A Stochastic Model for Differential Side Channel Cryptanalysis

Werner Schindler[1], Kerstin Lemke[2,⋆], and Christof Paar[2]

[1] Bundesamt für Sicherheit in der Informationstechnik (BSI),
Godesberger Allee 185-189, 53175 Bonn, Germany
Werner.Schindler@bsi.bund.de
[2] Horst Görtz Institute for IT Security,
Ruhr University Bochum, 44780 Bochum, Germany
{lemke, cpaar}@crypto.rub.de

Abstract. This contribution presents a new approach to optimize the efficiency of differential side channel cryptanalysis against block ciphers by advanced stochastic methods. We approximate the real leakage function within a suitable vector subspace. Under appropriate conditions profiling requires only one test key. For the key extraction we present a 'minimum principle' that solely uses deterministic data dependencies and the 'maximum likelihood principle' that additionally incorporates the characterization of the noise revealed during profiling. The theoretical predictions are accompanied and confirmed by experiments. We demonstrate that the adaptation of probability densities is clearly advantageous regarding the correlation method, especially, if multiple leakage signals at different times can be jointly evaluated. Though our efficiency at key extraction is limited by template attacks profiling is much more efficient which is highly relevant if the designer of a cryptosystem is bounded by the number of measurements in the profiling step.

Keywords: Differential Side Channel Cryptanalysis, Stochastic Model, Minimum Principle, Maximum Likelihood Principle, Power Analysis, DPA, Electromagnetic Analysis, DEMA, Template Attack.

1 Introduction

Side channel cryptanalysis exploits physical information that is leaked during the computation of a cryptographic device. The most powerful leakage consists of instantaneous physical signals which are direct responses on the internal processing. These instantaneous observables can be obtained by measuring the power dissipation or the electromagnetic emanation of the cryptographic device as a function of time. Power analysis, which was first introduced in [9] and electromagnetic analysis ([8]) are based on the dependency of the side channel information on the value of intermediate data, which is in turn caused by the physical implementation.

⋆ Supported by the European Commission through the IST Programme under Contract IST-2002-507932 ECRYPT, the European Network of Excellence in Cryptology.

J.R. Rao and B. Sunar (Eds.): CHES 2005, LNCS 3659, pp. 30–46, 2005.

Advanced stochastic methods have turned out to be efficient tools to optimize pure timing and combined timing and power attacks. Using such methods, the efficiency of some known attacks could be increased considerably (up to a factor of fifty), some attacks could be generalized and new attacks were conceived ([12,13,14]). The understanding of the source of an attack and its true risk potential is important for a designer of a cryptographic system for implementing effective and reliable countermeasures that prevent also privileged attacks.

This contribution gives a thorough stochastic approach to optimize the efficiency of differential side channel analysis applied against block ciphers. In our work, the quantification of side channel leakage is done in a chosen vector subspace. Under suitable conditions it requires only measurements under one test key, and even this test key need not be known. Our approach aims to achieve the efficiency of the template attacks in the key extraction phase but requires far less measurements in the profiling phase, e.g., in case of AES we guess that savings in the order of up to one hundred are feasible. This is surely interesting for designers of cryptosystems in order to assess the susceptibility of their implementations towards attacks. The mathematical model is supported by an experimental analysis of an AES implementation on an 8-bit microcontroller. Further, we show how our model can be generalized to comprehend both masking countermeasures as well as the usage of multiple physical channels.

1.1 Related Work

Differential side channel cryptanalysis identifies the correct key value by statistical methods for hypothesis testing. Differential Power Analysis (DPA) ([9]) turned out to be a very powerful technique against unknown implementations. The single measurements are partitioned accordingly to the result of a selection function that depends both on known data and on key hypotheses. [9] suggested to just use the difference of means for the two sets of single measurements. Improved statistics are the student's T-Test and the correlation method which are given in [2]. Additional guidelines for testing the susceptibility of an implementation are presented in [3].

Other contributions assume that the adversary is more powerful, e.g, that the adversary is able to load key data into the cryptographic device. Profiling as a preparation step of power analysis was first described by [6]. Probably the most sophisticated strategy is a template based attack ([4]) which aims to optimize Simple Power Analysis (SPA) and requires a precise characterization of the noise. Moreover, physical information can be captured simultaneously by different measurement set-ups, e.g., by measuring the EM emanation and the power consumption in parallel ([1]).

2 The Mathematical Model

In this section we introduce a new mathematical model for differential side channel attacks against block ciphers. We investigate this model (Subsect. 2.1)and

exploit these insights to derive optimal decision strategies (Subsects. 2.2 and 2.3). The success probability (or equivalently, the risk potential) and the efficiency of our approach are considered.

We assume that the adversary (e.g., the designer) measures physical observables at time t in order to guess a subkey $k \in \{0,1\}^s$. The letter $x \in \{0,1\}^p$ denotes a known part of the plaintext or the ciphertext, respectively. We view a measurement at time t as a realization of the random variable

$$I_t(x,k) = h_t(x,k) + R_t. \tag{1}$$

The first summand $h_t(x,k)$ quantifies the deterministic part of the measurement as far it depends on x and k. The term R_t denotes a random variable that does not depend on x and k. Without loss of generality we may assume that $E(R_t) = 0$ since otherwise we could replace $h_t(x,k)$ and R_t by $h_t(x,k) + E(R_t)$ and $R_t - E(R_t)$, respectively. We point out that (1) does not cover masking techniques. A generalization of (1) and the main results in Subsects. 2.2 and 2.3, however, is straight-forward (cf. Subsect. 2.4). From now on we assume that the plaintext is known by the adversary but our results can be directly transferred to 'known-ciphertext' attacks.

Example 1. In Sect. 3 an AES implementation targeting one S-Box is analyzed. Then t is an instant, e.g., during the first round and $x, k \in \{0,1\}^8$.

2.1 Fundamental Theorems

The central goal of Subsect. 2.2 is to estimate the distribution of the random vector $(I_{t_1}(x,k), \ldots, I_{t_m}(x,k))$ where $t_1 < \cdots < t_m$ are different instants that are part of the side-channel measurements We work out important facts that will be used in the next subsection.

Definition 1. *As usual* $\|\cdot\|\colon \mathbb{R}^n \to \mathbb{R}$ *denotes the Euclidean norm, that is* $\|(z_1, z_2, ..., z_n)\|^2 = \sum_{j=1}^{n} z_j^2$. *In this work, terms* $\boldsymbol{b}^T$ *and* A^T *stand for the transpose of the vector* $\boldsymbol{b}$ *and the matrix* A*, respectively. The term* $\widetilde{f}$ *denotes an estimator of a value* f*. Random variables are denoted with capital letters while their realizations, i.e. values assumed by these random variables, are denoted with the respective small letters.*

Mathematical Model. The random variables R_t, X and K (resp. R_t, $X_1, X_2, \ldots, X_N$, and K) are defined over the same probability space $(W, \mathcal{W}, P)$, where W is a sample space, $\mathcal{W}$ a σ-algebra consisting of subsets of W and P a probability measure on $\mathcal{W}$. More precisely, $R_t\colon W \to \mathbb{R}$; $X, X_1, \ldots, X_N\colon W \to \{0,1\}^p$ (random plaintext parts) and $K\colon W \to \{0,1\}^s$ (random subkey). By assumption, the random variables R_t, X and K (resp. R_t, $X_1, X_2, \ldots, X_N$, and K) are independent. For the sake of readability in (2), for instance, we suppress the subscript $_{X,R_t,K=k}$ as this should be obvious.

Theorem 1. *Let $k \in \{0,1\}^s$ denote the correct subkey. Then the following assertions are valid:*

(i) The minimum

$$\min_{h'\colon \{0,1\}^p \times \{0,1\}^s \to \mathbb{R}} E\left(\left(I_t(X,k) - h'(X,k)\right)^2\right) \tag{2}$$

is attained at $h' = h_t$. If $\mathrm{Prob}(X = x) > 0$ for all $x \in \{0,1\}^p$ (e.g., if X is equidistributed on $\{0,1\}^p$) the minimum is exclusively attained for $h' = h_t$.
(ii) Let $t_1 < t_2 \cdots < t_m$. Then the minimum

$$\min_{h'_1,\ldots,h'_m\colon \{0,1\}^p \times \{0,1\}^s \to \mathbb{R}} E\left(\| \left(I_{t_1}(X,k) - h'_1(X,k), \ldots, I_{t_m}(X,k) - h'_m(X,k)\right) \|^2\right) \tag{3}$$

is attained at $(h'_1, \ldots, h'_m) = (h_{t_1}, \ldots, h_{t_m})$.
(iii) For each $x \in \{0,1\}^p$ we have $h_t(x,k) = E_{X=x}\left(I_t(X,k)\right)$.

Proof. Clearly, $I_t(X,k) - h'(X,k) = \Delta h(X,k) + R_t$ with $\Delta h = h_t - h'$. Squaring both sides and evaluating their expectations yields

$$E\left(\left(I_t(X,k) - h'(X,k)\right)^2\right) = E\left(\Delta h(X,k)^2\right) + E\left(R_t^2\right) \geq E\left(R_t^2\right)$$

since $E(R_t) = 0$, and since $\Delta h_t(X,k)$ and R_t are independent by assumption. If $\mathrm{Prob}(X = x) > 0$ for all $x \in \{0,1\}^p$ then $E(\Delta h(X,k)^2) > 0$ for $h' \neq h_t$ which completes the proof of (i). Similarly,

$$\begin{aligned} &E\left(\| \left(I_{t_1}(X,k) - h'_1(X,k), \ldots, I_{t_m}(X,k) - h'_m(X,k)\right) \|^2\right) \\ &= \sum_{j=1}^{m} E\left(\left(\Delta h(X,k) + R_{t_j}\right)^2\right) \geq \sum_{j=1}^{m} E\left(R_{t_j}^2\right), \end{aligned}$$

which verifies (ii), while (iii) follows immediately from (1).

Note that Theorem 1 (ii) says that we may determine the unknown functions $h_{t_1}, \ldots, h_{t_m}$ separately although we are interested in the joint distribution of $(I_{t_1}(X,k), \ldots, I_{t_m}(X,k))$. Principally, the 2^{p+s} unknown function values $h_t(x,k)$ could be estimated separately using Theorem 1(iii). Though satisfactory from a theoretical point of view this approach is impractical.

Considering the concrete implementation a designer (resp., an adversary) should be able to determine a (small) subset $\mathcal{F}_t \subset \mathcal{F} := \{h'\colon \{0,1\}^p \times \{0,1\}^s \to \mathbb{R}\}$ that either contains the searched function h_t itself or at least a function h_t^* that is sufficiently 'close' (to be made precise below) to h_t. For simplicity we restrict our attention to the case $\mathcal{F}_t = \mathcal{F}_{u;t}$, where this set of functions is a real vector subspace that is spanned by u known functions $g_{jt}\colon \{0,1\}^p \times \{0,1\}^s \to \mathbb{R}$. More precisely,

$$\mathcal{F}_{u;t} := \{h'\colon \{0,1\}^p \times \{0,1\}^s \to \mathbb{R} \mid \sum_{\mathrm{j}=0}^{\mathrm{u}-1} \beta'_{\mathrm{j}} \mathrm{g}_{\mathrm{jt}} \text{ with } \beta'_{\mathrm{j}} \in \mathbb{R}\} \tag{4}$$

We may assume that the functions g_{jt} are linearly independent so that $\mathcal{F}_{u;t}$ is isomorphic to $\mathbb{R}^u$. In particular, the minimum on the right-hand side of (6) always exists. Theorem 2 will turn out to be crucial for the following. In the following h_t^* will always denote an element in $\mathcal{F}_{u;t}$ where (6) and (7) attain their minimum.

Theorem 2. *As in Theorem 1 let $k \in \{0,1\}^s$ denote the correct subkey.*
(i) For each $h' \in \mathcal{F}_{u;t}$ we have

$$\begin{aligned} &E\left(\left(I_t(X,k) - h'(X,k)\right)^2\right) - E\left(\left(I_t(X,k) - h_t(X,k)\right)^2\right) \\ &= E_X\left(\left(h_t(X,k) - h'(X,k)\right)^2\right) \geq 0 \end{aligned} \tag{5}$$

where $E_X(\cdot)$ denotes the expectation with respect to the random variable X, i.e. the right-hand term equals $\sum_{x\in\{0,1\}^p} \mathrm{Prob}(X = x)\left(h_t(x,k) - h'(x,k)\right)^2$.

$$(ii) \quad E_X\left(\left(h_t(X,k) - h_t^*(X,k)\right)^2\right) = \min_{h'\in\mathcal{F}_{u;t}} E_X\left(\left(h_t(X,k) - h'(X,k)\right)^2\right) \tag{6}$$

implies

$$E\left(\left(I_t(X,k) - h_t^*(X,k)\right)^2\right) = \min_{h'\in\mathcal{F}_{u;t}} E\left(\left(I_t(X,k) - h'(X,k)\right)^2\right). \tag{7}$$

(iii) Let $t_1 < t_2 \cdots < t_m$. If $h'_j \in \mathcal{F}_{t_j}$ for all $j \leq m$ then

$$\begin{aligned} &E\left(\| \left(I_{t_1}(X,k) - h'_1(X,k), \ldots, I_{t_m}(X,k) - h'_m(X,k)\right) \|^2\right) \\ &= E\left(\| \left(I_{t_1}(X,k) - h_{t_1}(X,k), \ldots, I_{t_m}(X,k) - h_{t_m}(X,k)\right) \|^2\right) + \\ &\sum_{j=1}^{m} E_X\left(\left(h_{t_j}(X,k) - h'_j(X,k)\right)^2\right). \end{aligned} \tag{8}$$

Proof. Assertion (i) can be shown similarly as Theorem 1(i) while (ii) and (iii) are immediate consequences from (i).

Remark 1.

(i) If X is equidistributed on $\{0,1\}^p$ and if we interpret $h_t(\cdot,k)$ and $h'(\cdot,k)$ as 2^p-dimensional vectors the L^2-distance $\sqrt{E_X((h_t(X,k) - h'(X,k))^2)}$ between $h_t(\cdot,k)$ and $h'_t(\cdot,k)$ equals (apart from a constant) the Euclidean distance, and $h_t^*(\cdot,k)$ is the orthogonal projection of $h_t(\cdot,k)$ onto $\mathcal{F}_{u;t}$.
(ii) It is natural to select the function $h_t^* \in \mathcal{F}_{u;t}$ that is 'closest' to h_t, i.e. that minimizes $E_X((h_t(X,k) - h'(X;k))^2)$ on $\mathcal{F}_{u;t}$. Theorem 2 says that h_t^* can alternatively be characterized by another mimimum property (7), and that the approximators $\widetilde{h}_{t_1}^*, \ldots, \widetilde{h}_{t_m}^*$ may be determined separately. Theorem 3 below provides a concrete formula to estimate the unknown coefficients $\beta_{0,t}^*, \ldots, \beta_{u-1,t}^*$ of h_t^* with respect to the base $g_{0,t}, \ldots, g_{u-1,t}$.

(iii) An appropriate choice of the functions $g_{0,t}, \ldots, g_{u-1,t}$, i.e. of $\mathcal{F}_{u;t}$, is essential for the success rate of the attack. Of course, the vector subspace $\mathcal{F}_{u;t}$ should have a small L^2-distance to the unknown function h_t. An appropriate choice may require some insight in the qualitative behaviour of the side channel observables. Clearly, $\mathcal{F}_{u_1,t} \subseteq \mathcal{F}_{u_2,t}$ implies that $h^*_{u_2,t}$ is at least as good $h^*_{u_1,t}$ but the number of measurements in the profiling phase increases with the dimension of $\mathcal{F}_{u;t}$.

Definition 2. *Let V denote an arbitrary set and let $\phi\colon \{0,1\}^p \times \{0,1\}^s \to V$ be a mapping for which the images $\phi(\{0,1\}^p \times k') \subseteq V$ are equal for all subkeys $k' \in \{0,1\}^s$. We say that the function h_t has Property (EIS) ('equal images under different subkeys') if $h_t = \overline{h}_t \circ \phi$ for a suitable mapping $\overline{h}_t\colon V \to \mathbb{R}$, i.e. $h_t(x,k)$ can be expressed as a function of $\phi(x,k)$.*

Example 2. $p = s$, $\phi(x,k) := x \odot k$ where $\odot$ denotes any group operation on $\{0,1\}^p =: V$ (e.g. '$\oplus$').

Lemma 1. *Assume that $h_t(\cdot,\cdot)$ has property (EIS). Then for any pair $(x',k') \in \{0,1\}^p \times \{0,1\}^s$ there exists an element $x'' \in \{0,1\}^p$ with $h_t(x',k') = h_t(x'',k)$.*

Proof. By assumption, $\phi(\{0,1\}^p, k) = \phi(\{0,1\}^p, k')$. Consequently, there exists an $x'' \in \{0,1\}^p$ with $\phi(x'',k) = \phi(x',k')$ and hence $h_t(x'',k) = h_t(x',k')$.

If considerations on the fundamental properties of the physical observables suggest that $h_t(\cdot,\cdot)$ meets (at least approximately) the invariance property (EIS) it is reasonable to select functions g_{jt} that allow representations of the form $g_{jt} = \overline{g}_{jt} \circ \phi$ with $\overline{g}_{jt}\colon V \to \mathbb{R}$. Then

$$h^*_t = \overline{h}^*_t \circ \phi \text{ with } \overline{h}^*_t(y) := \sum_{j=0}^{u-1} \beta_{jt} \overline{g}_{jt}(y) \tag{9}$$

(see Sect. 3.1). As an important consequence it is fully sufficient to determine $\widetilde{h}^*_t(\cdot,k) \in \mathcal{F}_{u;t}$ for any single subkey $k \in \{0,1\}^s$, which is an enourmous advantage over a pure template attack which requires 2^{p+s} templates. An advanced template attack that exploits Lemma 1 requires 2^p templates. If possible, we recommend to select plaintexts from a uniform distribution so that deviations $|h_t(x,k) - h^*_t(x,k)|$ count equally to the L^2-distance for all (x,k). Whether the invariance assumption (EIS) is really justified for $h_t(\cdot,\cdot)$ may be checked by a second profiling with another subkey.

2.2 The Profiling Phase

In this subsection we explain how to determine approximators of $h_t(\cdot,\cdot)$, or more precisely, of $h^*_t(\cdot,\cdot)$ and the distribution of the noise vector $(R_{t_1}, \ldots, R_{t_m})$. We interpret the 'relevant parts' $x_1, x_2, \ldots x_{N_1}$ (i.e. input for the function h_t) of known plaintexts as realization of independent random variables $X_1, X_2, \ldots, X_{N_1}$ that are distributed as X. The Law of Large Numbers implies

$$\frac{1}{N_1}\sum_{j=1}^{N_1}\left(i_t(x_j,k)-h'(x_j,k)\right)^2 \overset{N_1\to\infty}{\longrightarrow} E\left(\left(I_t(X,k)-h'(X,k)\right)^2\right) \tag{10}$$

with probability 1 for any $h'\colon \{0,1\}^p \times \{0,1\}^s \to \mathbb{R}$. Here $i_t(x_j,k)$ denotes the measurement at time t for curve j which has the plaintext part $x_j \in \{0,1\}^p$.

Theorem 3. *(Estimation of h_t) Again, let k denote the correct subkey. For any $h' := \sum_{j=0}^{u-1}\beta'_j g_{jt} \in \mathcal{F}_{u;t}$ we have*

$$\sum_{j=1}^{N_1}\left(i_t(x_j,k)-h'(x_j,k)\right)^2 = \|\boldsymbol{i}_t - A\boldsymbol{b}\|^2 \tag{11}$$

where $A = (a_{ij})_{1\le i\le N_1;0\le j<u}$ is a real-valued $(N_1 \times u)$-matrix, $\boldsymbol{b} \in \mathbb{R}^u$ and $\boldsymbol{i} \in \mathbb{R}^{N_1}$. More precisely, $a_{ij} := g_j(x_i,k)$, $\boldsymbol{b} := (\beta'_0,\ldots,\beta'_{u-1})^T$ and $\boldsymbol{i}_t := (i_t(x_1,k),\ldots,i_t(x_{N_1},k))^T$. Any solution $\boldsymbol{b}^* = (b_0^*,\ldots,b_{u-1}^*)^T$ of

$$A^T A\boldsymbol{b} = A^T\boldsymbol{i}_t \tag{12}$$

minimizes the right-hand side of (11). If the $(u\times u)$-matrix A^TA is regular then

$$\boldsymbol{b}^* = (A^TA)^{-1}A^T\boldsymbol{i}_t. \tag{13}$$

Due to (10) we use the approximator $\widetilde{h}_t^*(x,k) = \sum_{j=0}^{u-1}\beta_{jt}^* g_{jt}(x,k)$ with $\beta_{jt}^* := b_j^*$.

Proof. Equation (11) is obvious whereas (12) is well-known (cf. [7], Subsect. 6.2.1 with $X = A$, $Y = \boldsymbol{i_t}$ and $B = \boldsymbol{b}$; least square estimator) whereas the final assertions are obvious.

Remark 2. We already know that if h_t has the property (EIS) the profiling need only be done for one subkey k. We point out that the adversary need not even know this subkey. In fact, for a given measurement vector $\boldsymbol{i_t}$ the adversary applies Theorem 3 to all possible subkeys $k' \in \{0,1\}^s$ and computes the respective coefficient vectors $\boldsymbol{b}^{*'}$. If $k' \neq k$ Theorem 3 indeed determines an optimal function $\widetilde{h}_t^{*'} \in \mathcal{F}'_{u,t}$ which is spanned by the functions $g'_{jt}(x,k) := g_{jt}(x,k) + (g_{jt}(x,k') - g_{jt}(x,k))$ in place of the g_{jt} while the measurement vector $\boldsymbol{i_t}$ implicitly depends on the (unknown) correct subkey k. Hence it is very likely that $\mathcal{F}'_{u,t}$ has a larger L^2-distance to h_t than $\mathcal{F}_{u;t}$ and, consequently $\|\boldsymbol{i}_t - A\boldsymbol{b}^*\|^2 < \|\boldsymbol{i}_t - A\boldsymbol{b}^{*'}\|^2$ for all instances t. The adversary just adds these squared norms for each admissible subkey over several instants t, and decides for that subkey for which this sum is minimal (see Sect. 3.1 for an experimental verification). In fact, the determination of k is a by-product of the profiling phase which costs no additional measurements. At least principally, this observation could also be used for a direct attack without profiling, which yet requires a sufficient number of measurements.

Definition 3. *$\boldsymbol{R_t}$ denotes the random vector $(R_{t_1},\ldots,R_{t_m})$ in the following. Similarly, we use the abbreviations $\boldsymbol{I_t}(x,k)$, $\boldsymbol{i_t}(x_j,k)$, $\boldsymbol{h_t}(x,k)$ and $\boldsymbol{h_t^*}(x,k)$, where $\boldsymbol{t}$ stands for $(t_1,\ldots,t_m)$.*

After having determined the approximators $\widetilde{h}^*_{t_1}, \ldots, \widetilde{h}^*_{t_m}$ the adversary uses a second set that consists of N_2 measurement curves to estimate the distribution of the m-dimensional random vector $\boldsymbol{R_t} = \boldsymbol{I_t}(X,k) - \boldsymbol{h_t}(X,k)$. We point out that in general the components $R_{t_1}, \ldots, R_{t_m}$ of $\boldsymbol{R_t}$ are not independent, and unlike the functions h_{t_j} they hence cannot be guessed separately. In the most general case the adversary interpolates the N_2 vectors $\{\boldsymbol{i_t}(x_j,k) - \widetilde{\boldsymbol{h}}^*_{\boldsymbol{t}}(x_j,k) \mid j \leq N_2\}$ by a smooth probability density f_0. In the experimental part of this paper we assume that the random vector $\boldsymbol{R_t}$ is jointly normally distributed with covariance matrix $C = (c_{ij})_{1 \leq i,j \leq m}$, i.e. $c_{ij} := E(R_{t_i}R_{t_j}) - E(R_{t_i})E(R_{t_j}) = E(R_{t_i}R_{t_j})$ since $E(R_{t_i}) = E(R_{t_j}) = 0$. If the covariance matrix C is regular the random vector $\boldsymbol{R_t}$ has the m-dimensional density $f_0 := f_C$ with

$$f_C \colon \mathbb{R}^{\mathrm{m}} \to \mathbb{R} \qquad \mathrm{f_C}(\boldsymbol{z}) = \frac{1}{\sqrt{(2\pi)^{\mathrm{m}} \det \mathrm{C}}} \mathrm{e}^{-\frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\mathrm{C}^{-1}\boldsymbol{z}} \tag{14}$$

(cf. [7], for instance). Note that the adversary merely has to estimate the components c_{ij} for $i \leq j$ since the covariance matrix is symmetric.

2.3 The Key Extraction Phase

By our mathematical model $\boldsymbol{I_t}(x,k) - \boldsymbol{h_t}(x,k) = \boldsymbol{R_t}$ for all $(x,k) \in \{0,1\}^p \times \{0,1\}^s$, and $E(R_{t_j}) = 0$ for each $j \leq m$. If $\boldsymbol{R_t}$ has the density $f_0 \colon R^m \to [0,\infty)$ (e.g., $f_0 = f_C$ for a suitable covariance matrix C), and if k° denotes the (unknown) correct subkey of the attacked device then for each $x \in \{0,1\}^p$ we have

$$\boldsymbol{I_t}(x,k^\circ) \text{ has density } f_{k^\circ} \text{ with } f_{k^\circ}(\boldsymbol{z}) := f_0(\boldsymbol{z} - \boldsymbol{h_t}(x,k^\circ)). \tag{15}$$

After having observed N_3 measurement curves (with known parts $x_1, \ldots, x_{N_3}$) the adversary evaluates the product

$$\alpha(x_1, \ldots, x_{N_3}; k) := \prod_{j=1}^{N_3} \widetilde{f}_k\left(\boldsymbol{i_t}(x_j,k^\circ)\right) = \prod_{j=1}^{N_3} \widetilde{f}_0\left(\boldsymbol{i_t}(x_j,k^\circ) - \widetilde{\boldsymbol{h}}^*_{\boldsymbol{t}}(x_j,k)\right) \tag{16}$$

for all subkeys $k \in \{0,1\}^s$ where $\widetilde{f}_0$ denotes the approximation of the exact density f_0 that the adversary has determined in the second step of the profiling phase. Note that $\boldsymbol{i_t}(x_j,k^\circ)$ are observables that depend implicitly on the correct subkey k°. Note further that

$$\widetilde{f}_k\left(\boldsymbol{z}\right) = \widetilde{f}_0\left(\boldsymbol{z} - \widetilde{\boldsymbol{h}}^*_{\boldsymbol{t}}(x,k')\right) = \widetilde{f}_{k^\circ}\left(\boldsymbol{z} + (\boldsymbol{h_t}(x,k^\circ) - \widetilde{\boldsymbol{h}}^*_{\boldsymbol{t}}(x,k'))\right). \tag{17}$$

If the profiling phase has been successful $\boldsymbol{h_t}(x,k^\circ) - \widetilde{\boldsymbol{h}}^*_{\boldsymbol{t}}(x,k') \approx \widetilde{\boldsymbol{h}}^*_{\boldsymbol{t}}(x,k^\circ) - \widetilde{\boldsymbol{h}}^*_{\boldsymbol{t}}(x,k') \approx \boldsymbol{h_t}(x,k^\circ) - \boldsymbol{h_t}(x,k')$ and $\widetilde{f}_0 \approx f_0$. The adversary decides for k' if the term $\alpha(x_1, \ldots, x_{N_3}; k')$ is maximal (maximum likelihood principle).

We point out that the correct subkey k° also fulfils a minimum property:

$$\min_{k' \in \{0,1\}^s} E\left(\|\boldsymbol{I_t}(X,k^\circ) - \boldsymbol{h_t}(X,k')\|^2\right) = E\left(\|\boldsymbol{I_t}(X,k^\circ) - \boldsymbol{h_t}(X,k^\circ)\|^2\right). \tag{18}$$

The situation is similar to Theorem 1 where the correct function $\boldsymbol{h_t}(X, \cdot)$ attains a minimum for the given (correct) subkey. Equation (18) can be verified as Theorem 1. In fact, the left-hand terms in (18) equal $\sum_{j=1}^{m}(E_X(h_{t_j}(x, k^\circ) - h_{t_j}(x, k))^2) + E(R_{t_j}^2))$. As an alternative to the maximum likelihood approach described above the adversary may decide for that subkey $k' \in \{0,1\}^s$ that minimizes

$$\frac{1}{N_2}\sum_{j=1}^{N_2} \|\boldsymbol{i_t}(x_j, k^\circ) - \widetilde{\boldsymbol{h}}_{\boldsymbol{t}}^*(x_j, k')\|^2 \tag{19}$$

This key extraction is less efficient than the maximum likelihood approach as it (explicitly) only considers the deterministic part $\boldsymbol{h}_t$. On the other hand it saves the second part of the profiling phase which may be costly for large m (cf. Sect. 3).

To perform the overall attack the adversary subsequently applies (16) or (19) to obtain the ranking of the candidates for all subkeys. Assuming that one plaintext-ciphertext pair is known, 'candidate vectors' consisting of probable subkey candidates can be checked.

Template attacks aim at h_t itself whereas our approach estimates h_t^*. Hence the key extraction efficiency of the template attacks gives an upper bound for our approach. However, if the vector subspace $\mathcal{F}_{u;t}$ has been chosen appropriately this efficiency gap should be small, especially due to the presence of noise.

We point out that the designer may estimate the risk potential against template attacks by a stochastic simulation. If $\mathcal{F}_{u;t}$ was chosen suitably the $\widetilde{f}_{k'}$ should be close to the true densities $f_{k'}$ and in particular of similar shape. In the simulation the designer yet assumes that the estimated densities $\widetilde{f}_{k'}$ were exact, which corresponds to a template attack with large sample size.

If the attacked device processes several subkeys simultaneously, the efficiency of the overall attack can be further increased by applying a two-step stochastic sieving process, viewing the key extraction process as a sequence of statistical decision problems. The interested reader is referred to [14], Sect. 4 (see also [13], Sect. 7) where such a sieving algorithm was introduced for a timing attack on a weak AES implementation. This sieving process is applicable to hardware-based cryptographic implementations since all subkeys are processed in parallel, but it is not detailed in this contribution.

2.4 Generalizations of Our Model

Our model in equation (1) is not appropriate if the device under test applies algorithmic masking mechanisms that use (pseudo-)random numbers. However, (1) allows a straight-forward generalization. We merely have to replace $h_t\colon \{0,1\}^p \times \{0,1\}^s \to \mathbb{R}$ by $h_{b,t}\colon \{0,1\}^p \times \{0,1\}^v \times \{0,1\}^s \to \mathbb{R}$ where the second argument denotes the random number that is used for masking. Analogously to (3) the minimum

$$\min_{\boldsymbol{h}'_{b,\boldsymbol{t}}\colon \{0,1\}^p \times \{0,1\}^v \times \{0,1\}^s \to \mathbb{R}^m} E\left(\|\boldsymbol{I_t}(X, Y, k) - \boldsymbol{h}'_{b,\boldsymbol{t}}(X, Y, k)\|^2\right) \tag{20}$$

is attained at $\boldsymbol{h}_{b,\boldsymbol{t}}$ where Y denotes a random variable (independent of X and $\boldsymbol{R}_{\boldsymbol{t}}$) that models the random numbers used for masking. Under the reasonable assumption that the designer has access to these random numbers the profiling works analogously as in Subsect. 2.2, yielding a density $\widetilde{f}_{b;0}\colon \mathbb{R}^{\mathrm{m}} \to \mathbb{R}$. In Definition 2 the function ϕ is simply replaced by $\phi_b\colon \{0,1\}^p \times \{0,1\}^v \times \{0,1\}^s \to V$. Of course, in the key extraction phase knowledge of the masking random numbers $y_1, \ldots, y_{N_3}$ cannot be assumed. The designer, resp. the adversary, hence decides for the subkey k' that maximizes the product

$$\alpha_b(x_1, \ldots, x_{N_3}; k) := \prod_{j=1}^{N_3} \sum_{y' \in \{0,1\}^v} \mathrm{Prob}(y_j = y') \widetilde{f}_0\Big(\boldsymbol{i}_{\boldsymbol{t}}(x_j, y, k^\circ) - \widetilde{\boldsymbol{h}}^*_{b,\boldsymbol{t}}(x_j, y', k)\Big) \quad (21)$$

among all $k \in \{0,1\}^s$ (cf. (16)). The mixture of densities on the right-hand side expresses the fact that the true density also depends on the unknown random numbers $y_1, \ldots, y_{N_3}$. If these random numbers are unbiased and independent then $\mathrm{Prob}(Y_j = y') = 2^{-v}$ for all $j \leq N_3$ and $y' \in \{0,1\}^v$. Due to lack of space we skip a formal proof of (21). The generalized model can be used for high-order differential side-channel attacks. One possible goal is to quantify the efficiency of particular masking techniques.

Reference [1] considers the case where signals from several side-channels can be measured simultaneously. Our model can also be generalized to this situation in a natural way: We just have to replace the scalar function $h_t(x,k)$, or more generally $h_{b,t}(x,y,k)$, by the q-dimensional vector $h_{[q],b,t}(x,y,k) := (h_{1,b,t}(x,y,k), \ldots, h_{q,b,t}(x,y,k))$ where $h_{n,b,t}(x,y,k)$ quantifies the deterministic part of the n^{th} side-channel. Similarly, instead of I_t and R_t we consider q-dimensional random vectors $I_{[q],b,t}$ and $R_{[q],b,t}$ for each instant. The correct vector-valued function $\boldsymbol{h}_{[q],b,\boldsymbol{t}}$ minimizes

$$E\left(\sum_{j=1}^{m}\sum_{n=1}^{q}(I_{n,b,t_j}(X,Y,k) - h'_{n,b,t_j}(X,Y,k))^2\right) \quad (22)$$

among all $\boldsymbol{h}'_{[q],b,\boldsymbol{t}}\colon \{0,1\}^p \times \{0,1\}^v \times \{0,1\}^s \to (\mathbb{R}^{\mathrm{q}})^m$.

3 Experimental Analysis

An AES implementation on an 8-bit ATM163 microcontroller was developed for the experimental evaluation of the efficiency achieved by our new decision strategies. The AES was implemented in Assembly language and does not include any countermeasures. The side channel information was gained by measuring the instantaneous current consumption in the ground line. Four measurement series were recorded using 2000 single measurements with a different fixed AES key $\boldsymbol{k} = \{k_1, ..., k_{16}\}$ in each series. The random input data $\boldsymbol{x} = \{x_1, ..., x_{16}\}$ were chosen independently from a uniform distribution. It is $x_l \in \{0,1\}^8$ and $k_l \in \{0,1\}^8$ with $l \in \{1, ..., 16\}$.

The following list summarizes the steps in the profiling (Steps 1 to 4) and key extraction phase (Steps 5 to 7). Note that for the minimum principle Step 4 is skipped ($N_2 = 0$) and Step 6 is applied at key extraction. Instead of Step 6 the maximum likelihood principle uses Step 7.

1. Perform $N_1 + N_2$ measurements using a static key $\boldsymbol{k}$ and known data $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$.
2. With regards to the attacked device select for each instant t the functions $g_{i,t}(\cdot,\cdot)$ that span the vector subspace $\mathcal{F}_{u;t}$.
3. Choose a selection function that combines k_l and x_l and apply Theorem 3 to a subset of N_1 measurements to obtain the estimators $\widetilde{h}_t^*(\cdot,\cdot)$. (Optionally: Repeat Steps 1 to 3 for another test key $\boldsymbol{k}_2$ and compare the results in order to verify the assumption (EIS).)
4. Choose instants $t_1 < \cdots < t_m$. Use the complementary subset of N_2 measurements to obtain the density $\widetilde{f}_0 \colon \mathbb{R}^m \to \mathbb{R}$. (maximum likelihood principle only)
5. Perform N_3 measurements using the target device with the unknown static key $\boldsymbol{k}^\circ$ and known data $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$.
6. Choose instants $t_1 < \cdots < t_m$ and apply (18) and (19) to guess the correct subkey k_l° of the attacked device. (minimum principle only)
7. Apply (16) to guess the correct subkey k_l° of the attacked device. (maximum likelihood principle only)

For comparison, even when exploiting (EIS) template attacks require $2^8 \cdot N_2$ single measurements for an AES implementation.

3.1 The Profiling Phase: Estimation of h_t^*

For profiling we chose the selection function $S(\phi(x,k))$ for the AES S-Box S with $\phi(x,k) = x \oplus k$ where we suppress the byte-number indicating index l of plaintext and subkey. For the vector subspaces we tested different choices, that are evaluated regarding their efficiency in Section 3.3. The chosen vector subspace is applied to the overall time frame, i.e., we do not use a combination of several vector subspaces at different instants.

In this Section, profiling is presented in more detail for the nine-dimensional bit-wise coefficient model, referenced as vector subspace $F_9 = \mathcal{F}_{9;t}$ for all instants t. According to equation (9) with $u = 9$, Theorem 3 and Lemma 1 the deterministic side channel contribution $h_t(\phi(x,k))$ is approximated by

$$\widetilde{h}_t^*(\phi(x,k)) = b_{0t} + \sum_{i=1}^{8} b_{it} \cdot g_i(\phi(x,k)) \tag{23}$$

wherein $g_i(\phi(x,k)) \in \{0,1\}$ is the i-th bit of $S(\phi(x,k))$. The coefficient b_{0t} gives the expectation value of the non-data dependent signal part and the coefficients b_{it} with $i \neq 0$ are the bitwise data dependent signal portions. Though the internal processing of the implementation is deterministic, the measurands are not: noise is an important contribution to the physical signal. The coefficients b_{it}

are revealed by solving an overdetermined system of N_1 linear equations (see Theorem 3).

The experimental results show that the resulting coefficients b_{it} differ in amplitude, so that the use of the Hamming weight model can not be of high quality. The coefficients b_{it} were computed on all four measurement series independently. As it can be exemplary seen in Fig. 1 the deviations of coefficients revealed at the four series are relatively small. As the four series were done with different AES keys, these experimental results confirm the assumptions of Lemma 1 saying that it is justified to perform the profiling of $h_t^*(\cdot, k)\colon \{0,1\}^p \to \mathbb{R}$ for only one subkey $k \in \{0,1\}^s$.

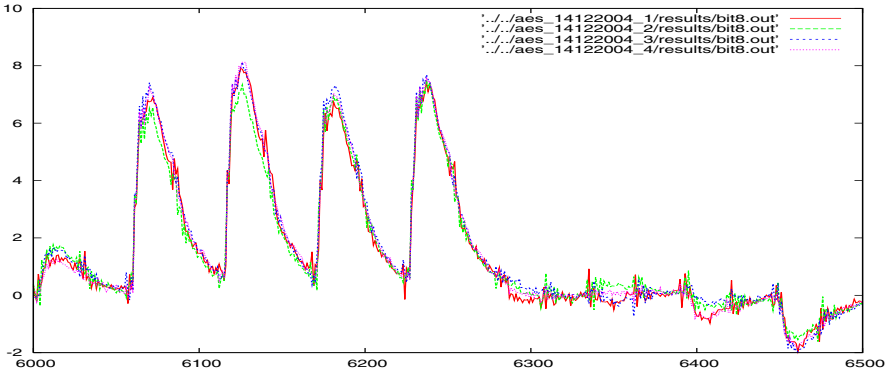

Fig. 1. Coefficient $b_{8,t}$ for all four measurement series as a function of time t. The signals of bit no. 8 (least significant bit) turned out to be the most significant ones. It is $N_1 = 2000$.

Profiling Without Knowing the Key. In case that the subkey k is unknown the estimation of h_t^* may be performed for all possible key values $k' \in \{0,1\}^8$ (cf. Remark 2 in Sect. 2.2). It was experimentally confirmed that the term $\|(i_t(x,k) - \widetilde{h}_t^{*'}(x,k'))^2\|$ indeed was minimal for the correct subkey k. By analyzing the relevant time frame of 6500 instants the difference between the first and the second candidate was 1.9 times larger than the difference between the second and the last candidate. However, we note that the usage of the correlation method [2] to determine k needs less computational efforts.

3.2 The Profiling Phase: Estimation of the Noise

The characterization of the noise was done independently of the estimation of the coefficients b_{it}. Concretely, as preparation step for the maximum likelihood principle we used $N_1 = 1000$ for the extraction of the coefficients b_{it}. The computations of the covariance matrix $C = (c_{ij})_{1 \le i,j \le m}$ for sets of m points were done with $N_2 = 1000$ and $N_2 = 5000$. For the case $N_2 = 5000$ we combined three measurement series, except for the one that is used for the key extraction later on.

3.3 The Key Extraction Phase: Minimum Principle

For the minimum principle given by equations (18) and (19) the estimation of h_t^* is needed, but not the estimation of the noise contribution. If not stated otherwise, only one measurement series served for the profiling step ($N_1 = 2000$) and the key extraction is applied at another series.

First, a suitable choice of m points in time t has to be found[1]. We used $\|b\| = \|(b_{1,t}, b_{2,t}, ..., b_{8,t})\|$ as the measure for our decision. Concretely, we chose the threshold $\tau = 30$ in the following selections for F_9.

S_1: By selecting all instants with $\|b\| \geq \tau$ we obtained seven different signals[2] and the number of instants was $m = 147$. For each signal, most instants are in series.
S_2: At each signal with $\|b\| \geq \tau$ we took the time yielding the maximum value of $\|b\|$. Here, we obtained 7 different instants.
S_3: We chose only one point in time yielding the maximum value of $\|b\|$.
S_4: We chose points that fulfill $\|b\| \geq \tau > var_t$ with $var_t := empVar(i_t(x_j, k) : j \leq N_1)$ denoting the empirical variance. Here, we obtained $m = 100$ different positions in time, but only at five different signals.
S_5: We chose points that fulfill $\|b\| \geq \tau > var_t$ yielding the same result as selection S_4 and we add additionally all points in time that fulfill $\|b\| > \tau$ at the remaining two signals. Altogether, we obtained $m = 120$.
S_6: For each of the seven signals with $\|b\| \geq \tau$ we chose three points by visual inspection, so that the instants chosen are spread over one signal. For the selection S_6 it is $m = 21$.

The minimum value of equation (19) is computed for all subkeys $k' \in \{0,1\}^8$. In this contribution we assess the efficiency by the average number of single measurements needed to achieve a certain success rate using a given number N_3 of single measurements taken from the same measurement set. The success rate (SR) was tested by ten thousand random choices of N_3 single measurements from one series. It can be seen in Table 1 that 10 single measurements yield already a success rate of about 75 % and beyond 30 single measurements the success rate can be above 99.9 %. The best results were gained at the selections S_5 and S_6.

Choice of Vector Subspaces. Different vector spaces are evaluated regarding their efficiency. The choice of high-dimensional vector spaces, e.g, by including all terms of $g_i(\phi(x,k))g_{i'}(\phi(x,k))$ $(i \neq i')$ (see (9) and (23)) did not lead to great improvements. We observed only weak contributions of second-order coefficients that even vanish at many combinations. We present results for

$F_2 = \mathcal{F}_{2;t}$ for all t: the Hamming weight model ($u = 2$),
$F_5 = \mathcal{F}_{5;t}$ for all t: a set of four bit-wise coefficients ($u = 5$) (these are the most significant bit-wise coefficients of F_9),

[1] Note, that we do not consider the covariance of the noise at the chosen points in this approach for key extraction.
[2] We assign all instants that occur during one instruction cycle to one signal.

Table 1. Success Rate (SR) that the correct subkey value is the best candidate as result of (18) and (19) by using N_3 randomly chosen measurements for the analysis at the set of instants S_1 to S_6. The vector space used was F_9

N_3	SR for S_1	SR for S_2	SR for S_3	SR for S_4	SR for S_5	SR for S_6
2	5.57 %	5.64 %	1.06 %	3.31 %	6.35 %	6.36 %
3	12.06 %	11.14 %	1.65 %	7.49 %	13.21 %	13.57 %
5	29.14 %	28.47 %	3.00 %	21.43 %	32.81 %	33.40 %
7	50.39 %	48.20 %	4.39 %	39.41 %	54.23 %	53.88 %
10	75.29 %	73.45 %	8.29 %	65.45 %	78.97 %	78.69 %
15	94.27 %	92.92 %	14.68 %	89.22 %	95.77 %	95.15 %
20	98.57 %	98.31 %	22.26 %	97.59 %	99.17 %	98.82 %
30	99.92 %	99.89 %	39.34 %	99.85 %	99.97 %	99.95 %

Table 2. Success Rate (SR) that the correct key value is the best candidate as result of (18) and (19) by using N_3 randomly chosen measurements in different vector spaces

N_3	SR for F_2 ($\tau = 1$)	SR for F_5 ($\tau = 8$)	SR for F_{10} ($\tau = 30$)	SR for F_{16} ($\tau = 70$)
2	2.59 %	4.22 %	5.18 %	4.81 %
3	4.75 %	9.03 %	11.27 %	9.73 %
5	11.63 %	21.97 %	27.28 %	23.69 %
7	21.66 %	37.61 %	47.66 %	41.04 %
10	37.77 %	62.22 %	72.94 %	65.05 %
15	62.46 %	86.36 %	93.57 %	88.69 %
20	80.36 %	95.71 %	98.41 %	96.17 %
30	96.23 %	99.74 %	99.88 %	99.81 %

$F_{10} =$ $\mathcal{F}_{10;t}$ for all t: a set of the bit-wise coefficient model and one carefully chosen second-order coefficient ($u = 10$), and

$F_{16} =$ $\mathcal{F}_{16;t}$ for all t: the bit-wise coefficient model and seven consecutive second order coefficients ($u = 16$).

For Table 2 the time instants are chosen in the same way as described for F_9 with S_1 at the beginning of Section 3.3 and the thresholds τ are indicated. High-dimensional vector spaces require more measurement curves than low-dimensional ones: There is a trade-off between the number of measurements used during profiling and the dimension of a suitable vector space. In our case, F_9 (see Table 1 and 2) seems to be a good choice though there is some space left for optimization, e.g., by using $N_1 = 5000$, $N_3 = 10$, and $\tau = 10$ the success rate of F_{10} was 80.19% and superseded the corresponding result for F_9 (77.31%). Another optimization would be to select only contributing functions $g_{i,t}(\cdot, \cdot)$ for the chosen vector subspace at the relevant instants.

Comparison with the Correlation Method. Herein, the efficiency gain of the minimum principle is compared with the correlation method of [2] on base of the same pool of measurement data. The correlation method checks for the max-

Table 3. Success Rate (SR) obtained for the correlation method using the 8-bit Hamming weight and the least significant bit (lsb-Bit) as the selection function. The last column shows the SR if the weighted estimated coefficients b_{it} using F_9 are used for the correlation.

N_3	SR (Hamming weight)	SR (lsb-Bit)	SR (estimated b_{it})
5	0.82 %	0.51 %	1.12 %
7	1.31 %	0.84 %	2.37 %
10	2.74 %	1.17 %	4.60 %
15	6.04 %	2.11 %	9.33 %
20	9.70 %	3.55 %	16.67 %
30	19.67 %	6.54 %	31.99 %
50	41.27 %	16.53 %	62.84 %
100	82.85 %	45.22 %	96.13 %

imum correlation peak obtained and it does not evaluate joined sets of multiple instants.

The success rate obtained with the correlation method is illustrated in Table 3 and can be compared with selection S_3 in Table 1 which was restricted to the same instant. In comparison, the correlation method yields worse success rates than the minimum principle. By taking, e.g., $N_3 = 10$ the minimum principle yields an improvement by a factor of 3.0 regarding the Hamming weight prediction and by a factor of 7.1 regarding the best result of one bit prediction of the correlation method. Even, if the estimated coefficients b_{it} of the minimum principle are known an improvement by a factor of 1.8 is achieved. (Note that the relative factor depends on N_3.) As the minimum principle uses the adaptation of probability densities it is advantageous if compared to the correlation method that exploits the linear relationship. Moreover, we point out that the success rate of the minimum principle increases greatly, if multiple signals are jointly evaluated.

3.4 The Key Extraction Phase: Maximum Likelihood Principle

For the maximum likelihood principle as described in Section 2.3 and equation (16) both the estimation of h_t^* and the estimation of the noise is needed. The profiling was done as described in the corresponding parts of Section 3.1 and 3.2.

The m-dimensional random vector $\boldsymbol{Z} = (I_{t_1}(X,k) - \widetilde{h}^*_{t_1}(X,k), \ldots, I_{t_m}(X,k) - \widetilde{h}^*_{t_m}(X,k))$ is assumed to be jointly normally distributed with covariance matrix C. The strategy is to decide for the key hypothesis k' that maximizes equation (16) for the multivariate Gaussian distribution using N_3 measurements which is equivalent to find the minimum of the expression $\sum_{i=1}^{N_3} \boldsymbol{z_i^T} C^{-1} \boldsymbol{z_i}$.

The analysis was done by using the vector subspace F_9 with the selections S_2 and S_6 defined at the beginning of Section 3.3. Note, that for the single instant selection S_3 the maximum likelihood principle reduces to the minimum principle.

Again, the success rate (SR) was computed using ten thousand random choices of one measurement series. As shown in Table 4, based on $N_2 = 1000$

Table 4. Success Rate (SR) that the correct key value is the best candidate as result of equation (16) by using N_3 randomly chosen single measurements for the analysis. All results are based on F_9 with $N_1 = 1000$. If not explicitly stated it is $N_2 = 1000$.

N_3	SR for S_2	SR for S_6	SR for S_2 (N_2=5000)	SR for S_6 (N_2=5000)
2	6.06 %	4.73 %	7.39 %	6.55 %
3	13.93 %	10.45 %	17.06 %	16.00 %
5	36.30 %	28.04 %	43.70 %	41.43 %
7	61.12 %	51.48 %	70.51 %	68.34 %
10	84.33 %	78.26 %	91.08 %	90.17 %
15	97.97 %	95.86 %	99.14 %	99.25 %
20	99.85 %	99.49 %	99.97 %	99.96 %
30	99.99 %	>99.99 %	>99.99%	>99.99 %

a significant improvement was achieved for the selection S_2 regarding Table 1, but not for the selection S_6. This decrease by using the maximum likelihood principle if $N_3 < 15$ and $N_2 = 1000$ for S_6 can be explained by our limited profiling process: the estimation error at the profiling of a 7×7 covariance matrix is significantly lower than the error committed for a 21×21 matrix on the base of $N_2 = 1000$. This assessment is confirmed by the corresponding columns in Table 4 for $N_2 = 5000$. Both the success rates for S_2 and S_6 were further enhanced. As result, a high value for N_2 can be crucial for the maximum likelihood principle, especially if high dimensions are used for the covariance matrix.

The maximum likelihood method needs typically twice the number of measurements during profiling. Therefore, even though key extraction is less efficient under certain circumstances the 'minimum principle' might be preferred. Given 15 measurements, it can be read out from Table 4 that the maximum probability to find the correct key value is 99.25 %. The resulting probability to decide for the correct AES key is $(0.9925)^{16} = 0.8865$.

The number N_3 of measurements can be further reduced if it is tolerated that the correct key value is 'only' among the first best candidates as result of differential side channel cryptanalyis and a plaintext-ciphertext pair is available. E.g., if the correct key value is among the first four subkey candidates with high probability, up to 2^{32} tries remain to localize the correct key value. In case of S_2 and $N_3 = 10$ the corresponding success rate that the correct subkey is at least at the fourth position of the subkey ranking is 97.58 %, if $N_2 = 1000$, and 99.42 %, if $N_2 = 5000$.

4 Conclusion

This contribution proposes a new mathematical approach to optimize the efficiency of differential side channel cryptanalysis by stochastic methods. The quantification of side channel leakage is done in a chosen vector space and does not even (necessarily) require knowledge of one test key. For the key extraction we present a 'minimum principle' that solely uses deterministic data dependencies and the 'maximum likelihood principle' that additionally incorporates the char-

acterization of the noise revealed during profiling. We have shown how our model can be generalized to comprehend both masking countermeasures as well as the usage of multiple physical channels. The theoretical predictions derived from our mathematical model are accompanied and confirmed by experiments. We conclude that the adaptation of probability densities by our methods is clearly advantageous regarding the correlation method, especially, if multiple leakage signals at different instants can be jointly evaluated. Though our efficiency at key extraction is limited by template attacks profiling is much more efficient.

References

1. D. Agrawal, J.R. Rao, P. Rohatgi: Multi-Channel Attacks. In: C.D. Walter, Ç.K. Koç, C. Paar (eds.): Cryptographic Hardware and Embedded Systems — CHES 2003, Springer, LNCS 2779, Berlin 2003, 2–16.
2. M. Aigner, E. Oswald: Power Analysis Tutorial, Technical Report, TU Graz.
3. J.-S. Coron, P. Kocher, D. Naccache: Statistics and Secret Leakage. In: Y. Frankel (ed.): Financial Cryptography (FC 2000), Springer,157–173. LNCS 1962, Berlin 2001.
4. S. Chari, J.R. Rao, P. Rohatgi: Template Attacks. In: B.S. Kaliski Jr., Ç.K. Koç, C. Paar (eds.): Cryptographic Hardware and Embedded Systems — CHES 2002, Springer, LNCS 2523, Berlin 2003, 13–28.
5. J.-S. Coron, P. Kocher, D. Naccache: Statistics and Secret Leakage. In: Y. Frankel (ed.): Finacial Cryptography - FC 2000, Springer, LNCS 1962, Berlin 2001, 157–173.
6. P.N. Fahn, P.K. Pearson: IPA: A New Class of Power Attacks. In: Ç.K. Koç and C. Paar: Cryptographic Hardware and Embedded Systems - CHES 1999, Springer, Lecture Notes in Computer Science 1717, Berlin 1999, 173–186.
7. Fang, K.-T. and Zhang, Y.-T.: Generalized Multivariate Analysis, Berlin, Springer 1990.
8. K. Gandolfi, C. Mourtel, F. Olivier: Electromagnetic Analysis: Concrete Results. In: Ç Koç, D. Naccache, C. Paar (eds.): Cryptographic Hardware and Embedded Systems - CHES 2001, Springer, LNCS 2162, Berlin 2001, 251–261.
9. P.C. Kocher, J. Jaffe, B. Jun: Differential Power Analysis. In: M. Wiener (ed.): Advances in Cryptology – CRYPTO '99, Springer, LNCS 1666, Berlin 1999, 388–397.
10. K. Lemke, K. Schramm, C. Paar: DPA on n-Bit Sized Boolean and Arithmetic Operations and Its Application to IDEA, RC6, and the HMAC-Construction. In: M. Joye and J.-J. Quisquater (eds.): Cryptographic Hardware and Embedded Systems - CHES 2004, Springer, LNCS 3156, Berlin 2004, 205–219.
11. W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery: Numerical Recipes in C — The Art of Scientific Computing. Second Edition, Cambridge University Press, 1992.
12. W. Schindler: A Timing Attack against RSA with the Chinese Remainder Theorem. In: Ç.K. Koç, C. Paar (eds.): Cryptographic Hardware and Embedded Systems — CHES 2000, Springer, LNCS 1965, Berlin 2000, 110–125.
13. W. Schindler: On the Optimization of Side-Channel Attacks by Advanced Stochastic Methods. In: S. Vaudenay (ed.): Public Key Cryptography — PKC 2005, Springer, LNCS 3386, Berlin 2005, 85–103.
14. W. Schindler, F. Koeune, J.-J. Quisquater: Improving Divide and Conquer Attacks Against Cryptosystems by Better Error Detection / Correction Strategies. In: B. Honary (ed.): Cryptography and Coding — IMA 2001, Springer, LNCS 2260, Berlin 2001, 245–267.