), the MIA can be used as a metric to compare different cryptographic devices.

2 Background

2.1 Information theoretic definitions

Entropy. The entropy [7] of a random variable X on a discrete space $\mathcal{X}$ is a measure of its uncertainty during an experiment. It is defined as:

$$\mathrm{H}\left[X\right] = -\sum_{x \in \mathcal{X}} \Pr\left[X = x\right] \log_2(\Pr\left[X = x\right]).$$

The joint entropy of a pair of random variables X, Y expresses the uncertainty one has about the combination of these variables:

$$\mathrm{H}\left[X, Y\right] = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \Pr\left[X = x, Y = y\right] \log_2(\Pr\left[X = x, Y = y\right]).$$

The joint entropy is always greater than that of either subsystem, with equality only if Y is a deterministic function of X. The joint entropy is also sub-additive. Equality occurs only in the case where the two variables are independent.

$$\mathrm{H}\left[X\right] \leq \mathrm{H}\left[X, Y\right] \leq \mathrm{H}\left[X\right] + \mathrm{H}\left[Y\right].$$

Finally, the conditional entropy of a random variable X given another variable Y expresses the uncertainty on X which remains once Y is known.

$$\mathrm{H}\left[X|Y\right] = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} \Pr\left[X = x, Y = y\right] \log_2(\Pr\left[X = x|Y = y\right]).$$

The conditional entropy is always greater than zero, with equality only in the case where X is a deterministic function of Y. It is also less than the entropy of X. Equality only occurs if the two variables are independent.

$$0 \leq \mathrm{H}\left[X|Y\right] \leq \mathrm{H}\left[X\right].$$

All these measures can be straightforwardly extended to continuous spaces by differentiation. For example, the differential entropy is defined as:

$$\mathrm{H}\left[X\right] = -\int_{x\in\mathcal{X}} \Pr\left[X = x\right] \log_2(\Pr\left[X = x\right]).$$

The differential entropy can be negative, contrary to the discrete entropy.

Mutual information The mutual information is a general measure of the dependence between two random variables. On a discrete domain, the mutual information of two random variables X and Y is defined as:

$$\mathrm{I}\left(X;Y\right) = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \Pr\left[X = x, Y = y\right] \log_2\left(\frac{\Pr\left[X = x, Y = y\right]}{\Pr\left[X = x\right]\cdot\Pr\left[Y = y\right]}\right).$$

It is directly related to Shannon's entropy, and can be expressed using entropies:

$$\begin{aligned}\mathrm{I}\left(X;Y\right) &= \mathrm{H}\left[X\right] - \mathrm{H}\left[X|Y\right]\\ &= \mathrm{H}\left[X\right] + \mathrm{H}\left[Y\right] - \mathrm{H}\left[X,Y\right]\\ &= \mathrm{H}\left[X,Y\right] - \mathrm{H}\left[X|Y\right] - \mathrm{H}\left[Y|X\right]\end{aligned}$$

It can also be straightforwardly extended to the continuous case:

$$\mathrm{I}(X;Y) = \int_{x\in\mathcal{X}}\int_{y\in\mathcal{Y}} \Pr\left[X = x, Y = y\right] \log_2\left(\frac{\Pr\left[X = x, Y = y\right]}{\Pr\left[X = x\right]\cdot\Pr\left[Y = y\right]}\right).$$

2.2 Pearson's correlation coefficient

This coefficient is a simpler measure of dependence between two random variables X and Y. Computing it does not require the knowledge of the probability density functions of X and Y but it only measures the linear dependence between these variables (whereas mutual information is able to detect any linear or non-linear dependence). It is defined as follows (with $\overline{X}$ the mean value of X):

$$\rho\left(X,Y\right) = \frac{\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\left(x-\overline{X}\right)\cdot\left(y-\overline{Y}\right)}{\sqrt{\sum_{x\in\mathcal{X}}\left(x-\overline{X}\right)^2\cdot\sum_{y\in\mathcal{Y}}\left(y-\overline{Y}\right)^2}}.$$

2.3 Side-channel analysis

In a side-channel attack, an adversary tries to recover secret information from a leaking implementation, *e.g.* a software program or an IC computing a cryptographic algorithm. The core idea is to compare key-dependent models of the leakages with actual measurements. Typically, the adversary first defines the subkeys that he aims to recover. For example, in a block cipher implementation, those subkeys could be one byte of the master key. Then, for each subkey candidate, he builds models that correspond to the leakage generated by the encryption of different plaintexts. Eventually, he evaluates which model (*i.e.* which subkey) gives rise to the best prediction of the actual leakages, measured for the same set of plaintexts. As a matter of fact and assuming that the models can be represented by a random variable X and the leakages can be represented by a random variable Y, the side-channel analysis can simply be seen as the problem of detecting a dependence between those two variables. Pearson's coefficient and the mutual information can be used for this purpose.

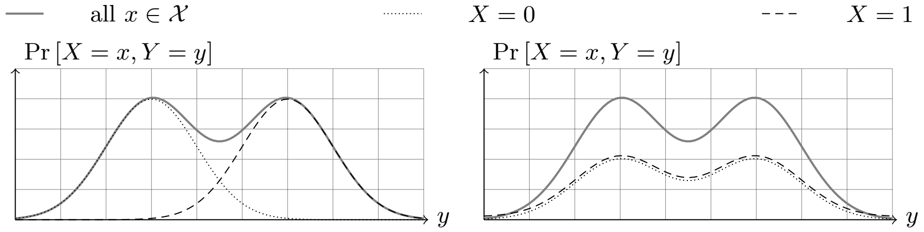
In the following, we consider side-channel attacks restricted by two important assumptions. First, we investigate *univariate attacks*, *i.e.* attacks in which one compares the leakage models X with a single sample in the leakage traces. It means that the variable Y has only one dimension. Second, we consider *discrete leakage models*, *i.e.* we assume that the variable X is discrete (by contrast, the actual leakage variable Y can be continuous). We note that univariate attacks are typical scenarios in standard DPA attacks such as [14] and discrete leakage models are also a very natural assumption as long as the side-channel attacks cannot be enhanced with profiling and characterization [6]. Hence, these two assumptions can be seen as reasonable starting points for the analysis of MIA.

3 How to use MIA: the information theoretic toolbox

Following the previous informal description, let us denote the subkey candidates in a side-channel attack as k_j and the models corresponding to those subkeys as X_j. The distinguisher used in a mutual information analysis is defined as:

$$d_j = \hat{\mathrm{I}}(X_j; Y).$$

For simplicity, we will omit the j subscript in the following of the paper. The idea behind this procedure is that a meaningful partition of Y where each subset corresponds to a particular model value will relate to a side-channel sample distribution $\hat{\Pr}[Y|X = x]$ distinguishable from the global distribution of $\hat{\Pr}[Y]$. The estimated mutual information will then be larger than zero. By contrast, if the key guess is incorrect, the false predictions will form a partition corresponding to a random sampling of Y and therefore simply give scaled images of the total side-channel probability density function (pdf for short). Hence, the estimated mutual information will be equal (or close) to zero in this case.

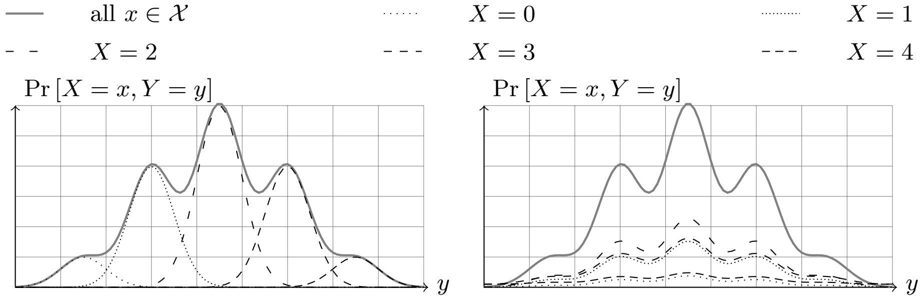


$X =$	0	1
	y_0	y_1
	y_2	y_3
	y_4	y_5
	y_6	y_7
	...	...

$X =$	0	1
	y_0	y_1
	y_3	y_2
	y_5	y_4
	y_6	y_7
	...	...

Fig. 1. Probability densities and associated leakage partitions for correct (left) and wrong (right) subkey hypotheses in the case of a single bit DPA attack.

Example. Let us imagine a target implementation in which the adversary receives leakages of the form $y = \mathrm{H_W}(\mathsf{S}(p \oplus k)) + n$ where $\mathrm{H_W}$ is the Hamming weight function, S the 4-bit S-box of the block cipher Serpent, p a known plaintext, k the target subkey of the attack and n is a Gaussian noise. Let us also assume two different attacks: in the first one, the model X corresponds to a single bit of $\mathsf{S}(p \oplus k)$; in the second one, the model X corresponds to $\mathrm{H_W}(\mathsf{S}(p \oplus k))$. Figures 1 and 2 illustrate what happens asymptotically to the correct and a wrong subkey hypotheses in the case these two attacks. They clearly show the higher dependence for the correct subkey (*i.e.* the left figures) that is expected by [12].



$X =$	0	1	2	3	4
	y_0	y_1	y_3	y_7	y_{15}
	...	y_2	y_5	y_{11}	...
		y_4	y_6	y_{13}	
		y_8	y_9	y_{14}	
		...	y_{10}	...	
			y_{12}		
			...		

$X =$	0	1	2	3	4
	y_7	y_0	y_1	y_5	y_6
	...	y_2	y_3	y_8	...
		y_{13}	y_4	y_{10}	
		y_{15}	y_9	y_{12}	
		...	y_{11}	...	
			y_{14}		
			...		

Fig. 2. Probability densities and associated leakage partitions for correct (left) and wrong (right) subkey hypotheses in the case of a 4-bit DPA attack.

In theory, the MI distinguisher tests a *null hypothesis* stating that the predicted leakage values and the side-channel samples are independent if the subkey hypothesis is false. When this hypothesis is not verified, the adversary assumes that he found the correct subkey. However, in practice there may exist certain dependencies between a wrong subkey candidate and the actual leakages (*e.g.* ghost peaks as in [5]). Hence, the adversary generally selects the subkey that leads to the highest value of the distinguisher. This description underlines that a MIA is essentially composed of the two problems listed in introduction:

1. An estimation of some probability density functions, namely those of the global samples and of the samples corresponding to each modeled leakage.
2. The test of a null hypothesis stating that the predicted leakages and their actual side-channel values are independent.

As a matter of fact, different solutions can be considered for this purpose. Therefore, in the remainder of this section, we first review some possible techniques to estimate the probability density functions used in a side-channel attack. Then we present various probability-distance measures that can replace the usual relative entropy in mutual information analysis. Eventually, we discuss the possibility to compare two pdf without explicitly estimating them and briefly mention alternative attack techniques inspired from "all-or-nothing" multiple-bit DPA.

3.1 Probability density function estimation

The problem of modeling a probability density function from random samples of this distribution is a well studied problem in statistics, referred to as density estimation. A number of solutions exist, ranging from simple histograms to kernel density estimation, data clustering and vector quantization. The authors of [12] used histograms for density estimation as a proof of concept for MIA. But in certain contexts, an attack can be greatly improved by using more advanced techniques. In the following, we summarize a few density estimation tools that have been initially suggested in [3] as relevant to side-channel attacks and then applied to MIA in [19]. They are detailed in Appendix A.

Non-parametric methods. One interesting feature of the MIA is that it does not rely on particular assumptions on the leakages. Hence, it is natural to consider non-parametric estimation techniques first since, *e.g.* assuming Gaussian leakages would again reduce the genericity of the distinguisher. In practice, two techniques can generally be used for this purpose:

- Histograms perform a partition of the samples by grouping them into bins. Each bin contains the samples of which the value falls into a certain range. The respective ranges of the bins have equal width and form a partition of the range between the extreme values of the samples. Using this method, one approximates a probability by dividing the number of samples that fall within a bin by the total number of samples (see Appendix A.1).

- Kernel density estimation is a generalization of histograms. Instead of bundling samples together in bins, it adds (for each observed sample) a small kernel centered on the value of the leakage to the estimated pdf. The resulting estimation is a sum of small "bumps" that is much smoother than the corresponding histogram. It usually provides faster convergence towards the true distribution. Note that although this solution requires to select a Kernel and a bandwidth (details are given in Appendix A.2), it does not assume anything more about the estimated pdf than histograms.

Parametric methods. Contrary to the previous techniques, parametric methods for density estimation require certain assumptions about the leakages. They consequently trade some of the genericity of the MIA for a hopefully better effectiveness, *i.e.* they are an intermediate solution between attacks using the correlation coefficient and the original MIA of [12]. In this context, a particularly interesting tool is the finite mixture estimation. A mixture density is a probability density function that consists in a convex combination of probability density functions. Given a set of densities $p_1(x), \ldots, p_n(x)$, and positive weights $w_1, \ldots, w_n$ verifying $\sum w_i = 1$, the finite mixture is defined as:

$$\hat{\Pr}[x] = \sum_{i=0}^{n-1} w_i \; p_i(x).$$

A typical choice is to assume a mixture of Gaussian densities (see, *e.g.* [15]), which leads to an efficient parametric estimation of the pdf (see Appendix A.3).

3.2 Probability-distance measures

Once the probability densities have been estimated, one has to test whether the predicted leakages are correlated with the actual measurements. This dependence is tested using a probability-distance measure which allows deciding which subkey is the most likely to be the correct one. As in the previous section, different solutions can be used, that we detail and connect to the original MIA.

Kullback-Leibler divergence. The Kullback-Leibler divergence, or relative entropy [7], is a measure of the difference between two probability density functions P and Q. It is not a distance, as it is non-commutative and does not satisfy the triangle inequality. The KL divergence of Q from P, where P and Q are two probability functions of a discrete random variable X, is defined as:

$$D_{\mathsf{KL}}\left(\mathsf{P}\|\mathsf{Q}\right) = \sum_{x \in \mathcal{X}} \Pr\left[X = x, X \sim \mathsf{P}\right] \log \frac{\Pr\left[X = x, X \sim \mathsf{P}\right]}{\Pr\left[X = x, X \sim \mathsf{Q}\right]},$$

where $\Pr\left[X = x, X \sim \mathsf{P}\right]$ denotes the probability that the random variable X equals x when it follows the density function P. The mutual information can be defined in terms of Kullback-Leibler divergence, as being the divergence between the joint distribution $\Pr[X=x, Y=y]$ and the product distribution $\Pr[X=x] \cdot \Pr[Y=y]$, or as the expected divergence between the conditional distribution $\Pr[Y=y|X=x]$ and $\Pr[Y=y]$. In other words:

$$\begin{aligned} \mathrm{I}(X;Y) &= D_{\mathsf{KL}}\left(\Pr[X=x, Y=y] \,\|\, \Pr[X=x] \cdot \Pr[Y=y]\right) \\ &= E_{x\in\mathcal{X}}\left(D_{\mathsf{KL}}\left(\Pr[Y=y|X=x] \,\|\, \Pr[Y=y]\right)\right) \end{aligned}$$

Hence, it can be seen as the expected value of the divergence between the leakage distributions taken conditionally to the models and the marginal distribution.

F-divergences. The f-divergence [9] is a function of two probability distributions P and Q that is used to measure the difference between them. It was introduced independently by Csiszàr [8] and Ali and Silvey [1] and is defined as:

$$\mathrm{I}_f(\mathsf{P},\mathsf{Q}) = \sum_{x\in\mathcal{X}} \Pr[X=x, X\sim\mathsf{Q}] \times f\left(\frac{\Pr[X=x, X\sim\mathsf{P}]}{\Pr[X=x, X\sim\mathsf{Q}]}\right),$$

where f is a parameter function. Some classical examples include:

- Kullback-Leibler divergence: $f(t) = t\log t$
- Inverse Kullback-Leibler: $f(t) = -\log t$
- Pearson χ^2–divergence: $f(t) = (t-1)^2$
- Hellinger distance: $f(t) = 1-\sqrt{t}$
- Total variation: $f(t) = |t-1|$

As detailed in [12], the qualitative motivation for using the mutual information as a metric of dependence is sound. But one can wonder about its effectiveness. That is, all the previous f functions ensure an asymptotically successful attack. But are there significant differences in the convergence of the corresponding distinguishers? We note that the previous list is not exhaustive. For example, one could consider the Jensen-Shannon divergence that is a popular method based on the Kullback-Leibler divergence, with the useful difference that it is always a finite value: $D_{\mathsf{JS}}(P\|Q) = \frac{1}{2}\left(D_{\mathsf{KL}}(P\|M) + D_{\mathsf{KL}}(Q\|M)\right)$, where $M = \frac{1}{2}(P+Q)$. Similarly, the earth mover's or Mallow distances [4, 17] could also be used.

3.3 Distinguishing without explicit pdf estimation

Interestingly, an explicit pdf estimation is not always necessary and there also exist statistical tools to compare two pdfs directly from their samples. The Kolmogorov-Smirnov test is typical of such non parametric tools. For different samples x_i and a threshold x_t, it first defines an empirical cumulative function:

$$F(x_t) = \frac{1}{n}\sum_{i=1}^{n} \chi_{x_i\le x_t}, \text{ where } \chi_{x_i\le x_t} = \begin{cases} 1 \text{ if } x_i \le x_t \\ 0 \text{ otherwise.} \end{cases}$$

Then, the Kolmogorov-Smirnov distance is defined by:

$$D_{\mathsf{KS}}(P\|Q) = \sup_{x_t} |F_P(x_t) - F_Q(x_t)|.$$

This distance can then be used to test a null hypothesis. Since it is based on a supremum rather than a sum as the previous distances, it is better integrated to the following (MIA-inspired) distinguisher:

$$E_{x\in\mathcal{X}}\left(D_{\mathsf{KS}}\left(\Pr\left[Y=y|X=x\right]\|\Pr\left[Y=y\right]\right)\right)$$

This is further improved by normalizing each KS distance with the number of samples used in its computation, taking into account the convergence:

$$E_{x\in\mathcal{X}}\left(\frac{1}{|Y|X=x|}D_{\mathsf{KS}}\left(\Pr\left[Y=y|X=x\right]\|\Pr\left[Y=y\right]\right)\right),$$

where $|Y|X=x|$ is the number of leakages samples with modeled value x. Finally, an even more efficient alternative to the KS test is the two sample Cramér-von-Mises test [2], which is also based on the empirical cumulative function.

$$D_{\mathsf{CvM}}\left(P\|Q\right)=\int_{-\infty}^{+\infty}\left(F_P(x_t)-F_Q(x_t)\right)^2dx_t.$$

3.4 All-or-nothing comparisons

Eventually, we mention that the MIA is defined as the expected value of a divergence between the leakage distributions conditionally to the model values and the marginal leakage distribution, *i.e.* $E_{x\in\mathcal{X}}\left(D_{\mathsf{KL}}\left(\Pr\left[Y=y|X=x\right]\|\Pr\left[Y=y\right]\right)\right)$. But divergences between the conditional distributions could be considered as well, as in "all-or-nothing" DPA attacks (see, *e.g.* [3] for an example).

3.5 How much does it matter? Experimental results

The previous sections illustrate that MIA is in fact a generic tool in which different statistics can be plugged in. A natural question is to evaluate the extend to which different pdf estimations and definitions of divergence affect the effectiveness of the distinguisher. For this purpose, we carried out attacks based on the traces that are publicly available in the DPA Contest [10] and computed the success rate defined in [22] in function of the number of traces available to the adversary (*i.e.* encrypted messages), over 1000 independent experiments, using a Hamming weight leakage model. The results of these experiments are in Figure 3 from which we can extract different observations: First, classical attacks using the correlation coefficient are the most effective in this simple context. Second, the pdf estimation tools have a stronger impact than the notion of divergence on the MIA-like attacks. In particular and as far as non-parametric pdf estimations are concerned, the Kernel-based MIA performs significantly better than its counterpart using histograms. Eventually, it is worth noting the good behavior of the normalized KS and Cramér-von-Mises tests for which pdf estimation is not required. They are interesting alternatives to the other tests because of their simple implementation which makes them comparable to plain histograms in terms of processing workload. The Cramér-von-Mises criterion seems to behave as efficiently as the kernel-based methods, while avoiding the (hard) problem of choosing the kernel bandwidth. Hence, an intriguing open problem is to determine wether this test can be as efficient in more challenging contexts (e.g. implementations protected with masking or other countermeasures).

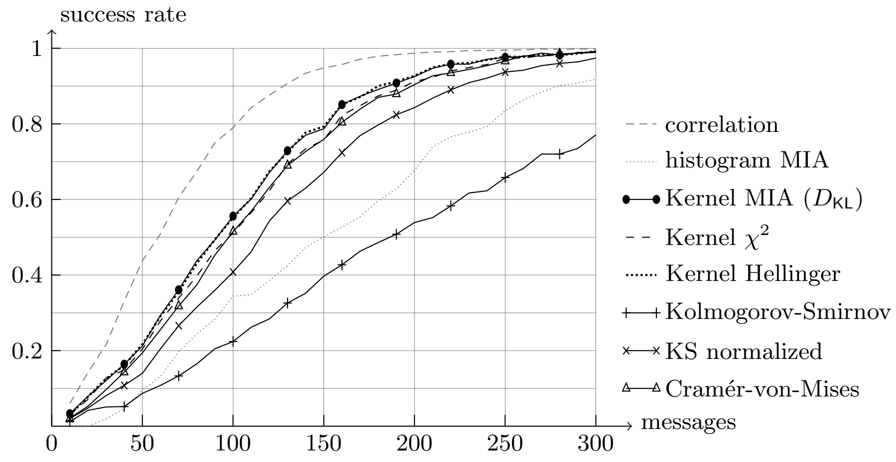


Fig. 3. Success rate of different attacks against the first DES S-box in the DPA Contest.

4 When to use it: MIA versus correlation

The experiments of Figure 3 suggest (as already emphasized by the authors in [12]) that when a reasonable leakage model is known by the adversary, standard DPA techniques such as using Pearson's correlation coefficient are more efficient than MIA. Hence, an obvious question is to determine the existence of contexts in which MIA would be necessary. With this respect, it is shown in [19] that higher-order attacks against masking schemes are good examples of such situations. This is essentially because MIA easily generalizes to multivariate statistics and hence does not need to worry about the combination of the leakages such as, *e.g.* [18]. In this section, we aim to show that MIA can even be useful in a univariate context, as soon as the adversary's leakage model is sufficiently imprecise.

Theoretically, this can be easily explained as follows. Let us assume that the leakages Y can be written as the sum of a deterministic part X_P (representing a perfect model) and a gaussian distributed random part R (representing some noise in the measurements): $Y = X_P + R$ and that a side-channel adversary exploits a leakage model $X_A = f(X_P)$. In ideal scenarios, we have $X_A = X_P$ but in practice, there generally exist deviations between the adversary's model and the perfect model, here represented by the function f. Correlation attacks are asymptotically successful as long as $\rho(X_A^g, Y) > \rho(X_A^w, Y)$, *i.e.* the correlation for the model corresponding to a correct subkey (with g superscript) is higher than the one for a wrong subkey candidate (with w superscript). If the adversary's model can again be written as $X_A = X_P + R'$ with R' another additive Gaussian noise, then correlation attacks will obviously remain the best solution. But in general, imprecisions in the models can take any shape (not only additive). This may lead correlation attacks to fail where, *e.g.* MIA can still succeed.

As an illustration, an interesting case that is reasonably connected to practice is to assume a data bus in a micro-controller such that one bit (say the LSB) leaks significantly more than the others (*e.g.* because of a larger capacitance). Taking the example of Section 3, this time with the 8-bit AES S-box, we could imagine that the leakages equal: $y = \sum_{i=1}^{8} a_i \cdot [\mathsf{S}(p \oplus k)]_i$. If the bit coefficients $a_i = 1$ for all i, we have Hamming weight leakages again. But by increasing a coefficient (*e.g.* a_1) and keeping the same Hamming weight model for the adversary, we can force this model to be arbitrarily wrong. Figure 4 illustrates the results of attacks that simulate this scenario. It shows that the number of messages required to reach a given success rate always increases with a_1 for the attacks using the correlation coefficient. By contrast, it stabilizes at some point for the MIA and KS test. Hence, for a sufficiently "wrong" leakage model, MIA-like attacks become useful. It is worth noting that the stabilization observed for the MIA and KS tests can be understood by looking at the pdf for a correct subkey candidate in Appendix B (again simplified to a 4-bit example): once a_1 is sufficiently heavy for the global pdf to be made of two disconnected pdf (one for $[\mathsf{S}(p \oplus k)]_1 = 0$, one for $[\mathsf{S}(p \oplus k)]_1 = 1$), the effectiveness of these distinguishers remains constant. Eventually, it is worth mentioning that while the MIA better resists to incorrect models than correlation attacks, it is not immune against them. One still requires that $\mathrm{I}(X_A^g; Y) > \mathrm{I}(X_A^w; Y)$. In other words, choosing random models will obviously not lead to successful attacks.

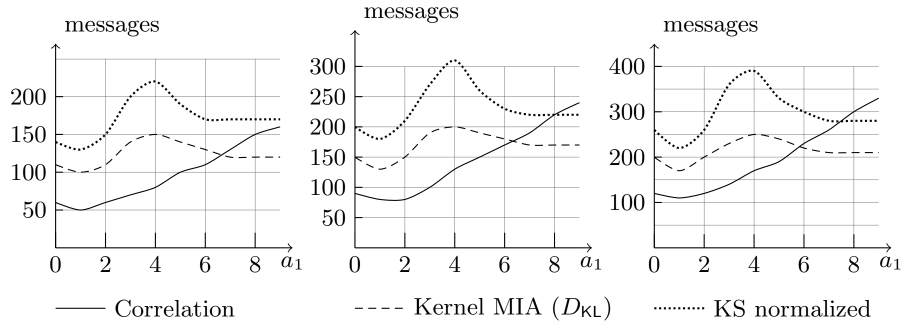


Fig. 4. Weight of the first leaking bit versus number of messages needed to reach a success rate of 50% (left), 75% (middle) and 90% (right), for different attacks.

5 Why to use it: MIA as an evaluation metric

Since the primary goal of the MIA is to distinguish subkeys, an adversary is not directly concerned with the value of $\mathrm{I}(X_A^g;Y)$ but rather with the fact that it is higher than $\mathrm{I}(X_A^w;Y)$. However, once a successful attack is performed, one can also wonder about the meaning of this value. In other words, can the mutual information $\mathrm{I}(X_A^g;Y)$ additionally be used as an evaluation metric for side-channel attacks, as the information theoretic metric suggested in [22]?

In order to discuss this question, we can again take the simple example of the previous section in which the leakages are the sum of a perfect model and a Gaussian noise: $Y = X_P + R$. Say the target subkey in a side-channel attack is denoted by a variable K. The model in [22] suggests to evaluate a leaking implementation with $\mathrm{H}[K|Y]$. Because of the additive noise, this can be written as: $\mathrm{H}[K|Y] = \mathrm{H}[K|X_P] + \mathrm{H}[X_P|Y]$. Additionally assuming that $R = 0$, we find: $\mathrm{H}[K|Y] = \mathrm{H}[K|X_P]$. By contrast, the MIA does not directly apply to the subkey variable, but to subkey-dependent leakage models. That is, assuming that an adversary performs MIA with a perfect leakage model, it computes: $\mathrm{I}(X_P;Y) = \mathrm{H}[X_P] - \mathrm{H}[X_P|Y]$ with $\mathrm{H}[X_P|Y] = 0$ if $R = 0$. Using the relation:

$$\mathrm{I}(K;X_P) = \mathrm{H}[K] - \mathrm{H}[K|X_P],$$

we have that if an adversary performs the MIA with a perfect leakage model and no noise (and a perfect pdf estimation tool), the following equation holds:

$$\mathrm{H}[K|Y] = \mathrm{H}[K|X_P] = \mathrm{H}[K] - \mathrm{I}(X_P;Y),$$

or similarly: $\mathrm{I}(K;Y) = \mathrm{I}(X_P;Y).$

It implies that MIA and the metric of [22] can be used equivalently in this case. Adding additive noise R to the leakages will not change the situation since it will simply add a term $\mathrm{H}[X_P|Y]$ to the previous equations. But as in Section 4, this equality does not hold anymore if the adversary's model is not perfect and the imperfections are not simply additive, *i.e.* if we have $Y = f(X_P) \neq X_P + R$. Then, the previous equality will turn into an inequality:

$$\mathrm{H}[K|Y] \leq \mathrm{H}[K] - \mathrm{I}(X_P;Y),$$

or similarly: $\mathrm{I}(K;Y) \geq \mathrm{I}(X_P;Y).$

That is, the mutual information computed by the MIA with an incorrect leakage model will tend to underestimate the amount of information leaked by the chip. In other words, MIA is a generic distinguisher while the conditional entropy $\mathrm{H}[K|Y]$ is a generic evaluation metric for side-channel attacks. The reason of this genericity comes from the information theoretic nature of these tools. In practice, MIA can be used to approach a fair evaluation metric if a perfect leakage model is available to the adversary but it deviates from this metric as soon as this conditions is not respected anymore[1]. This deviation essentially comes from the

[1] When moving to multivariate statistics, perfect models should be considered for each sample which yields the open question of how to efficiently exploit multiple models.

need to use an intermediate variable (corresponding to an intermediate value in the target algorithm, *e.g.* an S-box output) in non profiled side-channel attacks rather than considering the subkey leakages directly. That is, MIA computes $\mathrm{I}(X_P;Y)$ rather than $\mathrm{H}[K|Y]$. Summarizing, the MIA and the model of [22] have different objectives, namely recovering keys for MIA and allowing fair evaluations of leaking devices for the model. They also generally exploit different adversarial contexts, namely non-profiled attacks for the MIA and profiled attacks for the model. But eventually, the reason for using these tools is similar since they both allow capturing any kind of dependencies in the physical leakages and consequently lead to generic attacks and evaluation of the attacks and leakages.

References

1. S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B (Methodological).* vol 28, num 1, pp 131-142, 1966.
2. T. W. Anderson. On the distribution of the two-sample cramér-von mises criterion. *The Annals of Mathematical Statistics*, 33 (3) : 1148–1159, 1962.
3. Sébastien Aumonier. Generalized correlation power analysis. In *Ecrypt Workshop on Tools For Cryptanalysis.* Krakòw, Poland, September 2007.
4. P. Bickel, E. Levina. The earth's mover's distance is the mallows distance: some insights from statistics. In *Computer Vision 2001*, vol 2, pp 251-256.
5. E. Brier, C. Clavier, F. Olivier. Correlation power analysis with a leakage model. In *CHES 2004*, LNCS, vol 3156, pp 16-29, Boston, MA, USA, August 2004.
6. S. Chari, J. Rao, P. Rohatgi. Template attacks. In *CHES 2002*, Lecture Notes in Computer Science, vol 2523, pp 13-28, CA, USA, August 2002.
7. T.M. Cover, J.A. Thomas. *Elements of Information Theory.* Wiley, 1991.
8. Imre Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Sci. Math. Hungar.* vol 2, pp 229-318, 1967.
9. Imre Csiszár and Paul C. Shields. Information theory and statistics: a tutorial. *Commun. Inf. Theory*, vol 1, num 4, pp 417-528, 2004.
10. DPA Contest 2008/2009, http://www.dpacontest.org/
11. David Freedman and Persi Diaconis. On the histogram as a density estimator. *Probability Theory and Related Fields*, vol 57, num 4, pp 453-476, December 1981.
12. B. Gierlichs, L. Batina, P. Tuyls, B. Preneel. Mutual information analysis. In *CHES 2008*, LNCS, vol 5154, pp 426-442, Washington DC, USA, August 2008.
13. Wolfgang Härdle. *Smoothing Techniques: With Implementation in S.* Springer Series in Statistics. December 1990.
14. P. Kocher, J. Jaffe, B. Jun, Differential power analysis. In *Crypto 1999*, LNCS, vol 1666, pp 398-412, Santa-Barbara, CA, USA, August 1999.
15. K. Lemke, C. Paar. Gaussian mixture models for higher-order side channel analysis. In *CHES 2007*, LNCS vol 4227, pp 14-27, Vienna, Austria, September 2007.
16. Nan Laird, Arthur Dempster, Donald Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B (Methodological).* vol 39, num 1, pp 1-38, 1977.
17. C. L. Mallows. A note on asymptotic joint normality. *The Annals of Mathematical Statistics*, vol 43, num 2, pp 508-515, 1972.
18. T.S. Messerges. Using second-order power analysis to attack DPA resistant software. In *CHES 2000*, LNCS vol 1965, pp 238-251, Worcester, USA, August 2000.

19. Emmanuel Prouff and Matthieu Rivain. Theoretical and practical aspects of mutual information based side channel analysis. In *ACNS, Applied Cryptography and Network Security*, LNCS, vol 5536, pp 499-518, Paris, June 2009.
20. David W. Scott. On optimal and data-based histograms. *Biometrika*, vol 66, num 3, pp 605-610, December 1979.
21. Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman & Hall/CRC, April 1986.
22. Francois-Xavier Standaert, Tal G. Malkin, and Moti Yung. A unified framework for the analysis of side-channel key recovery attacks (extended version). Cryptology ePrint Archive, Report 2006/139, 2006. `http://eprint.iacr.org/`.
23. Berwin A. Turlach. Bandwidth selection in kernel density estimation: a review. In *CORE and Institut de Statistique*, 1993.
24. M.H. Zhang, Q.S. Cheng. Determine the number of components in a mixture model by the extended ks test. *Pattern Recogn. Lett.*, 25 (2), pp 211–216, 2004.

A Density estimation techniques

A.1 Histograms

For n bins noted b_i, the probability is estimated as:

$$\hat{\Pr}[\underline{b_i} \leq x \leq \overline{b_i}] = \frac{\#b_i}{q}, \text{ where } q = \sum_{0 \leq j \leq n} \#b_j$$

The optimal choice for the *bin width* h is an issue in Statistical Theory, as different bin sizes can greatly modify the resulting model. For relatively simple distributions, which is usually the case of side-channel leakages, reasonable choices are Scott's rule [20] ($h = 3.49 \times \hat{\sigma}(x) \times n^{-1/3}$) and Freedman-Diaconis' rule [11] ($d = 2 \times \mathrm{IQR}(x) \times n^{-1/3}$, IQR = interquartile range). While histograms are quite easy to implement, they generally provide a very slow convergence towards the target pdf, lack smoothness and heavily depend on bin width.

A.2 Kernel density estimation

The probability is estimated as:

$$\hat{\Pr}[X = x] = \frac{1}{nh} \sum_i K\left(\frac{x - x_i}{h}\right),$$

where the kernel function K is a real-valued integrable function satisfying $\int_{-\infty}^{\infty} K$ $(u)\ du = 1$ and $K(u) = -K(u)$ for all u. Some kernel functions are in Table 1. Similarly to histograms, the most important parameter is the *bandwidth h*. Its optimal value is the one minimizing the AMISE (Asymptotic Mean Integrated Squared Error), which itself usually depends on the true density. A number of approximation methods have been developed, see [23] for an extensive review. In our case , we used the modified estimator [21, 13]:

$$h = 1.06 \times \min\left(\hat{\sigma}(x), \frac{\mathrm{IQR}(x)}{1.34}\right) n^{-\frac{1}{5}}$$

Kernel	$K(u)$	Kernel	$K(u)$
Uniform	$\frac{1}{2}i(u)$	Triweight	$\frac{35}{32}(1-u^2)^3 i(u)$
Triangle	$(1-\lvert u\rvert)i(u)$	Tricube	$\frac{70}{81}(1-\lvert u\rvert^3)^3 i(u)$
Epanechnikov	$\frac{3}{4}(1-u^2)i(u)$	Gaussian	$\frac{1}{\sqrt{2\pi}}exp\left(-\frac{1}{2}u^2\right)$
Quartic	$\frac{15}{16}(1-u^2)^2 i(u)$	Cosinus	$\frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right) i(u)$

Table 1. Some kernel functions. i is defined as: $i(u) = 1$ if $|u| \leq 1$, 0 otherwise.

A.3 Gaussian mixtures

This parametric method models the pdf as:

$$\hat{\Pr}(X = x) = \sum_{i=0}^{n-1} w_i \; \mathcal{N}(x, \mu_i, \sigma_i),$$

where the μ_i and σ_i are the respective means and deviations of each mixture component. This method can be thought of as a generalization of the kernel density estimation with gaussian kernels, where one is not restricted to $w_i = \frac{1}{nh}$ or $\sigma_i = \frac{1}{h}$. The main advantage of the finite mixture method is that it usually leads to a number of mixture elements significantly smaller than the number of samples used to form the model in a kernel density estimation. An efficient algorithm called the *Expectation Maximization* (EM) algorithm [16] allows one to give a good approximation of a pdf in the form of a finite mixture. Given the number of components in the mixture, it computes their weights and gaussian parameters. Some additional procedures have been proposed that help choosing the number of components to be used in a mixture, for example in [24].

B Effect of incorrect leakage models

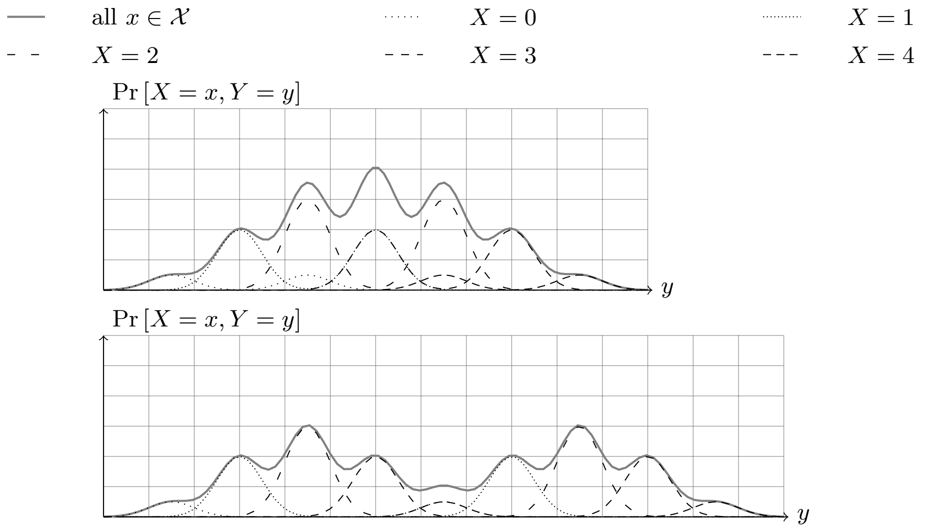


Fig. 5. Behavior of the probability densities for the correct subkey in a 4-bit DPA, assuming a Hamming weight leakage model and $a_1 = 3$ (up) and $a_1 = 5$ (down).