Mutual information analysis: higher-order statistical moments, efficiency and efficacy

Mathieu Carbone, Yannick Teglia, Gilles R. Ducharme, Philippe Maurine

Noname manuscript No.
(will be inserted by the editor)

Mutual Information Analysis: higher-order statistical moments, efficiency and efficacy

Mathieu Carbone · Yannick Teglia · Gilles R. Ducharme · Philippe Maurine

Received: date / Accepted: date

Abstract The wide attention given to the Mutual Information Analysis (MIA) is often connected to its *statistical genericity*, denoted flexibility in this paper. Indeed, MIA is expected to lead to successful key recoveries with no reliance on *a priori* knowledge about the implementation (impacted by the error modeling made by the attacker. and with as minimum assumptions as possible about the leakage distribution (*i.e.* able to exploit information lying in any statistical moment and to detect all types of functional dependencies), up to the error modeling which impacts its efficiency (and even its effectiveness). However, emphasis is put on the powerful generality of the concept behind the MIA, as well as on the significance of adequate *Probability Density Functions* (PDF) estimation which seriously impacts its performance. By contrast to its theoretical advantages, MIA suffers from underperformance in practice limiting its usage. Considering that this underperformance could be explained by suboptimal estimation procedures, we studied in-depth MIA by analyzing the link between the setting of tuning parameters involved in the commonly used nonparametric density estimation, namely *Kernel Density Estimation* (KDE) with respect to three criteria: the statistical moment where the leakage prevails, MIA's efficiency and its flexibility according to the classical Hamming weight model. The goal of this paper is therefore to cast some interesting light on the field of PDF estimation issues in MIA for which much work has been devoted to finding improved estimators having their pros and cons, while little attempt has been made to identify whether or not existing classical methods can be practically improved according to the degree of freedom offered by hyperparameters (when available). We show that some 'optimal' estimation procedures following a problem-based approach rather than the systemic use of heuristics following a accuracy-based approach can make MIA more efficient and flexible and a practical guideline for tuning the hyperparameters involved in MIA should be designed. The results of this analysis allowed us defining a guideline based on a detailed comparison of MIA's results across various simulations and real-world datasets (including publicly available ones such as DPA contest V2 and V4.1).

Keywords Side-Channel Analysis · Mutual Information · Bandwidth · Statistical moments

Mathieu Carbone · Yannick Teglia
ST Microelectronics - Advanced System Technology
Avenue Célestin Coq, 13790 Rousset, France
E-mail: firstname.lastname@st.com

Mathieu Carbone · Philippe Maurine
LIRMM - Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier
161, Rue Ada, 34090 Montpellier Cedex 5 France
E-mail: firstname.lastname@lirmm.fr

Gilles R. Ducharme
EPS - Institut de Mathématiques et de Modélisation de Montpellier
2, Place Eugène Bataillon, Université Montpellier 2, 34095 Montpellier Cedex 5, France
E-mail: firstname,lastname@univ-montp2.fr

Philippe Maurine
CEA - Centre Microélectronique de Provence Georges Charpak
880, Route de Mimet, 13541 Gardanne, France
E-mail: firstname.lastname@cea.fr

1 Introduction

The Side Channel Attacks (SCA) are a subclass of physical cryptanalysis giving *transparency* about the secret

of cryptographic device thanks the emanation of unintentional side channel observables. The core idea of SCA is to take advantage of key-dependent side channel observables (*e.g.* power consumption or Electromagnetic (EM) emissions) emanating from the cryptosystem, *i.e.* key recovery objective through exploring the data dependency between physical leakages and the internal state of the cryptosystem. Hence, the exploitation of these leakages by comparing them with key-dependent prediction models should lead to identify which key hypothesis is the most likely to derive to the leakage measurements. In practice, the two important requirements are to choose a suitable prediction model in order to highlight the dependency between the observations and predictions and to use an adequate comparison (statistical) tool, usually called *distinguisher* capable to detect this dependency and discriminate efficiently the correct key at the end.

Since the seminal work of Kocher *et al.* [19], the improvement of SCA from the statistical view point has gained prominence in research community. As a result, a refinement of the initial DPA, *i.e.* SCA using Difference of Means (DoM) as distinguisher, was proposed in [12] using the Pearson's Correlation as new distinguisher. This leads to the so-called Correlation Power Analysis (CPA) which aims at looking for linear (or at least largely monotonic) relationship between side-channel observations and predictions. More recently, a distinghuisher based on Linear Regression Analysis (LRA) was proposed in [15]. It can be viewed as a non-profiling variant of the stochastic approach [36]. LRA aims at fitting the best model which linearizes the leakage by minimizing the sum of squared residuals using Ordinary Least Squares (OLS) method. Although LRA is perceived by a large part of the community as the most efficient attack so far in the case where the model drifts away from the Hamming weight model [15], [22], this approach is originally limited as it only detects leakages hidden on the first-order statistical moment (*i.e.* mean) independently of the using basis functions. However, the preprocessing explained in [11,37,29] allows us to conduct the latter at higher orders in univariate scenario by raising centered (and standardized for orders ¿ 2) traces to some power corresponding to order. However, Knowing 'a priori' the leakage order can be challenging in non-profiled context. In 2008, a powerful SCA, called Mutual Information Analysis (MIA) has been proposed in [3] and independently formalized in [17]. Based on an information theoretic approach, MIA aims at being a 'generic' attack theoretically disclosing secret with no reliance on *a priori* knowledge about the implementation and no assumption on the leakage distribution. These so-called 'generic' methods are of interest because they indicate in some sense the inherent vulnerability of a cryptosystem independently from the physical implementation details. In [51], authors rethink the notion of 'genericity' in SCA context putting emphasis on the role of the leakage model rather than on the distinguisher. They state that generic strategies should follow from the properties of the used leakage model, *i.e.* injectivity of the chosen cryptographic functions as intermediate variable, and such a definition facilitates conclusive statements about attack outcomes independent of the distinguishing statistic chosen. Based on this definition, the term of 'genericity' for MIA is directly related to the identity leakage model [17] but this interesting approach turns out into an ineffective solution to mount MI-based attacks targeting an AES-Sbox (because of injective property of this target function) up to the work in [35] where authors propose to target as a suitable non-injective function, *i.e.* the MixColumns operation, for an effective attack in AES. Independently of the used leakage model, MIA is therefore likely to efficiently exploit any information lying on any statistical moment whatever is the functional dependency of these information with the adopted prediction model. In the literature, emphasis is put on the generality of the concept behind the MIA, as well as on the significance of adequate *Probability Density Functions* (PDF) estimation which seriously impacts its performance. Indeed, it is usually considered that the accuracy of PDF estimates determine MIA's efficiency. Several works have therefore proposed MIA enhancements according to this intuition. All of them aim at improving the accuracy of PDF estimates involved in the computation of MI index. Various ways to estimating PDF were followed and most publications point out the sensitivity of MI estimator to the chosen estimation method as well as to the setting of parameters involved in it. Among all methods at disposal, two approaches, referred to as *parametric* and *nonparametric* can be adopted according to the level of knowledge on the underlying distributions. Parametric approaches are especially suitable if a particular shape (*e.g.* Gaussian assumption) of the underlying PDF is identified as correct due to some application specific reasons. In contrast, nonparametric approaches do not rely on assumptions on the distributions from data which are drawn. Thus, their objective coincides with the one pursued by SCA community *i.e.* the quest of a 'flexible' distinguisher and this even at the cost of a significant computation overheads. Nevertheless, despite all the interesting and valuable former works, MIA still remains disappointing in practice. Indeed, empirical evaluations of MIA have indicated that, in scenarios favouring correlation DPA (such as those where the data-dependent

leakage is known to be well approximated by the Hamming weight) it is highly unlikely to offer any advantage over the latter against either simulated or real-world traces, that the correlation-based distinguisher continues to outperform the MI-based up to quite a high degree of divergence between the model and the true leakages. Besides, within this context, our contribution comes from the consideration that the underperformance of MIA could be explained from sub-optimal estimation procedures as it was noticed in [13] in which some interesting light were casted on the field of estimation issues in MIA. We

Our Contribution. Almost all publications on MIA have followed an accuracy-based approach. This means that they intuitively followed the idea according to which the use of accurate PDF estimators in the sense of *Mean Integrated Squared Error* (MISE), for which several heuristics have been derived, is the right direction to obtain efficient and flexible MIA. However, this does not guarantee maximal MIA's efficiency as shown in [13] since at the end an attacker normally cares more about distinguishing the correct key with a certain confidence rather than accurately estimating the PDF involved in the attack. In this paper, we therefore adopt a problem-based approach to give us insights on the right way to finetune hyperparameters in nonparametric approaches for MIA by the use of a refinement of distinguishing criterion introduced in [13].

2 Notations and Preliminaries

We use capital letters, like X, to denote a random variable (RV), calligraphic letters, like $\mathcal{X}$, to denote its support (set of possible values), and lowercase letters like x, for its realizations. The expectation operator of X is denoted by $\mathbb{E}[X]$. We further denote the term key, *i.e.* k, for the attacked round key byte.

We thereafter give a summary of an univariate and non-profiled vertical SCA (*i.e.* a DPA-like attack) workflow.

1. **Acquire** n physical leakage traces over $\mathcal{L}$ denoting as the observation space, *i.e.* $l_i \in \mathcal{L}^d$, $i = 1, \ldots, n$, corresponding to cryptographic computations operated by device during encryptions or decryptions. We hereafter suppose each leakage measurement consists in d physical realizations. Each of them contains information about intermediate values used internally $z_{k^*,i} = F(x_i, k^*)$, $i = 1, \ldots, n$, where $x_i \in \mathcal{X}$ is the i^{th} public value, *i.e.* input (plaintext) or output (ciphertext) byte of the cryptographic device and $k^* \in \mathcal{K}$ the secret key byte. At a sample point t, we assume the leakage L_t to be composed of two parts: a deterministic part $\phi_t(\cdot)$ and an independent additive noise B_t such that

$$L_t = \phi_t(Z_k^*) + B_t \;\hookrightarrow\; l_{i,t} = \phi_t(z_{k^*,i}) + b_{t,i} \tag{1}$$

   where $l_{i,t} \in \mathcal{L}$, $t = 0, \ldots, d-1$ denotes the leakage value in the i^{th} leakage trace at the sample point t and b_i denotes its leakage noise value.
2. **Predict**, for each key byte guess $k \in \mathcal{K}$ and for each $x_i \in \mathcal{X}$, $i = 1, \ldots, n$, a *sensitive* intermediate value $z_{k,i} \in \mathcal{Z}$, *e.g.* $z_{k,i} = \mathsf{Sbox}(x_i \oplus k)$ for block ciphers. Usually, $\mathcal{X}$,$\mathcal{K}$, $\mathcal{Z}$ are taken as $\mathbb{F}_2^m$, where m is the number of bits (for AES (resp. DES) $m = 8$ (resp. $m = 4$))
3. **Model**, for each predicted value, the physical leakage $\hat{\phi}(Z_k) = H_k \hookrightarrow \hat{\phi}(z_{k,i}) = h_{k,i} \in \mathcal{H}$, *e.g.* Hamming Weight (HW) [26] or Hamming Distance (HD) [12]. At this point, a common step often referred as *leakage partitioning* allows classifying the leakage samples, based on predictions $h_{k,i}$. The expectation is that predictions obtained from the correct key hypothesis (*i.e.* $k = k^*$) will lead to a meaningful leakage partition, *i.e.* there will be a dependency between L_τ and H_k, where τ represents a Point of Interest (PoI), *i.e.* leaking information about k^*. By contrast, a wrong key hypothesis $k \neq k^*$ should give rise to random predictions, so that the partition will only correspond to a random shuffling of leakages samples. Hence, $\hat{\phi}$ should be a good approximation of ϕ_τ to highlight the dependence between L_τ and H_k.
4. **Compare**, using a side-channel distinguisher D, the key-dependent models and the actual physical leakages and decide which is the most probable key byte guess $\hat{k^*} = \arg\max_{k \in \mathcal{K}} \left(\hat{\mathsf{D}}(h_{k,i}, l_{\tau,i}\right)$ where τ represents a Point of Interest (PoI), *i.e.* leaking information about k^*.

3 Mutual Information Analysis

Mutual Information Analysis (MIA) in SCA was introduced in [3,17] in order to catch any functional and/or statistical dependencies among random variables relaxing the linear assumption. Let (X, Y) be a hybrid random vector, that is X is discrete over $\mathcal{X}$ while Y is continuous with support $\mathcal{Y}$. The theoretical version of this index is defined as

$$\mathrm{MI}[Y; X] = \sum_x l(x) \int_{\mathcal{Y}} f(y|x) \log\left(\frac{f(y|x)}{g(y)}\right) dy, \tag{2}$$

where $f(y|x)$ is the conditional (on X) PDF of Y while $g(y)$ (resp. $l(x)$) is the marginal PDF of Y (resp. X)[1] and the symbol $\sum_x$ refers to a sum taken over values x of X such that $l(x) > 0$. There are other equivalent formulas defining the MI index, notably,

$$\mathrm{MI}[Y;X] = \mathrm{H}[Y] - \mathrm{H}[Y|X], \tag{3}$$

$$= \mathrm{H}[Y] - \sum_x l(x)\mathrm{H}[Y|x], \tag{4}$$

where $\mathrm{H}[Y] = -\int_{\mathcal{Y}} g(y)\log(g(y))\,dy$ is the (differential) entropy of random variable Y and similarly $\mathrm{H}[Y|x] = -\int_{\mathcal{Y}} f(y|x)\log(f(y|x))\,dy$.

Specializing formula (3), its application as an attack in $\mathcal{T}$-domain can be framed as the expectation (with respect to the conditioning value) of the Kullback-Leibler divergence between the global and partitioned traces at a leakage sample τ and for each key hypothesis $k \in \mathcal{K}$

$$\mathrm{MI}_k(\tau) = \mathrm{H}[L_\tau] - \mathrm{H}[L_\tau|H_k] \tag{5}$$

$$= \mathrm{H}[L_\tau] - \mathop{\mathbb{E}}_{h\in\mathcal{H}}\Big[\mathrm{H}[L_\tau|H_k = h]\Big]. \tag{6}$$

An estimate $\hat{k^*}$ of k^* is obtained as

$$\hat{k^*} = \arg\max_{k\in\mathcal{K}} \Big\{\hat{\mathrm{MI}}_k(\tau)\Big\}. \tag{7}$$

The main difficulty in implementing a MIA is in estimating the values $\mathrm{MI}_k(\tau)$. In contrast to Pearson's coefficient which is easily estimated via sample moments, the estimation of the MI index requires the estimation of the underlying PDF which is both theoretically and practically, a non trivial statistical problem for the studied nonparametric methods in this paper. Many methods have been proposed to estimate entropy as histograms [17], kernels [33,34,45], B-splines [48], maximal information coefficient [21], cumulants [20], *etc.* An overview of density estimation techniques applied in MIA is given in [20]. Neither parametric nor nonparametric estimators are universally preferable in all situations, however.

3.1 Estimating a PDF

Suppose a sample of independent copies $\{(x_i, y_i)\}_{i=1}^n$ of (X, Y) is at disposal. The problem of estimating the MI index in Eq. (4) requires estimators of the entropies $\mathrm{H}[Y]$ and $\mathrm{H}[Y|x]$, which in turn requires estimators of the PDF $g(y)$ and $f(y|x)$. As stated earlier, estimation of these underlying PDF is a difficult statistical problem. In general, a PDF estimator must offer a good trade-off between accuracy (bias) and variability (variance). In this section, we present the classical histograms and the more sophisticated KDE method. For the interested reader, details about another nonparametric method B-splines, can be found in [48]. Note that, for simplicity, we restrict attention to the case of univariate PDF.

3.1.1 Histograms.

The most used and studied nonparametric PDF technique is probably the *histogram method*. The histogram estimator of $g(y)$ is obtained by partitioning the support of Y, noted $\mathcal{Y}$, into m bins $B_j = [b_{j-1}, b_j)$, with $b_0 < b_1 < \cdots < b_m$ such that $\mathcal{Y} \in [b_0, b_m)$. Then

$$\hat{g}_{hist}(y) = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{I}\{y_i \in B(y)\}}{\ell(B(y))}, \tag{8}$$

where $\mathbb{I}\{A\} = 1$ if event A is realized and 0 otherwise, $B(y)$ is the bin that contains y, $\ell(B(y))$ is its length and n is the sample size. The resulting estimator of $\mathrm{H}[Y]$ is

$$\mathrm{H}_{hist}[Y] = -\sum_{j=1}^{m} \hat{g}_{hist}\,(q_j)\log \hat{g}_{hist}\,(q_j)\,\ell(B_j), \tag{9}$$

where $q_j = (b_{j-1} + b_j)/2$. Likewise, to obtain the histogram estimator of $\mathrm{H}[Y|x]$, let $n_x = \sum_{i=1}^n \mathbb{I}\{x_i = x\}$. Estimate $l(x)$ by n_x/n and $f(y|x)$ by

$$\hat{f}_{hist}(y|x) = \frac{1}{n_x}\sum_{i=1}^{n} \frac{\mathbb{I}\{y_i \in B(y)\}\mathbb{I}\{x_i = x\}}{\ell(B(y))}, \tag{10}$$

so that

$$\mathrm{H}_{hist}[Y|x] = -\sum_{j=1}^{m} \hat{f}_{hist}\,(q_j|x)\log \hat{f}_{hist}\,(q_j|x)\,\ell(B_j). \tag{11}$$

These quantities are plugged into (4) to get the histogram estimator of the MI index.

Various methods for optimal binning (*i.e.* bin width, bin size) in statistical literature have been proposed. When the underlying distribution is Gaussian, typical and reasonable choices are Scott's rule [40]

$$\ell_{scott} = 3.49\ \hat{\sigma}\ n^{-1/3} \tag{12}$$

and Freedman-Diaconis rule [16] $\ell_{fd} = 2\ \hat{\mathrm{IQR}}\ n^{-1/3}$ in case of equal binning where $\hat{\sigma}$ and $\hat{\mathrm{IQR}}$ are denoted as the sample standard deviation and interquartile range of the data $\{y_i\}_{i=1}^n$, respectively.

[1] Formally $l(x)$ is a probability mass function (PMF) because X is discrete. To simplify notation, we use the generic acronym PDF

3.1.2 Kernels.

Another commonly used and also well-studied PDF technique is the *kernel method*. The KDE of $g(y)$ is then given by

$$\hat{g}_{KDE}(y) = \frac{1}{n}\sum_{i=1}^{n} K_h\left(y - y_i\right),\ y \in \mathbb{R}, \tag{13}$$

where $K_h(y) = h^{-1}K(y/h)$ with $K : \mathbb{R} \to \mathbb{R}$ is a symmetric, non negative function called the *kernel function*, and satistyfing $\int_{\mathbb{R}} K(x)\ dx = 1$ and for which $h > 0$ is a real parameter, called the *kernel bandwidth* or also-called *smoothing parameter*. Regarding the kernel function, classical choices are the bell curve, best known as Gaussian function : $K(y) = \frac{1}{\sqrt{2\pi}}e^{-y^2/2}$ or the clipped upside-down parabola, *i.e.* Epanechnikov function : $K(y) = \frac{3}{4}(1 - y^2)$ for $|y| \leq 1$. In general, the shape of the kernel function K has little influence on the estimated density [41] compared to the choice of the bandwidth h which is crucial in controlling the trade-off between bias and variance (*i.e.* degree of smoothing). A relatively good estimator of an optimal bandwidth (in the sense of minimizing an approximation of the integrated mean squared error) is obtained by Silverman's rule [42]

$$h_S = c\ \hat{\sigma}\ n^{-1/5}, \tag{14}$$

where $c = 1.06$ (resp. 2.34) for Gaussian's (resp. Epanechnikov's) kernel and $\hat{\sigma}$ the sample standard deviation of the data $\{y_i\}_{i=1}^n$. More details can be found in [18] for interested readers. From Eq. (4), $\mathrm{H}_{KDE}[Y]$ can be estimated by

$$\mathrm{H}_{KDE}[Y] = -\int_{\mathcal{Y}} \hat{g}_{KDE}\left(y\right) \log \hat{g}_{KDE}(y)\, dy, \tag{15}$$

and similarly

$$\mathrm{H}_{KDE}[Y|x] = -\int_{\mathcal{Y}} \hat{f}_{KDE}\left(y|x\right) \log \hat{f}_{KDE}\left(y|x\right)\ dy, \tag{16}$$

where $\hat{f}_{KDE}\left(y|x\right)$ is obtained in the same manner as $\hat{g}(y)$ but with the part of the data having $x_i = x$ while $l(x)$ can be estimated by n_x/n where $n_x = \sum_{i=1}^{n} \mathbb{I}\{x_i = x\}$. At the end, these are plugged into Eq. (4) to produce the kernel estimator of the MI index.

At this stage, another hurdle is encountered because the above computations require integration. To reduce the computational cost, one can choose points $\mathcal{Q} = \{q_0 < \ldots < q_b\}$ (referred to as *query points*) and estimate $\mathrm{H}[Y]$ by

$$\hat{\mathrm{H}}_{KDE}[Y] = -\sum_{j=1}^{b} \hat{g}_{KDE}(q_j) \log \hat{g}_{KDE}(q_j)(q_j - q_{j-1}), \tag{17}$$

and similarly with $\hat{\mathrm{H}}_{KDE}[Y|x]$ in place of Eq. (16). It is noteworthy that this processing plays no role in the key discrimination process as it does not depend on the predictions. These query points in $\mathcal{Q}$ must be properly chosen to provide mathematical accuracy of the integral principally depending on their number which can be made arbitrarily good by increasing $|\mathcal{Q}|$, at the expense of computational costs. Hence, one should naturally choose these query points to be systematically fixed along a mesh grid covering all the sample points, whose coarseness depends on the available computing power and especially by the selected bandwidth value. Indeed, it is noteworthy that $|\mathcal{Q}|$ depends heavily on the bandwidth value, *i.e.* small bandwidths require a finer grid. These points are not chosen to provide statistical accuracy of the estimator (a difficult problem) but solely mathematical accuracy of the integral, a different problem for which various solutions exist, for example via the rectangular method (used in the remaining thesis) or through more sophisticated quadrature formulas. Besides, one makes KDE method *iterative* like histogram method differentiating reference data $\{y_i\}_{i=1}^n$ with query points $\{q_j\}_{j=0}^b$. We stress that heuristics have been developed with the view of getting a globally good estimate of a PDF [41] that should be unsuitable in all scenarios in SCA context yielding to a loss of efficiency and flexibility of MIA. Recall that selection of tuning parameters has always been a dilemma: on the one hand, asymptotic arguments and reference distributions lead to plug-in and reference rules whereby tuning parameters can be easily calculated, but these perform poorly on finite samples and when reference distributions do not match reality. On the other hand, data-driven selection criteria should give appropriate tuning parameters as proposed in [13] focusing on a suitable *adaptive* bandwidth selection from the viewpoint of the attack, *i.e.* a problem-based approach.

4 Influence of tuning parameters involved in KDE

4.1 Statement

To show the effect of the choice of the bandwidth value, the number of query points and the kernel type on

kernel-based MIA, a small simulation study in a *perfect* (classical) scenario with synthetic data was first conducted. Two thousand simulated leakage measurements were drawn from the following 'linear' leakage function according to the Hamming Weight leakage model of a typical AES S-box output $\mathrm{HW}(Z_{k^*})$ with additional independent additive Gaussian noise with mean 0 and variance σ^2

$$L = \mathrm{HW}(Z_{k^*}) + B \tag{18}$$

Although not always realistic, the preliminary investigation of this scenario is justified by the numerous works carried out under this assumption, as a reference. We used here synthetic data by fixing $\sigma = 8$ and considering that $\mathrm{HW}(Z_{k^*})$ follows binomial distribution with parameters 8 and 0.5, *i.e.* $\mathcal{B}(8, 0.5)$ so that the exact value of the theoretical MI index (= 0.0153) could be computed.

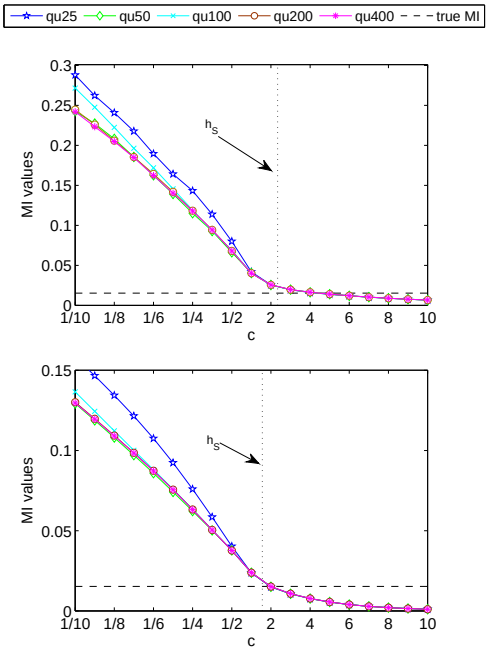

Fig. 1 Behavior of the estimator of the MI index with respect to the number of query points (qu) and to the constant c from Eq. (14) using Epanechnikov (top) and Gaussian (bottom) kernels. Each Silverman-based heuristic h_s depends of the used kernel function.

Figure 1 shows the results of estimating the MI index as the bandwidth h and the number of (equispaced) query points are changed. The setting of bandwidth value is evaluated over a grid ranging from some point of the neighborhood of the h_S (bandwidth obtained from the Silverman's rule) to some small/large multiple of this value by varying the constant c in Eq (14). For

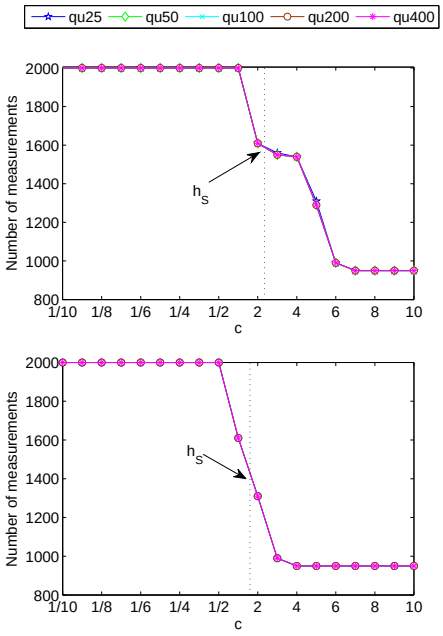

Fig. 2 Plots of Minimum Trace to Disclose (MTD) of kernel-based MIA with respect to the number of query points (qu), to the constant c from Eq. (14) using Epanechnikov (top) and Gaussian (bottom) kernels.

sake of readability, we stop at $c = 10$ but it is noteworthy that this practical choice does not affect the insights given by the further results. As expected, there is link between the bandwidth size and the choice of the number of query points. The higher the number of query points the more accurate the MI estimate is when the bandwidth value is lower than the Silverman's heuristic h_S. This circumvents the choice of their position when PDF are under-smoothed. However, this effect tends to vanish when the bandwidth value increases, *i.e.* near and greater than Silverman's heuristic h_S. This suggests that a reduction of the number of query points could be considered in order to reduce the computational costs. Besides, no impact is observed regarding the kernel type used and the main focus must be on the bandwidth value which notably influences the resulting MI estimate. Interestingly, Silverman's rule h_S with the constant c selected according to the kernel type (*i.e.* $c = 1.06$ (resp. $c = 2.34$) for Epanechnikov (resp. Gaussian) kernel) yields a good estimate of the actual MI index. Note also that as the bandwidth is increased, the bias of the MI estimator increases (hence its variance decreases) as the estimator (*i.e.* MI) decays to zero. This is explained by the fact that, as h increases, all KDE get over-smoothed and converge to the same function that resemble the initial kernel spreaded over the support $\mathcal{Y}$, with the entropies converging to the same value and the MI index vanishing. All this dove-

tails nicely with intuition and the admonishments in almost all publications on MIA that, in order to have a good estimator of the MI index, one should use adequate PDF estimators. However, this does not guarantee maximal MIA's efficiency.

By mounting an attack using Eq. (18), Figure 2 interestingly shows that increasing the bandwidth results in more efficient attacks, in terms of Minimum Trace to Disclose (MTD), *i.e.* less number of measurements is required to disclose the correct key hypothesis. The MTD condition is considered to be fulfilled when the correct key hypothesis is accurately disclosed after the processing of the 2000 measurements (*i.e.* with a relative margin $> 10\%$). Besides, the number of query points do not significantly impacts the attack's efficiency when it sufficiently sets large. Summarizing, this suggests that good PDF estimation does not necessarily translate in efficiency of the attack, where larger bandwidths and smoother PDF estimators, seem to yield better results in this simulation case.

To sustain these results, we performed the same framework using real-world measurements from DPA Contest v2 campaign [1] which is expected to be close to the latter simulation in Eq. (18). We used HD leakage model (word level) targeting the output of S-box 2 at the last round and setting Epanechnikov kernel and 400 equidistant query points to perform the KDE. The actual results in Figure 3 showing evolution of the success rates with respect to different bandwidth values (according to the constant c), match with those from the simulation.

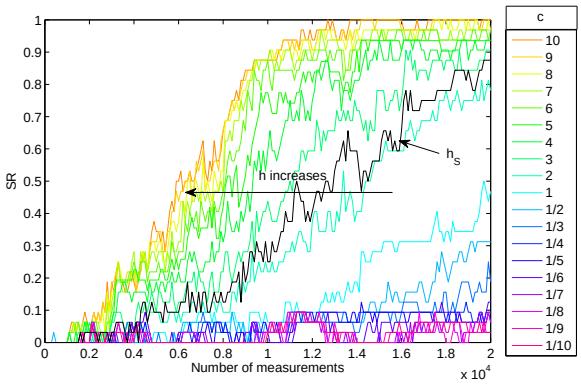

Fig. 3 Plot of success rates of kernel-based MIA with respect to the constant c from Eq. (14) using the measurements from DPA Contest V2.

It is this counterintuitive behavior that has led to the realization that the bandwidth could be seen, not as a nuisance parameter to be dealt with in a statistical estimation procedure, but more profitably as a lever that could be used to fine-tune a SCA. Some results in [6,34] go in the same direction showing that the traditional criterion for minimizing the approximation error is irrelevant since the most efficient attacks are those using the histogram method with a bad choice of bin parametrization for this criterion.

4.2 Distinguishing rule

Our proposed distinguishing rule follows a problem-based approach (specific to SCA) which explicitly exploits the fact that there is exactly one correct key hypothesis k^* trying to maximize the contrast between this key and all the remaining others according to a distinguisher and a leakage model. In MIA, to successfully distinguish the correct key hypothesis k^* from others, $\hat{\mathrm{MI}}_{k^*}(\theta) > \hat{\mathrm{MI}}_k(\theta)$, $\forall k \in \mathcal{K}\backslash\{k^*\}$ (except for a special case reported in [49,50] where authors showed that using 'identity' model based on drop-bit approach can lead to a different key candidate selection). Note that we now denote by θ the tuning parameter of a PDF estimation tool (*e.g.* number of bins for histogram or bandwidth value for KDE).

By letting $\hat{\mathrm{MI}}_k(\theta)$ be an estimator of MI_k parameterized by θ in all PDF involved in Eq. (6), an alternate *bounded* expression [2] into $[-1;1]$ of (15) in [13] is

$$\hat{k^*} = \arg\max_{k\in\mathcal{K}} \left\{ \max_{\theta\in\Theta} \left[\frac{|\hat{\mathrm{MI}}_k(\theta) - \overline{\hat{\mathrm{MI}}_{-k}(\theta)}|}{\left(\sum_{k\in\mathcal{K}} \left(\hat{\mathrm{MI}}_k(\theta) - \overline{\hat{\mathrm{MI}}_{-k}(\theta)}\right)^2 \right)^{1/2}} \right] \right\}, \tag{19}$$

where $\overline{\hat{\mathrm{MI}}_{-k}(\theta)}$ stands for the mean of all estimators except $\hat{\mathrm{MI}}_k(\theta)$, Θ represents the set of a running PDF tuning parameter (here, different bandwidth values). In order to prevent the special case reported in [49,50] and give a general definition of our distinguishing rule in view of other applications, we considered the absolute value of the numerator. The following maximization on $\theta \in \Theta$ aims at making this discrimination independent of its choice (value) of θ (*i.e.* quality of PDF estimation) resulting in an automatic (without 'a priori' knowledge k^*) selection procedure targeting the goal of getting a good estimate of k^*, in contrast to Silverman's rule that aims at getting good estimates of the PDF involved in $\hat{\mathrm{MI}}_k(\theta)$, *i.e.* focusing on a problem-based approach rather than an accuracy-based approach in the

[2] To avoid over-fitting θ in some cases, *i.e.* when leakage is embedded on higher-order statistical moment (not explored in [13]) leading to an incorrect key distinguishability

SCA context. Also, when analyzing a set of traces over many sample points $t \in \{0, \ldots, d-1\}$ in Eq. (19), a double maximization operation over θ and t must be performed, with the result being the operand of $\arg\max_{k \in \mathcal{K}}$. By looking for a *outlier behavior*, our approach aims at efficiently and methodically deriving 'optimal' setting of a tuning parameter in the sense of distinguishability rather than the distinguisher value itself. In practice, this distinguishing rule can be viewed as an 'on-the-fly' attack focusing on the flexibility aspect or as a way to characterize an optimal PDF tuning parameter value and give insights according to a scenario focusing on the efficiency aspect, *i.e.* the value of θ where the inner max operator over all $k \in \mathcal{K}$ is attained, denoted as

$$\theta_{opt} = \arg\max_{\theta \in \Theta} \left\{ \max_{k \in \mathcal{K}} \left[\frac{|\hat{\mathrm{MI}}_k(\theta) - \overline{\hat{\mathrm{MI}}_{-k}(\theta)}|}{\left(\sum_{k \in \mathcal{K}} \left(\hat{\mathrm{MI}}_k(\theta) - \overline{\hat{\mathrm{MI}}_{-k}(\theta)} \right)^2 \right)^{1/2}} \right] \right\}, \tag{20}$$

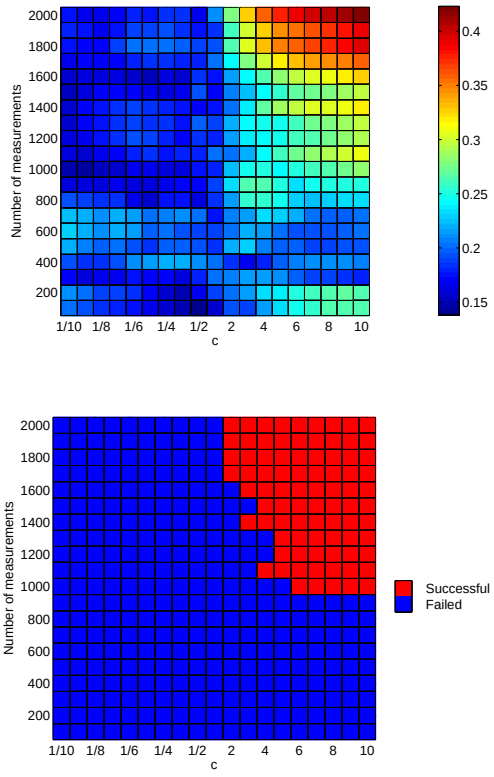

Fig. 4 Evolution of maximization over $k \in \mathcal{K}$ of the distinguishing rule (top) and the key recovery status (bottom) with respect of the constant c and number of measurements using simulated data from Eq. (18).

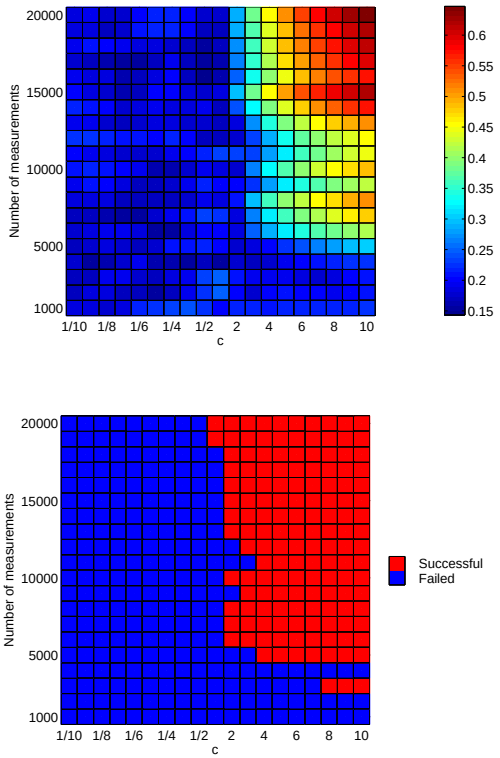

Fig. 5 Evolution of maximization over $k \in \mathcal{K}$ of the distinguishing rule (top) and the key recovery status (bottom) with respect of the constant c and number of measurements using real-world data from DPA contest v2.

The action of our distinguishing rule is illustrated in Figures (4) and (5) for simulated from Eq. (18) and real-world data from DPA contest v2, respectively. We used Epanechnikov kernel and 400 query points for KDE estimation procedure. It can be noticed that a maximization over $\theta \in \Theta = \{c\ \hat{\sigma}\ n^{-1/5} : c \in \{\frac{1}{10}, \frac{1}{9}, \ldots, 9, 10\}\}$ and $k \in \mathcal{K}$ allows a good probability of discrimination for k^* as *warm* colors in left part of Figures 4 match with location where k^* is disclosed in right part of Figures 5. In practice, since the adversary does not necessarily possess a copy of the DUT (*i.e.* profiling step) for which he knows '*a priori*' the secret key, the definition of our rule remains suitable in non-profiled scenario to find 'optimal' setting of a tuning parameter. In [31], authors assume a (semi-)profiled scenario for a similar optimization criterion of the filter coefficients on CPA.

Figure 6 displays the distinguishing rule in Eq. (19) (*i.e.* term into brackets in Eq. (19)) over $\Theta = \{c\,\hat{\sigma}\,n^{-1/5} : c \in \{\frac{1}{10}, \frac{1}{9}, \ldots, 9, 10\}\}$ and for each key hypothesis after the processing of 2000 measurements corresponding to the simulation from Eq. (18). It is noteworthy that this approach allows verifying previous observations regarding bandwidth setting, *i.e.* larger bandwidth than the commonly used Silverman's rule-of-thumb h_S was expected to give better results in this case. Even if the key

is disclosed using h_S, one should notice that increase bandwidth value lead to a more accurate discrimination as the relative margin also grows.

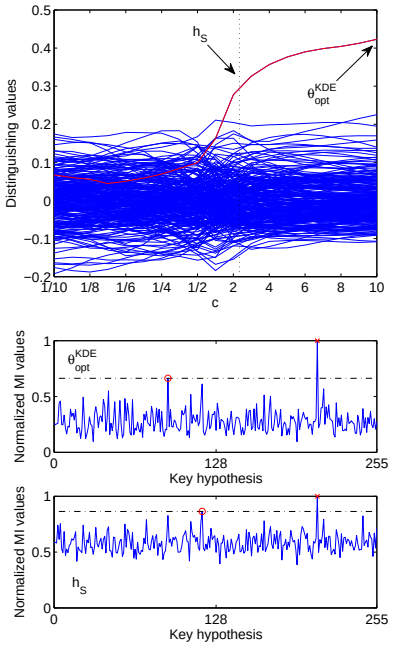

Fig. 6 Plot of distinguishing rule with respect to the constant c from Eq. (14), *i.e.* $h_S = c\ \hat{\sigma}\ n^{-1/5}$ and for each key hypothesis after the processing of 2000 measurements (top) (red line: correct key guess, blue lines: incorrect key guesses). Plots of normalized MI values over all key hypothesis according to different bandwidth values: h_S, *i.e.* $c = 2.34$ (for Epanechnikov kernel) and θ_{opt}^{KDE}, *i.e.* $c = 10$ after the same processing of 2000 measurements (bottom) (red cross: correct key guess, red circle: nearest-rival key guess).

Same observations were made using the real-world measurements from DPA contest v2 focusing on the first data set (among the 32 available data sets) and targeting the key byte 1 (output of the S-box 2) at the last round using a HD leakage model as depicted in Figure 7.

This behavior was replicated with many other data sets for which successful key recoveries were also observed with CPA [12] or LRA using *linear* basis function [15]. However at this stage, rather than displaying these results, it appears more interesting to further explore how to set the bandwidth h in a broader case where the latters could fail in the presence of more complex leakage.

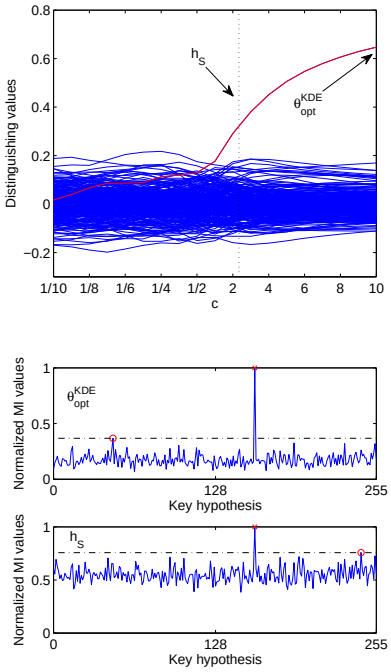

Fig. 7 Plot of distinguishing rule with respect to the constant c from Eq. (14), *i.e.* $h_S = c\ \hat{\sigma}\ n^{-1/5}$ and for each key hypothesis after the processing of 20000 measurements (top) (red line: correct key, blue lines: incorrect key guesses). Plots of normalized MI values over all key hypothesis according to different bandwidth values: h_S, *i.e.* $c = 2.34$ (for Epanechnikov kernel) and θ_{opt}^{KDE}, *i.e.* $c = 10$ after the processing of 20000 measurements (bottom) (red cross: correct key guess, red circle: nearest-rival key guess).

5 Practical investigation of efficiency/flexibility of MIA

In this section, we report results obtained by simulations before reporting experiments done with real-world datasets in order to give guidelines on how to set tuning parameters for MIA using nonparametric method: KDE. The goal was to analyze the evolution of both efficiency and flexibility of various distinguishers with respect to the statistical moment containing the leakage but also with respect to the values of the different tuning parameters of these distinguishers if any. These examples are expected to be reflective of the variety of leakage functions that one can find in SCA context. We proceed by a pragmatic approach based on a two-stage procedure as follows

Firstly, simulations (experiments on synthetic data sets) were conducted with different scenarios in order to control the drift from linearity of the functional dependency on the univariate leakage based on the first-order statistical moment. Secondly, leakage measurements have been simulated using on the specific countermeasure, called Rotating Sboxes Masking [30] which

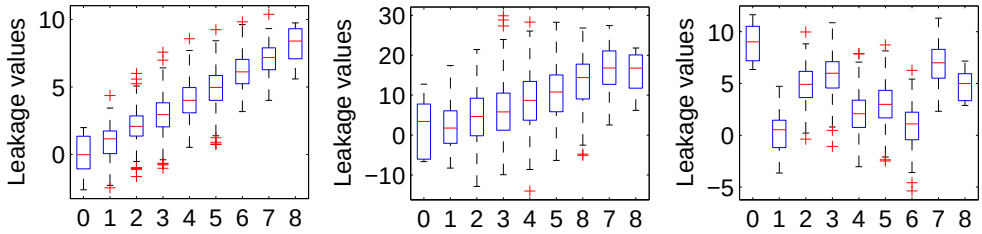

Fig. 8 An example of boxplot illustrating each scenario: 1 (left), 2 (middle) and 3 (right) with $\log_2(\text{SNR}) = 0$ for the correct key hypothesis.

is an instance of LEMS where the univariate leakage is embedded on higher-order statistical moment, *i.e.* 4^{th} according to the chosen mask table. The real world data sets were finally chosen to validate observations based on simulations.

Following the definition in [14,32], we recall that an univariate leakage is embedded in o^{th}-order statistical moment if the distributions of the random variables $(L_t|H_k = j),\ j \in \mathcal{H}$ differ when k ranges over $\mathcal{K}$ whereas the $(o-1)^{th}$-order statistical moments do not, *i.e.* minimal order statistical moment exploitable in the univariate leakage.

5.1 General setting of DPA-like attacks

Side Channel Distinguishers. We considered seven non-profiled distinguishers: DPA, [8,25], CPA [12], AoV [47,43]), LRA [15] and finally three MI-based distinguishers, *i.e.* one using parametric approach of cumulants [20] and two using nonparametric approaches based on histograms [17] and KDE [33,34,44] as PDF estimation tool.

For the Histogram-based MIA, we considered the Gierlichs' heuristic rule which consists in choosing number of bins according to size of predictions, *i.e.* prediction space: $\mathcal{H}$ corresponding to 9 bins for AES. This method was used as benchmark without taking into account an optimization of the number of bins.

For the related KDE-MIA,

- the number of query points was set to 400 so as to define a grid of equidistant points covering all the observations.
- we adopted the Epanechnikov kernel. It should be stated that similar results were obtained with the Gaussian kernel.
- we considered two different values of the kernel bandwidth. The first one was computed using Silverman's heuristic, *i.e.* h_S from Eq. (14) while the second one was computed using the distinguishing rule in Eq. 20, *i.e.* θ_{opt}^{KDE}.

Attacks target. The typical AES S-box output, *i.e.* $Z_{k^*} = \mathsf{Sbox}(X \oplus k^*)$, where $\mathsf{Sbox} : \mathbb{F}_2^8 \to \mathbb{F}_2^8$ corresponding to the SubBytes operation, X is uniformly distributed over $\mathbb{F}_2^8$, and represents a varying plaintext byte while $k^* \in \mathbb{F}_2^8$ represents the key byte to recover.

Model choice. We chose the very classical model proven efficient in practice, *i.e.* Hamming Weight function (HW), for the model-based attacks, *i.e.* all except LRA and DPA

5.2 Resuls on simulated datasets

In the SCA literature, most of the reported simulation results were obtained by considering an independent Gaussian noise added to a more or less complex function of the Hamming Weight representing the leakage behavior with processed data.

5.2.1 Leakage embedded on first-order statistical moment

Leakage simulations. We adopted the same approach to generate various functions integrating the leakage in the first-order statistical moment. More precisely, three different leakage functions $\phi(\cdot)$ in Eq. (1) were considered. Box plots of each scenario are given in Figure 8. This resulted in three leakage model scenarios, in which $B \sim \mathcal{N}(0, \sigma^2)$ models a Gaussian additive noise with mean 0 and variance σ^2:

- in ***scenario n°1***, $\phi(\cdot)$ is simply the Hamming Weight function; a *perfect* linear leakage model is therefore considered in this reference scenario:

$$L = \text{HW}(Z_{k^*}) + B \tag{21}$$

- in ***scenario n°2***, $\phi(\cdot)$ is assumed to be an unevenly weighted Hamming Weight function; this model therefore introduces an arbitrarily chosen error in the HW model. In our case, we adopted as in [45,50]:

$$L = \text{HW}([Z_{k^*}]_{1 \le i \le 7}) + 10 \cdot [Z_{k^*}]_8 + B \tag{22}$$

– in ***scenario n°3***, $\phi(\cdot)$ is a balanced non-linear leakage function obtained using a fixed permutation: $\mathsf{Perm} = \begin{pmatrix} 0\,1\,2\,3\,4\,5\,6\,7\,8 \\ 8\,0\,5\,6\,2\,3\,1\,7\,4 \end{pmatrix}$ of the Hamming weight function:

$$L = \mathsf{Perm}(\mathrm{HW}(Z_{k^*})) + B \tag{23}$$

Experimental results. In each scenario, when a distinguishability of a key candidate is sufficiently highlighted (heuristically determined) according to the SNR as defined in [24], we observed θ_{opt}^{KDE} is always significantly larger than the commonly used h_S value when looking at the optimal value in the set $\Theta = \{\frac{h_S}{20}, \frac{h_S}{19}, \ldots, 19 \cdot h_S, 20 \cdot h_S\}$. Thereafter we systematically set θ_{opt}^{KDE} to $12 \cdot h_S$ to obtain a version of KDE-MIA completely free of parameters. This suggests that over-smoothing PDF during a MIA leads to a better discrimination of the correct key. Note that results for LRA are evaluated using the linear basis functions but also discussed according to higher basis functions.

Figure 9 reports the Success Rate (SR) metric calculated from 100 independent attacks performed with the seven considered distinguishers for each leakage scenario. All attacks were evaluated with 2000 simulated values of the leakage. For all the scenarios, the efficiency curves of each attack have the same evolution. This suggests us that the noise similarly impacts the efficiency of the attacks. In ***scenario n°1***, CPA is the most efficient attack while the LRA, DPA, AoV, Cumulant-MIA and KDE-MIA parameterized by θ_{opt}^{KDE} are equivalently ranked second followed by the KDE-MIA using h_S and Histo-MIA ranked last. As already observed in [15], the dominance of CPA is due to the hypothesis made over $\hat{\phi}(\cdot)$ that induces an optimal tracking of the linear model, *i.e.* a model that exactly corresponds to the leakage function. As expected LRA outperforms all other attacks in ***scenario n°2*** because the hypothesis $\hat{\phi}(\cdot)$ imperfectly models the leakage (*i.e.* the model is built under the incorrect hypothesis $\phi(\cdot) = \mathrm{HW}(\cdot)$, resulting in a loss of efficiency for model-based attacks. Nevertheless, LRA and CPA are no longer the best attacks in ***scenario n°3***. This may be caused by the fact that the leakage is non-linear and that it is therefore becoming more and more difficult to find a monotonic trend in the data which could be exploited by the CPA and LRA using the linear basis to succeed. Besides, the failure of these latter attacks is expected when they face to the highly non-linear nature of a leakage, as reported in [49] using a (simulated) dual-rail logic style. This indicates that the impact of assuming false leakage behavior (in terms of non-linearity) can be as catastrophic as misleading the adversary. However, one should increase the size of the basis involved in LRA (*e.g.* using not only linear, but quadratic, cubic,... terms) but this leads to an unexpected feature of LRA regarding the key candidate selection, similarly observed in [49,50]. Indeed, the behavior of the LRA, when increasing the basis, can be interpreted/understood thanks to a careful analysis of the algebraic description of the target device leakage which does not take any advantage in this scenario. In our results, this observation regarding LRA is independent of the SNR level. Basically, LRA which requires a

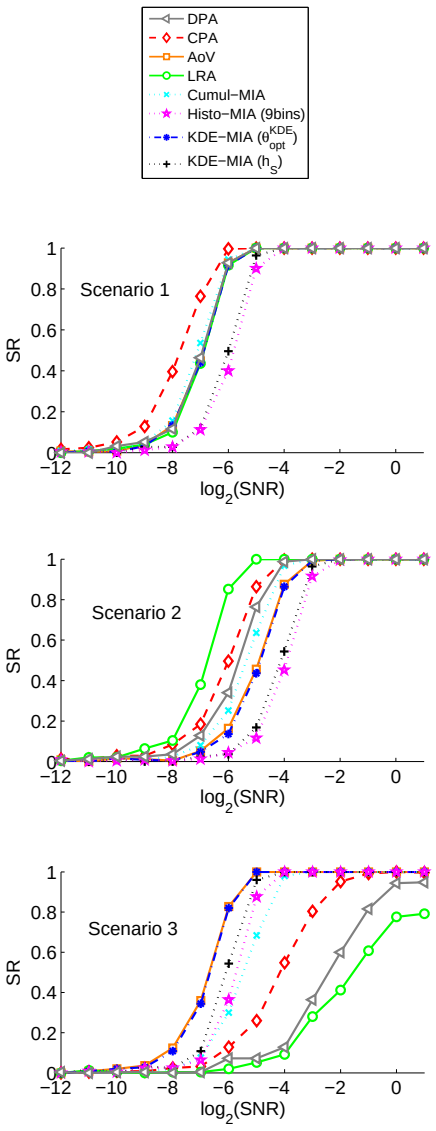

Fig. 9 Plots of SR for each scenario with respect to $\log_2(\mathrm{SNR})$.

set of well-chosen basis functions to perform efficiently more importantly needs to be justified by a reasonable physical intuition (*i.e.* having a connection with the ac-

tual leakages) as refining the model using larger basis is performed for all key candidates (*i.e.* not only the correct one). Consequently, we meet ideas developed in [49] in order to perform successful key recoveries, adversaries may be more interested in the 'outlier' behavior of the correct key model than in its distinguishing score with the measured leakages. Note that authors in [35] mitigate the impact of the negative results in [49,50] regarding the bit drop trick (for MIA using 'identity model'). It is noteworthy that our distinguishing rule could also be used to efficiently select the size of the basis functions for the LRA.

At this stage, the main observation resulting from these experiments is that choosing a bandwidth value larger than h_S leads to a more efficient KDE-MIA (but also more resistant to noise) when the leakage is enclosed in the first-order statistical moment. Moreover, it is noteworthy that MIA parameterized by θ_{opt}^{KDE} (*i.e.* using large h) is closely equivalent to AoV for all the scenarios. This surprising behavior related to the over-smoothing of PDF could be explained by the fact that when the Gaussian assumption is adopted, the key distinguishability is based on the respective shifts between the distributions associated to each HW values, and in this case the rate of convergence towards their real density functions does not play a key role. The possibility to use large bandwidth values when the leakage is embedded on the first-order statistical moment has also an impact on the computational costs. Indeed, it is not necessary to choose a large number of query points when the bandwidth value is large. Therefore, the computational costs of KDE-MIA can be drastically reduced when the first-order statistical moment embeds the leakage by simply reducing the number of query points. As regards other attacks, the three MIA using various nonparametric PDF approaches are more noise-resistant than parametric Cumulant-MIA. Interestingly, but as expected, KDE-MIA using θ_{opt}^{KDE} performs better than KDE-MIA using h_S and Histo-MIA when the leakage is enclosed in the first-order statistical moment. The explanation should lie in the power of PDF smoothing which is impacted by the choice of the estimation method and the setting of θ. Moreover, we observed that θ_{opt}^{histo} value is generally smaller than Gierlichs' heuristic (9 bins) over $\Theta = \{3, \ldots, 300\}$ considering θ^{histo} as the number of equal-width bins in each scenario. This empirically confirms our intuition of adopting an over-smoothing trend in PDF estimation procedure to obtain better MIA's efficiency. Note that using θ_{opt}^{histo} leads to similar results to using Gierlichs' heuristic. The choice of the bandwidth value when looking for a leakage embedded on the first-order statistical moment being clarified, let's analyze the impact of this choice when the leakage is embedded only in a higher-order statistical moment. As follows, we considered the fourth-order statistical moment related to the Low Entropy Masking scheme (LEMS) for which we consider the instance of Rotating Sboxes Masking (RSM) countermeasure [30].

5.2.2 Leakage embedded on higher-order statistical moment

Leakage simulations. To analyze the impact of the choice of h on the flexibility and efficiency of KDE-MIA when the leakage is enclosed on higher-order statistical moments, we generated synthetic leakage data associated to an instance of LEMS, called Rotating Sboxes Masking (RSM) countermeasure [30] which allows the cancellation of leakage based on the first three higher-order statistical moments. This was done using the following leakage scenario:

$$L^{\mathrm{RSM}} = \mathrm{HW}(Z_{k^*} \oplus M_i) + B, \quad i \in [0;15] \tag{24}$$

where $M_{i\in[0;15]}$ denotes the sixteen 8-bit base masks chosen in such a way that the exploited leakage of the masked variable is perceptible at the degree 4 (*i.e.* based on the fourth-order statistical moment). As explained in [9], a mask distribution taken as the 16 codewords of the $[8,4,4]$ linear code (extension with one parity bit of the $[7,4,3]$ Hamming code) satisfies this security level. We considered the same one used in the implementation of DPA Contest V4.1, *i.e.* $M_{i\in[0;15]} = \{$`0x00`, `0x0f`, `0x36`, `0x39`, `0x53`, `0x5c`, `0x65`, `0x6a`, `0x95`, `0x9a`, `0xa3`, `0xac`, `0xc6`, `0xc9`, `0xf0`, `0xff`$\}$ and an offset $i \in [0;15]$ randomly picked at each new simulated encryption.

Experimental results. In contrast with the case of a leakage embedded on the first-order statistical moment, we observed that the distinguishability of a key candidate is faster achieved in case of leakage embedded on the fourth-order statistical moment when $\theta_{opt}^{KDE} \approx h_S$. This suggests that a sufficient well-estimated PDF is required to catch a higher-order statistical moment. Figure 10 shows SR computed over 100 independent experiments in free-noise case. First of all, we observed that all the classical SCA exploiting leakage embedded on the first-order statistical moment in this context even in absence of noise, *i.e.* CPA, AoV, KDE-MIA with large bandwidth value $(12 \cdot h_S)$ and LRA, whatever higher basis functions are used, fail as well as DPA for which no model is required. Moreover, it can also be seen that Cumulant-MIA do not survive in this case making the Gaussianity assumption arguable. This sustains the results presented in [30] for DPA, CPA and AoV (called VPA in [30]) and grants

an immunity to LRA and (parametric) MIA using cumulants but it should be noticed that MI-based attacks using nonparametric methods can be easily mounted making questionable the countermeasure in this case as depicted in Figure 10. Indeed, Histo-MIA using 9 bins outperforms all other attacks due to the fact that the observation space is based on 9 possible values ensuring that no information is lost. This follows the first intuition in [17] which aimed at estimating the probability distributions as good as possible using as many bins as there are distinct values in the domain covered by the sample set. To sustain our purpose, we observed that $\theta_{opt}^{histo} \approx 9$ over $\Theta = \{3, \ldots, 300\}$ considering θ^{histo} as the number of equal-width bins. Nevertheless, Histo-MIA is less noise-resistant than KDE-MIA as displayed in Figure 11. Evidently, we also observed that alternative distribution based-attack to MIA, KSA succeed in recovering the correct key hypothesis.

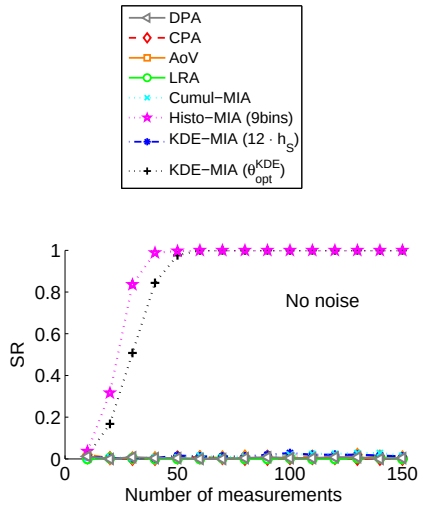

Fig. 10 Plot of Success Rate for all the attacks in free-noise case.

Besides, all the studied univariate first-order attacks can be extended to higher orders by introducing a preprocessing stage for the traces. In our context, This preprocessing involves the computation of mean-free standardized values which are then raised to power 4 (for a univariate 4th-order DPA-like attacks), e.g. as underlined in [38] for CPA. Unfortunately, as reported in [28,11], a higher-order attack typically requires huge number of traces, *i.e.* several (hundreds of) millions of traces to be successful and they become more susceptible to noise as the latter increases. Even in our free-noise case, all remains at a success rate of roughly 0 and a guessing entropy of 128, *i.e.* the quantity for a random guess without using side channel leakage after the processing of several thousands of traces thus giving an indubitable advantage for distribution-based attacks in this scenario. Besides the latters do not need to know at which order of the leakage prevails as they are expected to capture any PDF characteristics at the expense of using more traces by the increasing of the leakage order.

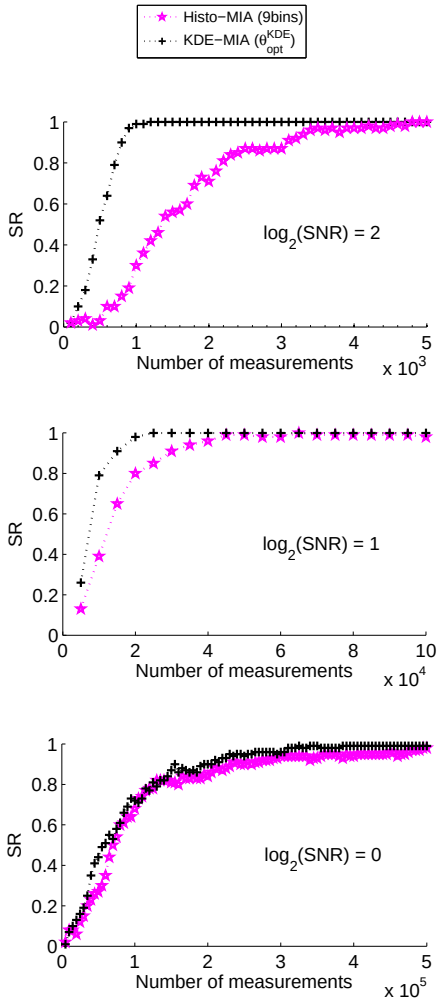

Fig. 11 Plots of SR for Histo-MIA (9 bins) and KDE-MIA using the distinguishing rule θ_{opt}^{KDE} (here, h_S) across different level noise $\log_2(\text{SNR})$.

We also used the class of Normal Inverse Gaussian (NIG) distributions [4] for which its parametrization allows controlling higher-order statistical moments and therefore building simulated leakages until fourth-order statistical moment. Let L^{NIG} be a RV following a NIG distribution with parameter vector $(\alpha, \beta, \mu, \sigma)$ as

$$L^{\text{NIG}} \sim \text{NIG}(\alpha, \beta, \mu, \sigma), \quad (25)$$
$$\mu \in \mathbb{R}, \alpha, \delta, \beta \in \mathbb{R}_+^*, 0 < |\beta| < \alpha,$$

with μ, σ, α and β the respective location, scale, tail heaviness and asymmetry parameters. For the sake of clarity, we set $\gamma = \sqrt{\alpha^2 - \beta^2}$ and its statistical moments are

$$\mathsf{Mean}(L^{\mathrm{NIG}}) = \mu + \sigma\frac{\beta}{\gamma} \qquad \mathsf{Var}(L^{\mathrm{NIG}}) = \sigma\frac{\alpha^2}{\gamma^3}$$

$$\mathsf{Skew}(L^{\mathrm{NIG}}) = 3\frac{\beta}{\alpha\sqrt{\sigma\gamma}} \qquad \mathsf{Kurt}(L^{\mathrm{NIG}}) = 3\frac{\alpha^2 + 4\beta^2}{\sigma\alpha^2\gamma},$$

The leakage measurements have been independently simulated for second, third and fourth-order statistical moment by evaluating one and taking the remaining ones (almost) constant. We noticed that θ_{opt}^{KDE} value is close to h_S when looking at the optimal value in the set $\Theta = \{\frac{h_S}{20}, \frac{h_S}{19}, \ldots, 19{\cdot}h_S, 20{\cdot}h_S\}$. This gives insights into the way of setting tuning parameters involved in PDF estimation when the leakage is embedded on higher-order statistical moment.

5.3 Results on Real-world datasets

To sustain the observations obtained on synthetic data, some experiments on real measurements were conducted.

Attacked Datasets. We focused on three different AES implementations (denoted by DUT #1, #2 and #3) with leakage characteristics similar to those of some of the scenarios we analyzed in section 5.2. Measurements of DUT #1 and #3 are publicly available from the second (v2) and fourth version (v4.1) of DPA contest campaign [1] while measurements of DUT #2 were acquired by ourselves. Here are some characteristics of these DUT:

- DUT #1: an unprotected hardware AES-128 implementation on Xilinx Virtex-5 FPGA.
- DUT #2: an unprotected AES-128 designed with a 65nm Low Power High Threshold Voltage CMOS technology integrating an in-house communication protocol and supplied by 16 pads so that the power consumed by the AES is not drawn from a single power pad. A Xilinx Spartan 3 FPGA board is used to drive the integrated chip. Our acquisition campaign setup provided us 150000 power traces which were acquired with a differential probe measuring the variations of Vdd and Gnd, using a 8-bit oscilloscope with a 20GS/s sampling rate and a bandwidth of 1MHz-4GHz. A picture of the IC showing the location of the AES on the die and the experiment set-up used to collect power traces is given in Figure 12.

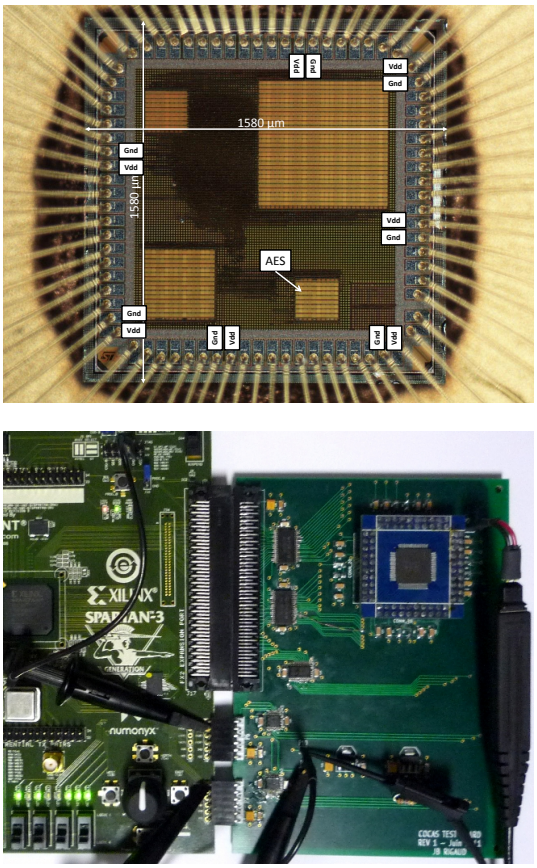

Fig. 12 Picture of the test chip DUT#2 showing the position of the 16 Vdd and Gnd pads (top) and measurement setup (left).

- DUT #3: a protected software AES-256 implementation on a 350nm metal-3 layer ATMEL AVR-163 microcontroller where the protection applied was an instance of low entropy masking scheme (LEMS): Rotating Sboxes Masking (RSM) [30].

Datasets Analysis. Before running further analyzes, it is noteworthy that each tested DUT represents a specific leakage simulation previously studied in Section 5.2. Indeed, DUT #1 should be identified to a linear leakage embedded on first-order statistical moment as in ***scenario n°1***, whereas DUT #2 to a non-linear leakage embedded on first-order statistical moment as in ***scenario n°3***. Indeed, a leakage precharacterization at the word level (HD) using the Akaike Criterion (AIC) [2] with the weighted least squares method, showed us a non-linear leakage for almost all the S-box according to high (resp. low) degree of the polynomials ranging from 4 to 8 for DUT #2 (resp. from 1 to 3 for DUT #1) at the word level. More details can be found in [47]. DUT #3 should be identified to leakage based on a higher-

order statistical moment, *i.e.* fourth, as also underlined in [52] and [27], corresponding to a real-world application of the simulated leakage in Section 5.2.2.

Leakage Detection. For DUT #1 and #2 (which is attempted to leak on the first-order statistical moment), we used the NICV [10] as a leakage detection technique. In contrast, for DUT #3, we naively selected τ as the sample point which maximizes the CPA results knowing the offset masks over the whole sample points of the online available campaign. For the sake of fairness, but not optimality, we selected one (the same) single PoI τ for the comparison of all the investigated attacks. An extension of NICV using MI to select PoI for which the leakage is embedded in higher-order statistical moment could be exploited.

Experimental results. We carried out univariate attacks on the key byte 0 (arbitrarily chosen) for each DUT. For sake of clarity and terminology in view of generalizing our insights to other nonparametric methods (*e.g.* histograms), we denote low (resp. high) resolution KDE-MIA, MIA for which the PDF estimation is done with a large (resp. small) bandwidth value. Here 'large' (resp. 'small') means typically that the bandwidth value $\approx 12 \cdot h_S$ (resp. h_S).

The results of these attacks are provided in Figure 13. To compute SR, we conducted on, DUT #1, 32 independent attacks since *public* acquisition campaign contains 640000 power measurements divided in 32 sets of 20000 measurements, *i.e.* each of them corresponds to a (different) random key used to encrypt 20000 random plaintexts. Regarding DUT #2 (resp. #3), we divided the 150000 (resp. 100000) acquired power (resp. EM) measurements with the same key into 100 subsets. As it can be noticed, results on real-world scenarios match with those obtained in simulated cases. More precisely, low resolution KDE-MIA performs better for DUT #1 and #2 than high resolution KDE-MIA since the linear and non-linear leakage is enclosed on the first-order statistical moment, respectively. High resolution KDE-MIA clearly outperforms all the other attacks for DUT #3. This can be explained by the fact that the leakage is based on the fourth-order statistical moment which can be captured by a MIA doing an accurate PDF estimation. The ease of this successful key recovery (*i.e.* the low number of measurements) is due to the used of a software implementation: ATMega 163 microcontroller for which the Signal-to-Noise ratio is very high [24]. Furthermore, less than 30 measurements were required to reach an success rate of 80% by univariate CPA attack knowing the mask as presented in [7]. All these experimental results confirm the observations we did in Section 5.2. We also conducted experiments on other key bytes following the same framework. Results are displayed in Figure 14. As attacks against masking are difficult [39] when the noise level is high, which is typically the case of hardware implementation (mainly due to algorithmic noise), one should rather implementing RSM countermeasure in hardware than in software implementation to increase its SCA robustness. In order to prevent SCA, another known technique called Shuffling, *i.e.* a random execution order of sensitive operations *e.g.* S-box. Note this countermeasure has been exactly used for the next version (v4.2) of the DPA contest [1].

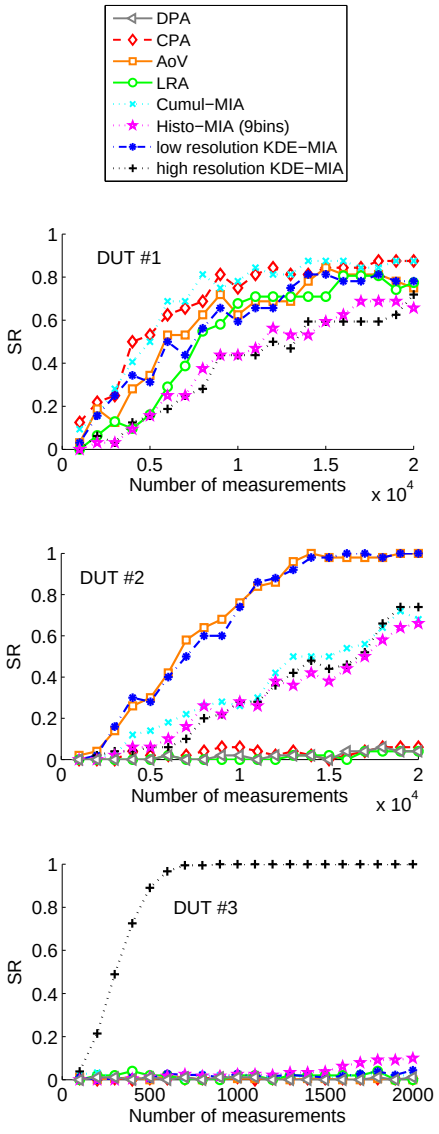

Fig. 13 Plots of SR for all investigated DUT.

Remark 1: By mounting a high resolution Histogram-based MIA using 256 bins for DUT #3 following the insights previously made, we obtained similar results with high resolution KDE-MIA sustaining the way of finely estimating probability distributions whatever the used PDF estimation tool. Moreover, we also observed that θ_{opt}^{histo} is high considering θ^{histo} as the number of equiwidth bins and $\Theta = \{3, \ldots, 300\}$.

Remark 2: For DUT #3, since the mask distribution is not uniformly distributed over $\mathbb{F}_2^8$, one can raise the *centered and standardized* leakage to the fourth power in order to exploit a *noisy* leakage embedded on the first-order statistical moment. Nevertheless, this leads in practice to an unsuccessful key recovery with CPA even when all the available measurements are used because of the noise which is also raised to the fourth power as well.

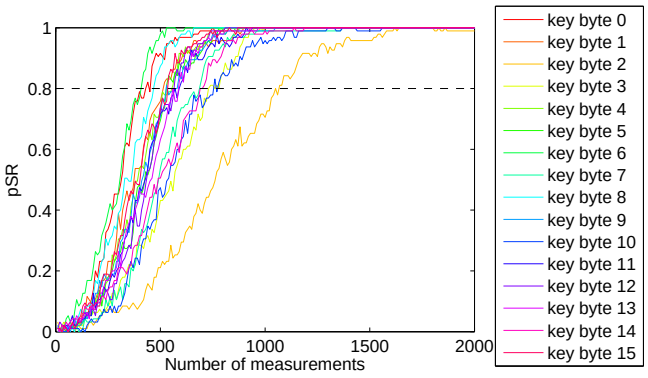

Fig. 14 Plots of partial Success Rate (pSR) for all the key bytes for DUT #3.

6 Conclusion

The introduction of the MIA was motivated by its theoretical ability in capturing all structures of functional and statistical dependencies between leakage and sensitive intermediate values; these latter being sometimes based on higher-order statistical moment because of the usage of some countermeasures. But the cost of this MIA feature lies in the difficulty in choosing adequately some tuning parameters. By focusing on the goal of optimizing KDE-MIA efficiency and flexibility instead of the auxiliary task of estimating PDF, we have defined practical guidelines for the implementation of efficient KDE-MIA through the proper selection of the bandwidth. These guidelines are based on the statistical moments in which the adversary aims at finding the leakage. The resulting bandwidths are usually larger than the commonly used h_S (obtained by Silverman's rule in Eq. (14)) and give better results in terms of attack efficiency across various simulations and real world experiments when the leakage is based on the first-order statistical moment (*i.e.* mean) which is usually the case in practice for unprotected implementations. In contrast smaller bandwidths (reported close to h_S) are required when the leakage is embedded on higher-order statistical moments in presence of some countermeasures (*e.g.* fourth-order in case of RSM with specific mask distribution) due to a need of an accurate PDF estimation. As a result, a trade-off between flexibility and efficiency can be adopted by adversaries according to a prior knowledge on the leakage characteristics. More generally, this work provides a characterization of MIA's efficiency/flexibility according to the resolution (based on tuning parameters) involved in nonparametric PDF estimation methods. Interestingly, one should note that low resolution MIA gives very similar results than AoV. We have shown that MIA conducted following these guidelines compares favorably with DPA [8,25], CPA [12], LRA [15] and AoV [43,5,23,46] and are even superior in some cases where the latters fail. The purpose of the distinguishing rule presented in this paper is mainly to formalize the intuition behind our results related to the PDF estimation step and provide study cases for which an *a priori* accuracy-based approach is not the straightforward way to achieve efficiency in SCA context. Note that the application of this rule depends on the *a priori* knowledge on the statistical moment embedding the leakage and finetune hyperparameters for some MI-based distinguishers according to the latter is not essential in practice since more general practical guideline can be drawn based on our conclusions. Through this work, we feel that various hyperparameters in the SCA contexts will be able to be set using this proposed rule and we believe that some benefits can be achieved in adapting the principles of statistical methods to the task at end: SCA in the present case. The formal investigation of our results to further talk about optimality for MIA should be an interesting perspective as well as formally define in which context MIA is more efficient than CPA.

References

1. TELECOM ParisTech SEN research group: DPA Contest. 2008-2014.
2. H. Akaike. Information theory and an extension of the Maximum Likelihood Principle. In B. N. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, Budapest, 1973. Akadémiai Kiado.
3. S. Aumonier. Generalized Correlation Power Analysis. In *ECRYPT Workshop on Tools For Cryptanalysis*, Kraków, Poland, September 2007.

4. O. Barndorff-Nielsen. Exponentially Decreasing Distributions for the Logarithm of Particle Size. *Royal Society of London Proceedings Series A*, 353:401–419, mar 1977.
5. L. Batina, B. Gierlichs, and K. Lemke-Rust. Differential Cluster Analysis. In C. Clavier and K. Gaj, editors, *Cryptographic Hardware and Embedded Systems*, volume 5747 of *Lecture Notes in Computer Science*, pages 112–127. Springer, 2009.
6. L. Batina, B. Gierlichs, E. Prouff, M. Rivain, F.-X. Standaert, and N. Veyrat-Charvillon. Mutual Information Analysis: a Comprehensive Study. In *J. Cryptology*, volume 24, pages 269–291, 2011.
7. P. Belgarric, N. Bruneau, J.-L. Danger, N. Debande, S. Guilley, A. Heuser, Z. Najm, O. Rioul, and S. Bhasin. Time-Frequency Analysis for Second-Order Attacks. In *CARDIS*, 2013.
8. R. Bevan and E. Knudsen. Ways to Enhance Differential Power Analysis. In *Information Security and Cryptology (ICISC)*, pages 327–342, Seoul, Korea, 2002.
9. S. Bhasin, C. Carlet, and S. Guilley. Theory of Masking with Codewords in Hardware: Low-Weight dth-order Correlation-Immune Boolean Functions. *IACR Cryptology ePrint Archive*, 2013:303, 2013.
10. S. Bhasin, J. Danger, S. Guilley, and Z. Najm. Side-Channel Leakage and Trace Compression using Normalized Inter-Class Variance. In R. B. Lee and W. Shi, editors, *HASP*, pages 7:1–7:9. ACM, 2014.
11. B. Bilgin, B. Gierlichs, S. Nikova, V. Nikov, and V. Rijmen. Higher-Order Threshold Implementations. In P. Sarkar and T. Iwata, editors, *ASIACRYPT 2014*, volume 8874 of *LCNS*, pages 326–343. Springer, 2014.
12. E. Brier, C. Clavier, and F. Olivier. Correlation Power Analysis with a Leakage Model. In *Cryptographic Hardware and Embedded Systems*, volume 3156 of *LNCS*, pages 16–29, Cambridge, MA, USA, August 2004. Springer, Heidelberg.
13. M. Carbone, S. Tiran, S. Ordas, M. Agoyan, Y. Teglia, G. R. Ducharme, and P. Maurine. On Adaptive Bandwidth Selection for Efficient MIA. In *COSADE*, 2014.
14. J. Coron, E. Prouff, and M. Rivain. Side Channel Cryptanalysis of a Higher Order Masking Scheme. In P. Paillier and I. Verbauwhede, editors, *Cryptographic Hardware and Embedded Systems*, volume 4727 of *LNCS*, pages 28–44. Springer, 2007.
15. J. Doget, E. Prouff, M. Rivain, and F.-X. Standaert. Univariate side channel attacks and leakage modeling. *J. Cryptographic Engineering*, 1:123–144, 2011.
16. D. Freedman and P. Diaconis. On the histogram as a density estimator:L2 theory. *Probability Theory and Related Fields*, 57:453–476, 1981.
17. B. Gierlichs, L. Batina, and P. Tuyls. Mutual Information Analysis : A Generic Side-Channel Distinguisher. In *Cryptographic Hardware and Embedded Systems*, volume 5141 of *LNCS*, pages 426–442, 2008.
18. B. E. Hansen. Lecture Notes on Nonparametrics. Unpublished lecture notes, 2009.
19. P. C. Kocher, J. Jaffe, and B. Jun. Differential power analysis. In *Proceedings of the 19th Annual International Cryptology Conference on Advances in Cryptology*, volume 1666 of *CRYPTO '99*, pages 388–397, London, UK, UK, 1999. Springer-Verlag.
20. T.-H. Le and M. Berthier. Mutual Information Analysis under the View of Higher-Order Statistics. In *IWSEC*, volume 6434 of *LNCS*, pages 285–300. Springer, 2010.
21. Y. Linge, C. Dumas, and S. Lambert-Lacroix. Maximal Information Coefficient Analysis. Cryptology ePrint Archive, Report 2014/012, 2014.
22. V. Lomné, E. Prouff, and T. Roche. Behind the Scene of Side Channel Attacks. In K. Sako and P. Sarkar, editors, *ASIACRYPT*, volume Kazue Sako and Palash Sarkar of *LNCS*, pages 506–525. Springer, 2013.
23. H. Maghrebi, S. Guilley, and J.-L. Danger. Leakage Squeezing Countermeasure against High-Order Attacks. In C. A. Ardagna and J. Zhou, editors, *WISTP*, Lecture Notes in Computer Science, pages 208–223. Springer, 2011.
24. S. Mangard, E. Oswald, and T. Popp. *Power Analysis Attacks: Revealing the Secrets of Smart Cards*, volume 31. Springer Publishing Company, Incorporated, 1st edition, December 2006.
25. T. S. Messerges. *Power Analysis Attacks and Countermeasures for Cryptographic Algorithms*. PhD thesis, University of Illinois, 2000.
26. T. S. Messerges, E. A. Dabbish, R. H. Sloan, T. S. Messerges, E. A. Dabbish, and R. H. Sloan. Investigations of Power Analysis Attacks on Smartcards. In *In USENIX Workshop on Smartcard Technology*, pages 151–162, 1999.
27. A. Moradi, S. Guilley, and A. Heuser. Detecting Hidden Leakages. In *ACNS*, pages 324–342, 2014.
28. A. Moradi, A. Poschmann, S. Ling, C. Paar, and H. Wang. Pushing the Limits: A Very Compact and a Threshold Implementation of AES. In K. G. Paterson, editor, *EUROCRYPT*, volume 6632 of *Lecture Notes in Computer Science*, pages 69–88. Springer, 2011.
29. A. Moradi and A. Wild. Assessment of Hiding the Higher-Order Leakages in Hardware - What Are the Achievements Versus Overheads? In T. Güneysu and H. Handschuh, editors, *Cryptographic Hardware and Embedded Systems*, volume 9293 of *LNCS*, pages 453–474. Springer, 2015.
30. M. Nassar, Y. Souissi, S. Guilley, and J.-L. Danger. RSM: A small and fast countermeasure for AES, secure against 1st and 2nd-order zero-offset SCAs. In *DATE*, pages 1173–1178, 2012.
31. D. Oswald and C. Paar. Improving Side-channel Analysis with Optimal Linear Transforms. In *Proceedings of the 11th International Conference on Smart Card Research and Advanced Applications*, CARDIS'12, pages 219–233, Berlin, Heidelberg, 2013. Springer-Verlag.
32. E. Prouff and R. P. McEvoy. First-Order Side-Channel Attacks on the Permutation Tables Countermeasure - Extended Version -. *IACR Cryptology ePrint Archive*, page 385, 2010.
33. E. Prouff and M. Rivain. Theoretical and Practical Aspects of Mutual Information Based Side Channel Analysis. In *ACNS 2009*, volume 5536 of *LNCS*, pages 499–518, Paris, France, June 2009.
34. E. Prouff and M. Rivain. Theoretical and Practical Aspects of Mutual Information Based Side Channel Analysis. *IJACT*, 2(2):121–138, 2010.
35. O. Reparaz, B. Gierlichs, and I. Verbauwhede. Generic DPA attacks: curse or blessing? In *Lecture Notes in Computer Science*. Springer-Verlag, 2014.
36. W. Schindler, K. Lemke, and C. Paar. A Stochastic Model for Differential Side Channel Cryptanalysis. In *Cryptographic Hardware and Embedded Systems*, volume 3659 of *Lecture Notes Computer Science*, pages 30–46. Springer, 2005.
37. T. Schneider and A. Moradi. Leakage Assessment Methodology - A Clear Roadmap for Side-Channel Evaluations. In T. Güneysu and H. Handschuh, editors, *Cryptographic Hardware and Embedded Systems*, volume 9293 of *LNCS*, pages 495–513. Springer, 2015.

38. T. Schneider, A. Moradi, and T. Güneysu. Robust and One-Pass Parallel Computation of Correlation-Based Attacks at Arbitrary Order. *IACR Cryptology ePrint Archive*, 2015:571, 2015.
39. K. Schramm and C. Paar. Higher Order Masking of the AES. In D. Pointcheval, editor, *CT-RSA*, volume 3860 of *Lecture Notes in Computer Science*. Springer, 2006.
40. D. W. Scott. On Optimal and Data-based Histograms. In *Biometrika*, volume 66, pages 605–610, December 1979.
41. D. W. Scott. *Multivariate Density Estimation: Theory, Practice and Visualization.* Wiley, New York, 1992.
42. B. Silverman. Density Estimation for Statistics and Data Analysis. *London: Chapman & Hall/CRC.*, page 48, 1998.
43. F.-X. Standaert, B. Gierlichs, and I. Verbauwhede. Partition vs Comparison side-channel distinguishers: An empirical evaluation of statistical tests for univariate side-channel attacks against two unprotected CMOS devices. In *Information Security and Cryptology*, volume 5461 of *LNCS*, pages 253–267, Seoul, Korea, December 2008.
44. F.-X. Standaert, T. G. Malkin, and M. Yung. A Unified Framework for the Analysis of Side-Channel Key Recovery Attacks. In *Proceedings of the 28th Annual International Conference on Advances in Cryptology: The Theory and Applications of Cryptographic Techniques*, EUROCRYPT '09, pages 443–461, Berlin, Heidelberg, 2009. Springer-Verlag.
45. F.-X. Standaert and N. Veyrat-Charvillon. Mutual Information Analysis: How, When and Why? In *Cryptographic Hardware and Embedded Systems*, volume 5747 of *LNCS*, pages 429–443, Lausanne, Switzerland, September 2009.
46. S. Tiran, S. Ordas, Y. Teglia, M. Agoyan, and P. Maurine. A model of the leakage in the frequency domain and its application to CPA and DPA. *J. Cryptographic Engineering*, 4(3):197–212, 2014.
47. S. Tiran, G. Reymond, J. Rigaud, D. Aboulkassimi, B. Gierlichs, M. Carbone, G. R. Ducharme, and P. Maurine. Analysis Of Variance and CPA in SCA. *IACR Cryptology ePrint Archive*, 2014:707, 2014.
48. A. Venelli. Efficient Entropy Estimation for Mutual Information Analysis Using B-Splines. In *WISTP*, volume 6033 of *LNCS*, pages 17–30, 2010.
49. N. Veyrat-Charvillon and F.-X. Standaert. Generic side-channel distinguishers: Improvements and limitations. In *CRYPTO 2011*, volume 6841 of *LNCS*, pages 354–372. Cryptology ePrint Archive, Report 2011/149, 2011.
50. C. Whitnall and E. Oswald. A Fair Evaluation Framework for Comparing Side-Channel Distinguishers. *J. Cryptographic Engineering*, 1(2):145–160, 2011.
51. C. Whitnall, E. Oswald, and F.-X. Standaert. The Myth of Generic DPA...and the Magic of Learning. In *CT-RSA*, pages 183–205, 2014.
52. X. Ye and T. Eisenbarth. On the Vulnerability of Low Entropy Masking Schemes. In *CARDIS*, pages 44–60, 2013.