Mutual Information Coefficient Analysis

Yanis Linge[1,2], Cécile Dumas[1], and Sophie Lambert-Lacroix[2]

[1] CEA-LETI/MINATEC, 17 rue des Martyrs,
38054 Grenoble Cedex 9, France
yanis.linge@emse.fr,cecile.dumas@cea.fr
[2] UJF-Grenoble 1 / CNRS / UPMF / TIMC-IMAG
UMR 5525, Grenoble, F-38041, France
Sophie.Lambert@imag.fr

Abstract In the domain of the Side Channel Attacks, various statistical tools have succeeded to retrieve a secret key, as the Pearson coefficient or the Mutual Information. In this paper we propose to study the Maximal Information Coefficient (MIC) which is a non-parametric method introduced by Reshef *et al.* [13] to compare two random variables. The MIC is based on the mutual information but it is easier to implement and is robust to the noise. We show how apply this tool in the particular case of the side channel attacks. As in statistics, benefits only appears with drawbacks, the computing complexity of the MIC is high. Therefore, we propose a way to efficiently compute the MIC. The obtained attack called the Maximal Information Coefficient Analysis is compared to the CPA [3] and the MIA [8]. The results show the interest of this approach when the leakage is noisy and bad modeleled.

1 Introduction

In order to describe the Side Channel Attacks (SCA) by means of the information theory, Gierlichs *et al.* presented the Mutual Information Analysis (MIA) in 2008. This attack takes advantage of mutual information (MI) combining the information contained in a signal and the information provided by a model. In the context of SCA, the signal is obtained by measuring the power consumption or electromagnetic radiation of an electronic device during the execution of a cryptographic algorithm. Since Kocher [9] it is well known that the form of these signals is partly related to the data handled by the component. For this reason the correlation power analysis (CPA) [3], which is based on the linear correlation, is a very efficient attack when implementations are not protected. Although the underlying model is very simple, as it is generally assumed that the signal amplitude is proportional to a combination of computed bits, the success of the CPA shows that it is often sufficient. Yet the leakage can be difficult to model, because the component or the implementation includes some countermeasures. In this case, a too simplistic model would no longer linearly related to the acquired signals. The MIA, that uses no assumption about the linear nature, seems therefore more suitable for processing signals whose leakage is roughly modeled. But when trying to implement this attack several problems

appear. First the MI computation needs to choose an estimator from a set of existing methods. Secondly, it is difficult to interpret the computed value for MI and compare different obtained values. For these reasons, Reshef *et al.* [13] have recently developed a new statistical tool, called Maximal Information Coefficient (MIC), which has the same advantages as the MI, since it considers any kind of relation, even non-linear, but has an exact definition, a soft interpretation and a better stability. In this paper, we present the work for adapting this new tool in the context of SCA and the experiments that confirm the interest of this attack, that we consequently call Maximal Information Coefficient Analysis (MICA). This paper is organized as follows. First we shortly remind the CPA and the MIA. Then, we present the MIC and its application to SCA in section 3. Before concluding, we expose some experimentations in sections 4 to compare the last attack to the two first ones.

2 Side Channel Attack

The Side Channel Attack or SCA targets a cryptographic algorithm, implemented on a device, which requires huge keys in order to prevent the brute force attack. But during the execution of this algorithm, the key is used piecewise. Since these parts of the key are small, we can enumerate all possible values and apply a brute force attack to determine the value of the targeted part. In classical cryptanalysis, we do not have access to the output of a cryptographic function before all the parts of the key are used. The SCA provides some information about the output of the function when only a part of the key is involved. Unfortunately for attackers, the acquired information is partial. Generally, attackers obtain information about the Hamming weight of the output.

We will now present a general model of SCA before introducing two common SCA: the Correlation Power Analysis and the Mutual Information Analysis.

2.1 General model for the SCA

Let K a random variable which represents a part of the secret key. We note $k^\star$ the right key value. X represents the input of the targeted algorithm. Let Z a random variable which represents the value of an intermediate state of the targeted algorithm. Finally, we denote the random variable representing the leakage by $L(Z) = f(X, K)$. The function f highly depends on the device. Notice that $L(Z)$ is a continuous variable.

To perform a SCA, we have n measurements of the leakage, $l_i = f(x_i, k^\star)$ with $i = 1, ..., n$. We suppose that these leakages give an independent random variable. In another hand, the leakage for a given key k, is modeled by a function $M(X, k)$. The variable $M(X, k)$ is discret. So we have n realizations $M(x_i, k)$ for each key in the set of all possible keys $\mathcal{K}$. The SCA uses these data to mesure the possible relation between the leakage and the model for each key. We consider that the SCA is successful if

$$\max_{k \in \mathcal{K}}(\mid D(M(x, k), l) \mid) = k^\star$$

where D is a distinguisher. Generally, the main difference between the attack methods is the choice of the distinguisher.

Now we present two SCA based on two different distinguishers.

2.2 Correlation Power Analysis or CPA

When we want to examine the relation between two variables, the first idea is to use the linear correlation coefficient(ρ). In our case, ρ is given by :

$$\rho_k = \frac{\frac{1}{n}\sum_{i=1}(l_i - \bar{l})(M(x_i,k) - \bar{M_k})}{\sqrt{\frac{1}{n}\sum_{i=1}(l_i - \bar{l})^2 \frac{1}{n}\sum_{i=1}(M(x_i,k) - \bar{M_k})^2}},$$

where $\bar{l}$ and $\bar{M_k}$ are arithmetic means. The more ρ_k is close to ± 1, the strongest is the linear relation between the leakage and the model. ρ_k^2 is a measure of the distance between the n points to the linear regression.

The Correlation Power Analysis has been introduced by Brier *et al.* in [3]. In this paper, the authors propose to use the Pearson correlation coefficient as a distinguisher. Nowadays the CPA is the most widely used approch to perform a SCA. Generally, the CPA works well because the relation between l and $M(x, k^\star)$ is close to the linear. On some particular devices, this relationship is not linear. We will now present a distinguisher allowing to detect both linear and nonlinear relations.

2.3 Mutual Information Analysis or MIA

The MIA is based on the mutual information or MI. The MI provides a way to estimate a relationship between two random variables even if the relation is not linear. Before describing the MIA, we need some tools of the information theory.

Preliminaries The Shannon entropy is a way to mesure the uncertainty associated to a random variable. Let A be a random variable in the space $\mathcal{A}$ and B a random variable in the space $\mathcal{B}$. For lighten notations, we consider that A and B are discrete variables and define by $Pr[E]$ the probability of the event E. For continuous variables, formulas are the same but the sum is replaced by an integral.

The entropy associated to A is defined by :

$$H(A) = -\sum_{a \in \mathcal{A}} Pr[A = a] \cdot log_2(Pr[A = a])$$

We can express the join entropy of a pair of random variables (A, B) in a similar way by :

$$H(A,B) = -\sum_{a \in \mathcal{A}, \in \mathcal{B}} Pr[A = a, B = b] \cdot log_2(Pr[A = a, B = b])$$

After all the conditional entropy of A given B by :

$$H(A|B) = -\sum_{a\in\mathcal{A},\in\mathcal{B}} Pr[A=a, B=b] \cdot log_2(Pr[A=a|B=b])$$

The mutual information mesures the mutual dependence of two random variables. It is defined by :

$$\begin{aligned} I(A,B) &= H(A) - H(A|B) \\ &= H(A) + H(B) - H(A,B) \end{aligned}$$

The maximal value of the MI is $H(A) + H(B)$. This measure does not depend on the form of the relation between the two variables. So we can detect if there exists a relation linking the variables even if this relation is not linear. If the MI is zero, A and B are independent. On the other hand, if the value of the MI is maximal, it exists a strong link between A and B. But the MI is not valuable alone. If we have a value $I(A,B) = 8$ nothing can be deduced on A and B, while if $\rho = 0.8$ we can deduce directly that A and B are linked.

MIA In [8], Gierlichs *et al.* suggest to use the Mutual information as a new distinguisher. With our previous notation, we defined the MI by :

$$I(l, M(x,k)) = H(l) + H(l|M(x,k))$$

The values of the MI obtained for different $k \in \mathcal{K}$ are comparable because L is constant. Since L is fixed, the difference between the values of the MI is only due to the entropy $H(l|M(x,k))$. Smaller is this entropy, greater is the link between L and $M(x,k)$. This distinguisher allows the detection of connection between the leakage and the model even if the relation is non-linear. The different studies and experimentations regarding the MIA[17] show most of the time that if the mutual information is used as distinguisher, the success of the attack requires more traces than using the linear correlation. When some particular devices are attacked or in some noisy environment, the use of the mutual information can be more effective than the CPA. The most important drawback of the MIA is that computing the mutual information can be tricky. In practice, we compute marginal entropies $H(l)$ and $H(M(x,k))$ and the join entropy $H(l, M(x,k))$. Since the variable $M(x,k)$ is discrete, the probability $Pr[M_k = m]$ is estimated by using the empirical frequency. The computation of the MI depends on the base used to determinate the probability density function (or *pdf*) of the leakage. The most common way to estimate the *pdf* of the leakage is by using bins, but there exists a lot of different estimators [17]. The mutual information provides an interesting distinguisher. Unfortunately, this distinguisher is not stable if the number of samples increases (or decreases). Moreover, the estimation of the *pdf* of the leakage has a great influence on the result of the MIA.

We will now introduce a new distinguisher based on the mutual information. This distinguisher overcomes the deficiencies of the MI while maintaining its generality.

3 Maximal Information Coefficient Analysis or MICA

3.1 Maximal Information Coefficient or MIC

In 2011, Reshef *et al.* [13] introduced a new measure of the independance of two variables: the Maximal Information Coefficient (MIC). The MIC is based on the mutual information (cf. 2.3). Computing the mutual information is tricky, when a continuous variable is involved. The authors propose to estimate the *pdf* of variables by using bins. Since there are many ways to choose the bins, Reshef *et al.* compute the maximal MI over all possible choices of bins. The main idea of the MIC is that if a relation exists between our two variables, there exists a partition of the data that will allow to include this relationship.

Compute the MIC The goal of the MIC is to detect if there exists relationship between two random variables A and B. Let define by D the couple (A, B) after ordering. Reshef *et al.* call *p-by-q* grid the partition of the couple (A, B) in p bins for the variable A and q bins for the variable B. There are a lot of different grids of size *p-by-q*. $\tilde{D}^G$ is the frequency distribution engendered by the couple (A, B) on the cell of the grid G.
We note by $\tilde{A}^G$ and $\tilde{B}^G$ the distribution of A and B over the grid G and $\mathcal{G}_{(p,q)}$ the set of all grids of size *p-by-q*.

As an example, let $D = ((0,1),(1,1),(2,1),(3,-1),(4,-1),(5,1))$ and $G = ([0,2[,[2,4[,[4,6[) \times ([-2,0[,[0,2[)$ a 3-*by*-2 grid.
So over G, D is given by Tab. 3.1 and $\tilde{A}^G = (2,2,2)$ and $\tilde{B}^G = (2,4)$.

	[-2,0[	[0,2[
[0,2[	0	2
[2,4[	1	1
[4,6[	1	1

For fixed p and q, the maximal mutual information over all grids *p-by-q* is defined by:

$$I^\star(D,p,q) = \max_{G \in \mathcal{G}_{(p,q)}} (I(\tilde{A}^G, \tilde{B}^G))$$

Since it is not possible to compare two maximal MI, $I^\star(D,p,q)$ and $I^\star(D,p',q')$, if the sizes of the grid are not the same, Reshef *et al.* propose a normalization. $I^\star(D,p,q)$ is bounded by $log_2(\min_{p,q})$. And the equality between $I^\star(D,p,q)$ and $log_2(\min_{p,q})$ is obtained if A and B are linked by a function. A natural normalization is given by :

$$M(D)_{p,q} = \frac{I^\star(D,p,q)}{log_2(\min_{p,q})}$$

Normalize the MI allows a comparaison between $I^\star$ for different grid sizes and a computation of the maximum for all possible p and q. As p and q is bounded by

n, the number of possible grids is bounded by n^n which is huge. So this maximal size of the grid has been reduced to $(n^{0.6})$ by Reshef *et al.*, thank to an empirical test. Finally, the Maximum Information Coefficient is:

$$MIC(D) = \max_{\forall p,q \ p \cdot q \leq n^{0.6}} (M(D)_{p,q})$$

with n the number of elements of A and B.

The MIC allows to identify a large type of relations. Moreover, the MIC is designed to maintain good results even in presence of noise. The MIC seems to be a good candidate for a generic distinguisher. Unfortunately, the computation of the MIC has a high time complexity. In the next section, we propose to use the specificity of our data to compute the MIC in order to use it as a distinguisher.

3.2 MICA

The MICA, Mutual Information Coefficient Analysis is naturaly defined by:

$$\max_{k \in \mathcal{K}} (MIC(M(x,k),l)) = k^\star$$

To simplify the computation complexity of the MIC for each key hypothesis, it is crucial to take into account of the data particularities.

First, it is important to remind that we study two different types of variables: L is continuous and $M(X,k)$ is discret. We know, *a priori*, the number of all possible values of the model. The idea is to bound q by this number, which generally is small. For example, the Hamming weight of a state byte of the AES is modeled by 9 values and the DES by 5 values.

Moreover, in [13] Reshef *et al.* propose an heuristic algorithm to compute $I^\star(D,p,q)$ for fixed p and q. They fix one partition of B of size q and they compute the maximum value of $I(A,B)$ over all the grids by varying the partition of A. Since the marginal entropy of A is maximal for the equipartition, the maximum value of $I(A,B)$ is given by the equipartition of A. This heuristic should be applied for the MICA, we choose to fix the bounded partition of $M(X,k)$.

In this case, it is easy to compute the maximal value of $I(L, M(X,k))$ when the partition of $M(X,k)$ is fixed regardless of the value of q and the partition of L is an equipartition of size p. The number of grids can be abruptly bounded by the product of the possible partitions for $M(X,k)$ and the maximal size of the grid$(n^{0.6})$. So if we have $n = 1,000,000$ samples of an AES, we need to explore less than $2,000,000$ grids which is feasible. Using these observations, we made an implementation in openCL to compute the MIC in our particular case. We compare our implementation with a C implementation, *minepy* [2] and the java implementation delivered by Reshef *et al.*. The implementation in java works only with a few number of samples (less than $1,000$). In Tab. 1, we present the needed time to compute the MIC using the three implementations.

Since our implementation uses the characteristics of the data handled in the SCA, it is more efficient. Moreover, we take part of the natural parallelism in the

Num. samples	java	minepy	openCL
1000	2.6	0.03	0.3
20,000	×	2	0.5

Table 1: Computation for the different implementations in seconds.

algorithm that calculates the MIC to increase the efficiency of the implementation.

In the next section, we present the result of our experimentations on the DPAContest v1 samples [6].

4 Experimentations

We made experiments on the public traces available in the DPAContest. The first version of the DPAContest delivers samples obtained by recording the power consumption of a DES [20]. First we studied the success rate of the MICA before comparing the results of the CPA and MICA.

4.1 Comparaison of the success rate for MICA, MIA and CPA

In the Fig. 1, we add to the Fig. 3 of [17], the success rate of the MICA for the DPAContest V1.

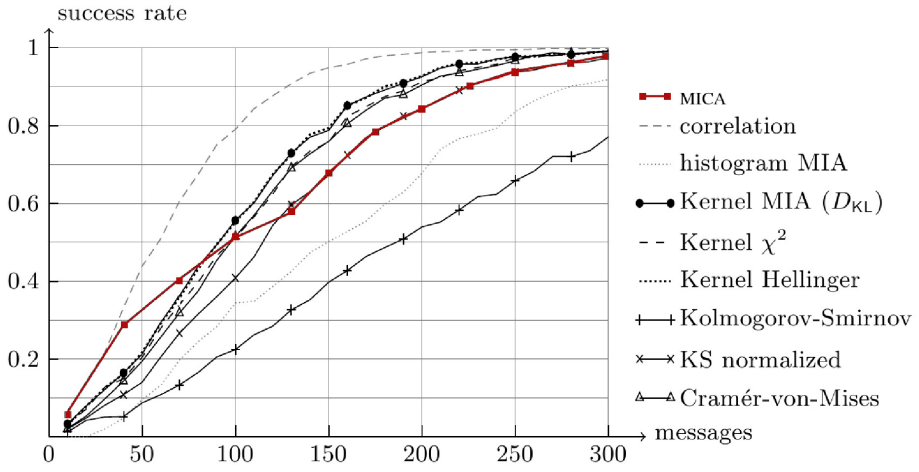

Figure 1: Success rate of different distinguishers

These experimentations are made on the first SBOX of the last round of the DES over 1000 independent experiments using a Hamming weight leakage

model. We can observe that, the CPA is still the most efficient distinguisher when the leakage model is well designed. The MICA is more efficient than the MIA when we have only few samples. However, the MICA is more effective than the histogram MIA. The main drawback of the MIA is the computation of the marginal and the conditional entropy. There exists a lot of different estimators, but none are "ideal". The result and the complexity of the MIA highly depends on the used estimator. As for the MIC provides an "out of the box" tool for the SCA, because the data does not need to be studied before using the MIC. Moreover, the impact of the noise on the MIA is not easy to study [18], whereas in [13], Reshef *et al.* showed a lot of examples of the robustness of the MIC in noisy environments.

For the rest of our experimentations, we compared only the CPA with the MICA. The success rate is not the only criterion to evaluate a distinguisher. In the next sections, we will more investigate the results of the MICA and compare them.

4.2 DPAContest V1

In Fig. 2 and 3, as previously, we targeted the first SBOX of the last round of the DES. For the Fig. 2 the number of samples exceeds the minimal number needed to have a success rate of 1. For the Fig. 3 the number of samples ensures that the success rate of the CPA and the MICA is greater than 0.9. Each curve represents the value of the distinguisher during the running time for different keys. The good key is the one with the higher value. We can note that $M(X, k^\star)$ and L are related in several instants of the execution.

On the top of the Fig. 2, we can note that the sign of the Pearson correlation coefficient changes. When comparing the two graphics, we note that the relationship between the leakage and the model of the right key appears at the same time. But the difference between the curve corresponding to the right key and the second highest is more important for the MICA. Moreover, the curves seems to be less noisy in the case of the MICA. In the Fig. 3 we show that the MICA and CPA are similar behavior when the number of samples becomes critical. In DPAContest V1, the relation between the Hamming weight of the handeled data and the leakage is close to the linear, if this relation was nonlinear, we can expec better result for the MICA.

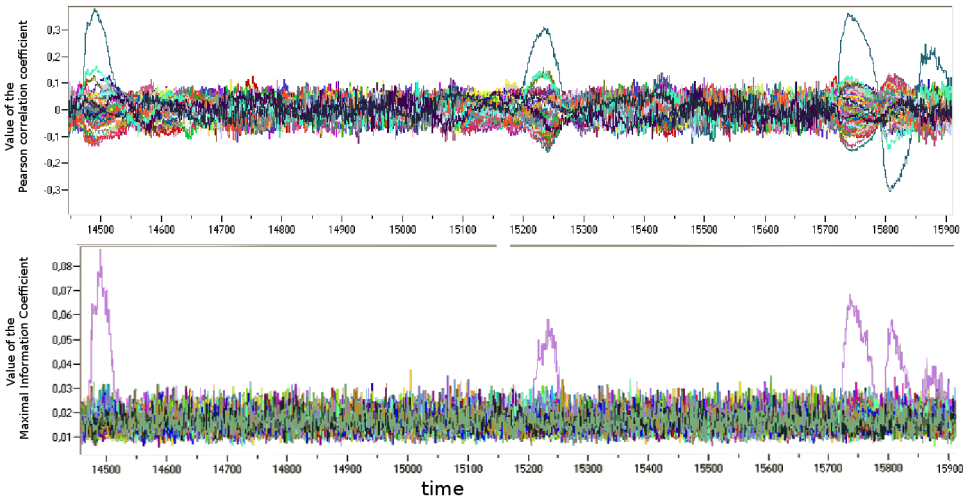

Figure 2: On the top the results of the CPA and on the bottom the results of the MICA, for 1,000 samples of DPAContestV1.

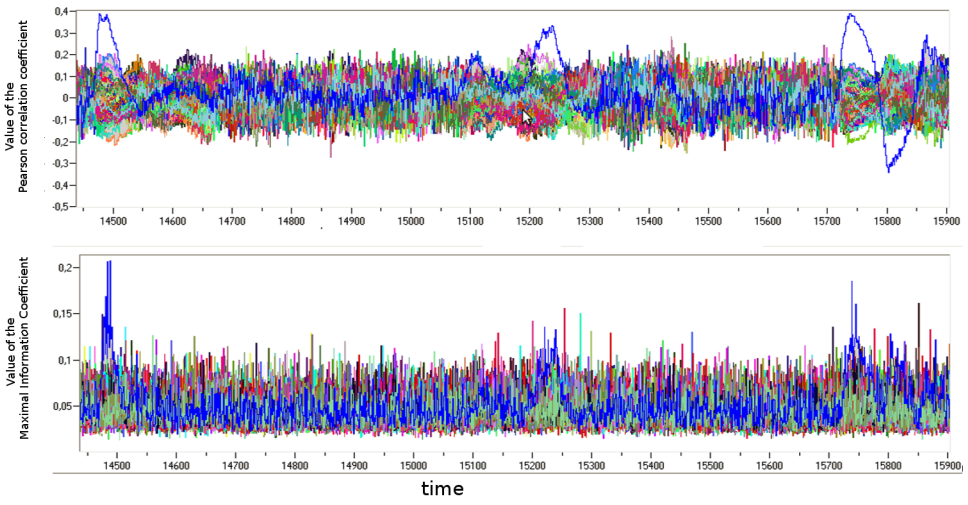

Figure 3: On the top the results of the CPA and on the bottom the results of the MICA, for 200 samples of DPAContestV1.

4.3 Other leakage models

Since the MIC provides a distinguisher more general than the linear correlation, it is interesting to study the MICA behaviour when the leakage is not well modeled. We suppose that the attacker uses a classical Hamming weight model while the leakage is a little different. We compare the three attacks based on the Pearson correlation coefficient, the MI and the MIC in simulation. For simulating the handling data, we pick 50 numbers x in $\{0, \cdots, 15\}$ and compute a simulated leakage, $f(x)$. We considered three different models for the leakage(f).

1. $f(x) =$ P(Hamming weight(x)) where P is a fixed permutation
2. $f(x) =$ Hamming weight of x
3. $f(x) =$ Hamming weight of x + the value of the first bit of x

In the Tab.2, we present the result of the three attacks for the simulated data.

leakage model	ρ	MI	MIC
1	0.14	3.88	0.61
2	0.81	2.03	0.77
3	0.70	2.58	1.0

Table 2: Values of ρ, MI, MIC for 50 simulated data for different leakage models.

As expected the CPA is the best method when the model is well defined. When the leakage model is completely different, as in the case 1, the MI distinguisher seems the best choice. But the MIC value is still good and it is not easy to compare the MI to the other coefficients. The MIC has great results when all the bits of the targeted variable have not the same contribution to the leakage. In this case, the value of the MIC is maximal because the leakage have more variations. If the model used in the attack is imperfect compared to the real leakage, the MIC assimilates the differences.

5 Conclusion

In this paper, we constructed a generic side channel distinguisher based on the Maximal Information Coefficient. This distinguisher, like the MIA, does not require a linear link between the leakage and the model. Like the CPA, it is a tool that can be easily used in the context of side channel. Taking into account of the data specifities we improve the computation complexity. Our OpenCL implementation is more efficient than the minepy implementation.

Although the MIC can only be applied into the univariate case, it presents main advantages, over the MI: it is clearly defined and seems get better results in noisy environments.

In our experiments, we observed that the proposed distinguisher is efficient to retrieve the key with few samples when the leakage model is well defined. Moreover we obtain great results when one bit leaks more than the others. In these cases and more generally when the leakage is bad modeled, the MIC could more easily detect a relationship than the MI or the Pearson correlation coefficient.

References

1. D. Agrawal, B. Archambeault, J.R. Rao and P. Rohatgi *The EM side-channel(s).* CHES 2002, LNCS, vol 2523, pp 24-45, 2002.
2. D. Albanese, M. Filosi, R. Visintainer, S. Riccadonna, G. Jurman, and C. Furlanello. *cmine, minerva & minepy: a C engine for the MINE suite and its R and Python wrappers.* arXiv:1208.4271[stat.ML], 2012.
3. E. Brier, C. Clavier and F.Olivier *Correlation power analysis with a leakage model.* CHES 2004, LNCS, vol 3156, pp 16 - 29, 2004.
4. S. Chari, J. Rao, P.Rohatgi. *Template Attack.* CHES 2002, LNCS, vol 2523, pp 13-28, 2002.
5. J. Daemen and V. Rijmen. *AES proposal: Rijndael*, 1998.
6. DPA Contest 2008/2009, http://www.dpacontest.org/
7. K. Gandolfi, C. Mourtel and F. Olivier. *Electromagnetic analysis: Concrete results.* CHES 2001, LNCS, vol 2162, pp 251-261, 2001.
8. B. Gierlichs, L. Batina, P. Tuyls and B. Preneel. *Mutual Information Analysis - A Generic Side-Channel Distinguisher.* CHES 2008, LNCS, vol 5154, pp 426-442, 2008.
9. P.C. Kocher, J. Jaffe and B. Jun. *Differential power analysis.* CRYPTO, pp 388-397, 1999.
10. T.-H. Le, J. Clédière, C. Servière, and J.-L. Lacoume. *Noise reduction in side channel attack using fourth-order cumulant.* IEEE Transactions on Information Forensics and Security, vol 2, no 4, pp 710–720, 2007.
11. http://www.khronos.org/opencl/
12. J-J. Quisquater and D. Samyde. *Electromagnetic analysis (ema): Measures and counter-measures for smart cards.* E-smart 2001, LNCS, vol 2140, pp 200–210, 2001.
13. D. Reshef, Y. Reshef, H. Finucane, S. Grossman, G. McVean, P. Turnbaugh, E. Lander, M. Mitzenmacher and P. Sabeti. *Detecting novel associations in large datasets.* Science, **6062**(334), pp.1518-1524, 2011.
14. M. Rivain. *On the Physical Security of Cryptographic Implementations.* PhD thesis, University of Luxembourg, 2009.
15. F.-X. Standaert *Partition vs. Comparison Side-Channel Distinguishers* http://www.dice.ucl.ac.be/fstandae/tsca/
16. F.-X. Standart, P. Bulens, G. de Meulenaer and N. Veyrat-Charvillon. *Improving the Rules of the DPA Contest.* Cryptology ePrint Archive, Report 2006/139, http://eprint.iacr.org/2006/139.
17. N. Veyrat-Charvillon, F.-X. Standart. *Mutual Information Analysis : How, When and Why?* CHES 2009, LNCS, vol 5747, pp. 429-443, 2009.
18. C. Whitnall, E. Oswald. *A Fair Evaluation Framework for Comparing Side-Channel Distinguishers.* Journal of Cryptographic Engineering, 1(2):145-160, August 2011.

19. C. Whitnall, E. Oswald and L. Mather. *An Exploration of the Kolmogorov-Smirnov Test as Competitor to Mutual Information Analysis*. Cryptology ePrint Archive, Report 2011/380, http://eprint.iacr.org/2011/380.
20. Federal Information Processing. *Data Encryption Standard. Standards Publication 46-1 National Technical Information Service, U.S. Dept. of Commerce, 1977.*