Less is More - Dimensionality Reduction from a Theoretical Perspective

Nicolas Bruneau, Sylvain Guilley, Annelie Heuser, Marion Damien, Olivier Rioul

Less is More
Dimensionality Reduction from a Theoretical Perspective

Nicolas Bruneau[1,2(✉)], Sylvain Guilley[1,3], Annelie Heuser[1], Damien Marion[1,3], and Olivier Rioul[1,4]

[1] Institut Mines-Télécom, Telecom ParisTech, Paris, France
{nicolas.bruneau,sylvain.guilley,annelie.heuser,damien.marion, olivier.rioul}@telecom-paristech.fr
[2] AST Division, STMicroelectronics, Rousset, France
[3] Threat Analysis Business Line, Secure-IC S.A.S., Rennes, France
[4] Applied Mathematics Department, École Polytechnique, Palaiseau, France

Abstract. Reducing the dimensionality of the measurements is an important problem in side-channel analysis. It allows to capture multi-dimensional leakage as one single compressed sample, and therefore also helps to reduce the computational complexity. The other side of the coin with dimensionality reduction is that it may at the same time reduce the efficiency of the attack, in terms of success probability.

In this paper, we carry out a mathematical analysis of dimensionality reduction. We show that optimal attacks remain optimal after a first pass of preprocessing, which takes the form of a linear projection of the samples. We then investigate the state-of-the-art dimensionality reduction techniques, and find that asymptotically, the optimal strategy coincides with the linear discriminant analysis.

1 Introduction

Side-channel analysis exploits leakages from devices. Embedded systems are targets of choice for such attacks. Typical leakages are captured by instruments such as oscilloscopes, which sample power or electromagnetic traces. The resulting leaked information about sensitive variables is spread over time.

In practice, two different attack strategies coexist. On the one hand, the various leaked samples can be considered individually—this is typical of *non-profiled attacks* such as Correlation Power Analysis [4]. On the other hand, *profiled attacks* characterize the leakage in a preliminary phase. An efficient leakage modelization should then involve a multi-dimensional probabilistic representation [6].

The large number of samples to feed into the model has always been a problematic issue for multi-dimensional side-channel analysis. One solution is to use techniques to select *points of interest*. Most of them, such as sum-of-square differences (SOSD) and t-test (SOST) [14], are *ad hoc* in that they result from

Annelie Heuser is a Google European fellow in the field of privacy and is partially founded by this fellowship.

a criterion which is independent from the attacker's key extraction objective. Recent criteria, such as leakage maximization by sensitive value [1], avoid this problem. Other formal criteria, related to *non-profiled* attacks, have also been proposed [18,23].

Therefore, there seems to be a converging effort, in both non-profiled and profiled attacks, to *reduce the dimensionality* of multi-dimensional measurements. This desirable property of dimensionality reduction achieves several goals simultaneously:

- it simplifies the side-channel problem (to a single multivariate pdf);
- it concentrates the information (to distinguish using fewer traces); and
- it improves computational speed.

It can be argued, however, that like every preprocessing technique, dimensionality reduction would lose information.

Contributions. In this paper, we tackle this problem of dimensionality reduction from a theoretical viewpoint. Provided that the attacker has full knowledge of the leakage model, we find that "less is more": the advantages of dimensionality reduction can come with no impact on the attack success probability, while improving computational speed.

We derive that the optimal dimensionality reduction process consists in a *linear combination* of samples, which we explicit as a projection on a specific one-dimensional space. For white noise, it turns out that the improved signal-to-noise ratio (SNR) *after* projection is simply the *sum* of the signal-to-noise ratios at the various samples *before* projection.

Finally, we show that the optimal dimensionality reduction technique asymptotically matches the linear discriminant analysis (LDA) preprocessing. We find that LDA generally outperforms principal component analysis (PCA) for which the SNR increases to a lesser extend than LDA, except in the case of white homoscedastic noise where PCA and LDA become equivalent.

We also validate in practice those results on the DPA CONTEST V2 traces [34].

Review of the State-of-the-Art. Dimensionality reduction is part and parcel of profiled attacks. The seminal paper on template attacks [6] is motivated by keeping covariance matrices involved in the training phase sufficiently well conditioned. Manual selection of *relevant leaking points* was discussed in [24] as *educated guesses*. Several automated techniques were proposed, such as sum-of-square differences (SOSD) and t-test (SOST) [14], and also wavelet transforms [11].

Several related metrics were proposed for *leakage detection*. The ANOVA (ANalysis Of VAriance) *F-test* is a ratio between the explained variance and the total variance—see e.g. [7,10] and [3] where it is named *Normalized Inter-Class Variance* (NICV). Also used for linear regression analysis, it is known as the *coefficient of determination*, denoted by the symbol "R^2". It is employed in the context of side-channel analysis in [33] as *multivariate regression analysis* in the presence of white noise, and in [29], where it is used as a distinguisher and as a linearity metric.

PCA has been used to compact traces in [2] and templates in [1]. The eigenvalues of PCA can be viewed as a security metric [15] or even as a distinguisher [30]. This technique is particularly attractive as it can be easily and accurately computed with no divisions involved. It is advocated in [21] that PCA aims at maximizing the inter-class variance, yet it is also important to take the intra-class variance into account. For this reason, LDA has been promoted as an improved alternative. Empirical comparisons were investigated in [26,31,32]. Unfortunately, despite some differences in terms of qualitative efficiency, there is no clear rationale to prefer one method over the other. In fact, it is unclear which of the intrinsic virtue of statistical tools, their implementation, or the dataset is actually responsible for the performance of dimensionality reduction.

Other works attempted to consider different *objective functions*. In [23], the correct key correlation is taken as the objective to be maximized. A similar goal is pursued in [16–19]. Still other dimensionality reduction techniques exist, such as quadratic discriminant analysis, but have not been studied in the side-channel literature. We mention that similar questions have also been raised in the presence of masking countermeasures [5,12,27].

Outline. The remainder of the paper is as follows. The optimal dimensionality reduction is derived theoretically in Sect. 2. Section 3 provides illustrative examples. A comparison with state-of-the-art techniques such as PCA, and LDA [31] is given in Sect. 4. Practical validations on real traces are in Sect. 5. Section 6 concludes.

2 Theoretical Solution in the Presence of Gaussian Noise

2.1 Notations

We adopt a matrix notation. The different queries are indexed by $q = 1, \ldots, Q$, where Q is the number of traces. The different samples in a given trace are indexed by $d = 1, \ldots, D$. Any matrix containing D samples from Q queries is denoted by:

$$M^{D,Q} = (M_{d,q})_{d,q},$$

where $d = 1, \ldots, D$ is a row index and $q = 1, \ldots, Q$ is a column index. We also denote all dth samples for all traces as $(M_{d,q})_q = M_d^Q$, and all the samples for the qth trace as $(M_{d,q})_d = M_q^D$. Thus, M_d^Q is a row vector and M_q^D is a column vector. Two matrices noted side-by-side are implicitly multiplied.

The notation $(\cdot)^\mathsf{T}$ is for transpose. For instance, if $u = u^D$ is $D \times 1$ matrix, then $u^\mathsf{T} = (u^D)^\mathsf{T}$ is a $1 \times D$ matrix. The usual scalar product on $\mathbb{R}^D$ is denoted by $\langle u \mid v \rangle = u^\mathsf{T} v \in \mathbb{R}$. The associated 2-norm of u is $\|u\|_2 = \sqrt{\langle u \mid u \rangle}$.

Random variables will be denoted by capital letters. The probability density function of a random variable X, as a function of x, is denoted by $p_X(x)$ or simply $p(x)$ if the context is clear.

2.2 Model

For most devices, the leakage signal may be represented as a continuous curve as illustrated in Fig. 1. The practical acquisition is done through a temporal series of D "discrete samples" within one clock period.

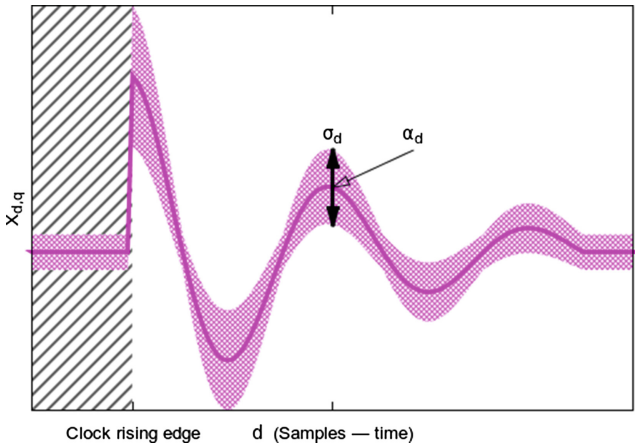

Fig. 1. Example of a modulated trace X_q^D

A sensitive variable that depends on the unknown secret key k^* is leaking through a leakage function ϕ. Typically, ϕ is the Hamming weight function, a sum of weighted bits, or its composition with a substitution box function. In order to further simplify the mathematical derivations, we assume that ϕ is centered. In deriving the optimal attack, it is assumed that the leakage model is perfectly known to the attacker. The model for a given key byte hypothesis k is given by

$$Y_q(k) = \phi(T_q \oplus k), \tag{1}$$

where the random variable T_q denotes a plain or cipher text byte, which is the same for all values of d. Without loss of generality we may assume that $Y_q(k)$ has normalized variance, i.e., $\mathrm{Var}(Y_q(k)) = \mathbb{E}(Y_q^2(k)) = 1$ for all values of q. The actual leakage can be written as

$$X_{d,q} = \alpha_d Y_q(k^*) + N_{d,q}, \tag{2}$$

where the weights α_d are not all zero, k^* is the (unknown) correct key, and $N_{d,q}$ is some random measurement noise. The α_d and noise distribution are assumed known to the attacker.

In matrix notation, we can summarize the equations for different values of d and q by a single matrix equation

$$X^{D,Q} = \alpha^D Y^Q(k^*) + N^{D,Q} \tag{3}$$

where α^D is a single column matrix and $Y^Q(k^*)$ is a single row matrix, whose product is a $D \times Q$ matrix.

We make the stationarity assumption that the noise distribution does not depend on the particular query, that is, the N_q^D are independent and identically distributed independently of the value of q. For a given q, however, the noise samples of N_q^D can very well be correlated. We assume that N_q^D follows a D-dimensional zero-mean Gaussian distribution $\mathcal{N}(0, \Sigma)$, where covariance matrix Σ is a symmetric positive definite $D \times D$ matrix. Therefore, there exists a matrix $\Sigma^{1/2}$, which is such that $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$. We assume that the matrix Σ is known by the attacker.

2.3 Optimal Attack

We focus on the optimal attack as part of our scientific approach to the problem. It is always possible that for some peculiar reason a suboptimal attack actually performs better in the presence of dimensionality reduction. But by the *data processing theorem* [9] any preprocessing like dimensionality reduction can only decrease information about the secret, and, therefore, degrade performance of the *optimal* attack. As a result, it does make sense to minimize the impact of dimensionality reduction on the success rate for this optimal attack so as not to be biased by performance loss or gain due to other factors.

The optimal attack, also known as the template attack [6], consists in applying the *maximum likelihood* principle [20]. Having collected Q traces of dimensionality D in a matrix $x^{D,Q}$, where each trace x_q^D corresponds to a known plaintext t_q, the best key guess that maximizes the probability of success is given by

$$\mathcal{D}(x^{D,Q}, t^Q) = \arg\max_k \; p(x^{D,Q} | t^Q, k^* = k) \tag{4}$$

$$= \arg\max_k \; p_{N^{D,Q}}(x^{D,Q} - \alpha^D y^Q(k)) \tag{5}$$

$$= \arg\max_k \; \prod_{q=1}^{Q} p_{N^{q,D}}(x_q^D - \alpha^D y_q(k)) \tag{6}$$

where

$$p_{N^{q,D}}(z^D) = \frac{1}{\sqrt{(2\pi)^D |\det \Sigma|}} \exp\Big(-\frac{1}{2}(z^D)^\mathsf{T} \Sigma^{-1} z^D\Big). \tag{7}$$

We have used the independence of the queries in (6) and the assumption that at each query, the noise distribution is the same in (7).

Notice that, the optimal attack can as well be a *simple power attack* (if $Q = 1$) or a *differential power attack* (if $Q > 1$), using the terminology from [22]. Still, in the sequel, we focus on attacks which require many traces ($Q \gg 1$).

2.4 Optimal Dimensionality Reduction

We state our main result in the following Theorem 1:

Theorem 1. *The optimal attack on the multivariate traces $x^{D,Q}$ is equivalent to the optimal attack on the monovariate traces $\tilde{x}^Q$, obtained from $x^{D,Q}$ by the formula:*

$$\tilde{x}_q = \frac{\left(\alpha^D\right)^\mathsf{T} \Sigma^{-1} x_q^D}{\left(\alpha^D\right)^\mathsf{T} \Sigma^{-1} \alpha^D} \qquad (q = 1, \ldots, Q). \tag{8}$$

Proof. By taking the logarithm of the expression to be maximized in Eqs. (4)–(7), the optimal distinguisher $\mathcal{D}(x^{D,Q}, t^Q)$ rewrites

$$\mathcal{D}(x^{D,Q}, t^Q) = \arg\min_k \sum_{q=1}^{Q} \left(x_q^D - \alpha^D y_q(k)\right)^\mathsf{T} \Sigma^{-1} \left(x_q^D - \alpha^D y_q(k)\right). \tag{9}$$

For each trace index q, the terms in the sum expand to

$$\begin{aligned} \underbrace{\left(x_q^D\right)^\mathsf{T} \Sigma^{-1} x_q^D}_{\text{cst. } C \text{ independent of } k} & - 2\left(\alpha^D\right)^\mathsf{T} y_q(k) \Sigma^{-1} x_q^D + \left(y_q(k)\right)^2 \left(\alpha^D\right)^\mathsf{T} \Sigma^{-1} \alpha^D \\ &= C - 2 y_q(k) \left[\left(\alpha^D\right)^\mathsf{T} \Sigma^{-1} x_q^D\right] + \left(y_q(k)\right)^2 \left[\left(\alpha^D\right)^\mathsf{T} \Sigma^{-1} \alpha^D\right] \\ &= \left[\left(\alpha^D\right)^\mathsf{T} \Sigma^{-1} \alpha^D\right] \left(y_q(k) - \frac{\left(\alpha^D\right)^\mathsf{T} \Sigma^{-1} x_q^D}{\left(\alpha^D\right)^\mathsf{T} \Sigma^{-1} \alpha^D}\right)^2 + C'. \end{aligned}$$

The latter division is valid since Σ is positive definite and α^D is a nonzero vector. Therefore,

$$\begin{aligned} \mathcal{D}(x^{D,Q}, t^Q) &= \arg\min_k \sum_{q=1}^{Q} \left(y_q(k) - \frac{\left(\alpha^D\right)^\mathsf{T} \Sigma^{-1} x_q^D}{\left(\alpha^D\right)^\mathsf{T} \Sigma^{-1} \alpha^D}\right)^2 \left[\left(\alpha^D\right)^\mathsf{T} \Sigma^{-1} \alpha^D\right] \\ &= \arg\min_k \sum_{q=1}^{Q} \frac{\left(\tilde{x}_q - y_q(k)\right)^2}{\tilde{\sigma}^2}, \end{aligned} \tag{10}$$

where

$$\begin{cases} \tilde{x}_q &= \dfrac{\left(\alpha^D\right)^\mathsf{T} \Sigma^{-1} x_q^D}{\left(\alpha^D\right)^\mathsf{T} \Sigma^{-1} \alpha^D}, \\ \tilde{\sigma} &= \left(\left(\alpha^D\right)^\mathsf{T} \Sigma^{-1} \alpha^D\right)^{-1/2}. \end{cases} \tag{11}$$

We have shown that (9) and (10) are equivalent expressions for the same optimal distinguisher, computed either:

- on multivariate traces x_q^D, with a noise covariance matrix Σ, or:
- on monovariate (i.e., scalar) traces $\tilde{x}_q$, with scalar noise of variance $\tilde{\sigma}^2$. □

Theorem 1 shows that in fact, the optimal attack already integrates an optimal dimensionality reduction. The maximal success rate is not altered.

Definition 2 (Projection vector). *Let V^D be a column of D elements. We call the projection of an acquisition campaign $X^{D,Q}$ on V^D the new mono-sample traces $(V^D)^\mathsf{T} X^{D,Q}$. That is, every trace X_q^D $(1 \leq q \leq Q)$ of the initial campaign is summarized as one sample $(V^D)^\mathsf{T} X_q^D = \langle V^D \mid X_q^D \rangle$.*

Based on this definition, Theorem 1 can be interpreted as follows.

Corollary 3. *The optimal dimensionality reduction is made by a linear combination of the samples where each multivariate trace is projected on the vector $V^D = \frac{\Sigma^{-1}\alpha^D}{(\alpha^D)^\mathsf{T}\Sigma^{-1}\alpha^D}$, of size $D \times 1$.*

Proof. By Theorem 1,

$$\underbrace{\tilde{x}^Q}_{1\times Q \text{ matrix}} = \underbrace{\frac{(\alpha^D)^\mathsf{T}\Sigma^{-1}}{(\alpha^D)^\mathsf{T}\Sigma^{-1}\alpha^D}}_{1\times D \text{ matrix } (V^D)^\mathsf{T}} \underbrace{x^{D,Q}}_{D\times Q \text{ matrix}} . \qquad \square$$

In addition, after this projection, the leakage becomes scalar and can be characterized by a signal-to-noise ratio as shown in the following.

Corollary 4. *After optimal dimensionality reduction, the signal-noise-ratio is given by*

$$\frac{1}{\tilde{\sigma}^2} = (\alpha^D)^\mathsf{T}\Sigma^{-1}\alpha^D.$$

Proof. This is in line with Eq. (10). The random leakage $X^{D,Q}$ is protected onto V^D to yield $\tilde{X}_q = Y_q(k) + \tilde{N}$ $(q = 1, \ldots, Q)$ where $\tilde{N}$ is an additive white Gaussian noise (AWGN) distributed as $\mathcal{N}(0, ((\alpha^D)^\mathsf{T}\Sigma^{-1}\alpha^D)^{-1})$. Recall that the variance of the leakage model has been assumed normalized $= 1$. Therefore, the signal-to-noise ratio equals

$$\frac{\mathrm{Var}(Y_q(k))}{\mathrm{Var}(\tilde{N})} = \frac{1}{((\alpha^D)^\mathsf{T}\Sigma^{-1}\alpha^D)^{-1}} = (\alpha^D)^\mathsf{T}\Sigma^{-1}\alpha^D \ . \qquad \square$$

The SNR is an interesting metric on its own, because it quantifies how much the signal has been concentrated (its power increased) for a given noise level. Furthermore, the SNR directly relates to the success rate of optimal attacks [13].

2.5 Discussion

It is interesting to note that the optimal dimensionality reduction does not depend on the actual distribution of $Y^D(k)$, the deterministic part of the leakage model. This means that irrespective of the leakage function ϕ, the best dimensionality reduction depends only on signal weights α^D and on noise covariance Σ.

Similarly, the optimal dimensionality reduction does not depend on the *confusion coefficient* of the leakage model [13]: for identical weight and noise distribution, the optimal linear combination of leakages is the same whether an XOR or a substitution box operation is targeted.

3 Examples

3.1 White Noise

One interesting situation is when the noise samples are uncorrelated (see for instance [33] for an experimental setup). The covariance matrix Σ is diagonal:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_D^2 \end{pmatrix}.$$

Proposition 5. *For white noise, the optimal dimensionality reduction takes the form:*

$$\tilde{x}_q = \frac{\sum_{d=1}^{D} \frac{\alpha_d}{\sigma_d^2} x_{d,q}}{\sum_{d=1}^{D} \frac{\alpha_d^2}{\sigma_d^2}} \qquad (q = 1, \ldots, Q) \tag{12}$$

Proof. Apply Theorem 1, where Σ^{-1} is diagonal with diagonal entries $1/\sigma_d^2$. □

Let $\mathrm{SNR}_d = \alpha_d^2/\sigma_d^2$ be the initial signal-to-noise ratio at the dth sample *before* dimensionality reduction.

Proposition 6. *For white noise, the equivalent signal-to-noise ratio* after *optimal dimensionality reduction is given by the sum*

$$\widetilde{SNR} = \sum_{d=1}^{D} SNR_d. \tag{13}$$

Proof. By Corollary 4, $\widetilde{\mathrm{SNR}} = (\alpha^D)^\mathsf{T} \Sigma^{-1} \alpha^D = \sum_{d=1}^{D} \frac{\alpha_d^2}{\sigma_d^2} = \sum_{d=1}^{D} \mathrm{SNR}_d$. □

Thus, combining independent multidimensional samples within one trace increases the signal-to-noise as if those samples were captured in D independent traces. In this case having Q traces of D samples each is simply the same as having $Q \times D$ independent monovariate traces.

3.2 Correlated Autoregressive Noise

A more general situation is when the samples are correlated like an autoregressive process. More precisely, assume that all samples share the same noise distribution of variance σ^2, and that two consecutive noise samples have correlation factor equal to $\rho \in]-1, +1[$. The correlation factors ρ typically models an autoregressive low-pass filtering of the acquisition setup (see Sect. 5.2 for a real-world example). The noise covariance matrix takes the Toeplitz form:

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{D-2} & \rho^{D-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{D-3} & \rho^{D-2} \\ \rho^2 & \rho & 1 & \rho & \cdots & \rho^{D-4} & \rho^{D-3} \\ \rho^3 & \rho^2 & \rho & 1 & \cdots & \rho^{D-5} & \rho^{D-4} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho^{D-2} & \rho^{D-3} & \rho^{D-4} & \rho^{D-5} & \cdots & 1 & \rho \\ \rho^{D-1} & \rho^{D-2} & \rho^{D-3} & \rho^{D-4} & \cdots & \rho & 1 \end{pmatrix} = \left(\sigma^2 \rho^{|d-d'|}\right)_{1 \le d,d' \le D}.$$

We emphasize that $|\rho|$ is strictly smaller than one in keeping with the assumption that Σ be positive definite. When $\rho = 0$, the noise becomes white as in the preceding subsection.

Proposition 7. *For autoregressive noise, the optimal dimensionality reduction takes the form:*

$$\tilde{x}_q = \tfrac{1}{\sigma^2(1-\rho^2)} \Big[(\alpha_1 - \rho\alpha_2) x_{q,1} + \textstyle\sum_{d=2}^{D-1} ((1+\rho^2)\alpha_d - \rho(\alpha_{d-1} + \alpha_{d+1})) x_{d,q} \\ + (\alpha_D - \rho\alpha_{D-1}) x_{q,D} \Big]. \quad (14)$$

Proof. It can easily be checked that Σ^{-1} is tridiagonal:

$$\Sigma^{-1} = \frac{1}{\sigma^2(1-\rho^2)} \begin{pmatrix} 1 & -\rho & 0 & 0 & \cdots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & \cdots & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & -\rho & \cdots & 0 & 0 \\ 0 & 0 & -\rho & 1+\rho^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & 0 & \cdots & -\rho & 1 \end{pmatrix}.$$

Then apply Theorem 1:

$$\tilde{x}_q = \tfrac{1}{\sigma^2(1-\rho^2)} \begin{pmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{D-1} & \alpha_D \end{pmatrix} \begin{pmatrix} 1 & -\rho & \cdots & 0 & 0 \\ -\rho & 1+\rho^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1+\rho^2 & -\rho \\ 0 & 0 & \cdots & -\rho & 1 \end{pmatrix} \begin{pmatrix} x_{q,1} \\ x_{q,2} \\ \vdots \\ x_{q,D-1} \\ x_{q,D} \end{pmatrix}$$

and expand. □

Notice that in the optimal dimensionality reduction, each leakage sample $x_{d,q}$ is not only weighted by its corresponding α_d but also by its two neighbor weights $\alpha_{d\pm1}$, provided the latter exist.

Proposition 8. *For autoregressive noise, the equivalent signal-to-noise ratio after optimal dimensionality reduction is given by*

$$\widetilde{SNR} = \tfrac{1}{\sigma^2(1-\rho^2)} \left[\alpha_1^2 + (1+\rho^2) \textstyle\sum_{d=2}^{D-1} \alpha_d^2 + \alpha_D^2 - 2\rho \sum_{d=1}^{D-1} \alpha_d \alpha_{d+1} \right]. \quad (15)$$

Proof. Apply Corollary 4:

$$\widetilde{\mathrm{SNR}} = \frac{1}{\sigma^2(1-\rho^2)} \begin{pmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{D-1} & \alpha_D \end{pmatrix} \begin{pmatrix} 1 & -\rho & \cdots & 0 & 0 \\ -\rho & 1+\rho^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1+\rho^2 & -\rho \\ 0 & 0 & \cdots & -\rho & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{D-1} \\ \alpha_D \end{pmatrix}$$

and expand. □

Corollary 9. *For equal weights $\alpha_1 = \cdots = \alpha_D = \alpha$, i.e., when initial signal-to-noise ratios $SNR_1 = \cdots = SNR_D = SNR$ are the same, one has*

$$\widetilde{SNR} = SNR \times \frac{D(1-\rho)+2\rho}{1+\rho}. \tag{16}$$

Proof. Proposition 8 reduces to

$$\begin{aligned} \widetilde{\mathrm{SNR}} &= \frac{\alpha^2}{\sigma^2(1-\rho^2)} \left(2 + (D-2)(1+\rho^2) - 2\rho(D-1)\right) \\ &= \frac{\alpha^2}{\sigma^2(1-\rho)(1+\rho)} \left((1-\rho)(D-\rho(D-2))\right) \\ &= \frac{\alpha^2}{\sigma^2} \frac{1}{1+\rho} \left(D-\rho(D-2)\right) = \mathrm{SNR} \times \frac{D(1-\rho)+2\rho}{1+\rho}. \end{aligned}$$ □

In other words, optimal dimensionality reduction has the effect of multiplying the monovariate SNR by the factor $\frac{D-\rho(D-2)}{1+\rho}$. This gain factor is of course equal to 1 for dimension $D = 1$, but becomes strictly greater than 1 for larger dimensions, since $\frac{D-\rho(D-2)}{1+\rho} > \frac{D-(D-2)}{2} = 1$ where we have used that $\rho > -1$ or $\frac{1}{1+\rho} > \frac{1}{2}$.

For very small values of correlation ρ, Taylor expansion about $\rho = 0$ gives $\frac{D-\rho(D-2)}{1+\rho} = D - 2(D-1)\rho + \mathcal{O}(\rho^2)$. The SNR gain is equal to the dimension D at first order, which is consistent with Proposition 6. In addition, that gain is never greater than D, since $\frac{D(1-\rho)+2\rho}{1+\rho} \leq \frac{D(1-\rho)+2D\rho}{1+\rho} = D$. Therefore, when $\mathrm{SNR}_1 = \ldots = \mathrm{SNR}_D$, nonzero values of correlation ρ decrease the efficiency of dimensionality reduction, the most favorable situation being the case of white noise samples.

4 Comparison with PCA and LDA

When the attacker does not precisely know the model given by Eq. (2), the optimal dimensionality reduction cannot be applied directly. In this section, we analyse theoretically two well-known engineering solutions to reduce the dimensionality: PCA and LDA. Both techniques are based on eigen decompositions.

4.1 Principal Components Analysis (PCA)

Principal components analysis aims at identifying directions in the *centered* data set $M^{D,Q} = (M_{d,q})_{d,q}$ defined by

$$M_{d,q} = X_{d,q} - \frac{1}{Q}\sum_{q'=1}^{Q} X_{d,q'} \qquad (1 \leq q \leq Q, 1 \leq d \leq D). \tag{17}$$

The directions of PCA are the eigenvectors of $M^{D,Q}(M^{D,Q})^{\mathsf{T}}$.

Proposition 10. *Asymptotically as* $Q \longrightarrow +\infty$,

$$\frac{1}{Q}M^{D,Q}(M^{D,Q})^{\mathsf{T}} \longrightarrow \alpha^D(\alpha^D)^{\mathsf{T}} + \Sigma. \tag{18}$$

Proof. By the law of large numbers,

$$\frac{1}{Q}\sum_{q=1}^{Q} M_{d,q}M_{d',q} \longrightarrow \mathrm{Cov}(M_{d,q}, M_{d',q})$$

almost surely, where the covariance term can be computed as: $\mathrm{Cov}(M_{d,q}, M_{d',q}) = \mathrm{Cov}(\alpha_d Y_q + N_{d,q}, \alpha_{d'} Y_q + N_{d',q})$. When expanding this expression, cross terms disappear by independence of Y^Q and $N^{D,Q}$. There remains:

$$\mathrm{Cov}(M_{d,q}, M_{d',q}) = \alpha_d\alpha_{d'} + \Sigma_{d,d'}$$

where we have used the hypothesis that Y_q has unit variance. □

The classical PCA has the drawback that $M^{D,Q}(M^{D,Q})^{\mathsf{T}}$ depends both on the *signal* and on the *noise*. *Inter-class PCA* has been introduced in [1]. The matrix $M^{D,Q}$ used in the PCA is traded for a more simple matrix $Z^{D,\#Y}$, where each column, indexed by y, is the centered column $\frac{1}{\sum_{\substack{1\leq q\leq Q\\ Y_q=y}} 1}\sum_{\substack{1\leq q\leq Q\\ Y_q=y}} X_q^D$. One advantage of this method is that it explicitly takes into account the sensitive variables Y.

It can be easily checked, that, asymptotically, each column Z_y^D tends to $\alpha^D y$ when $Q \longrightarrow +\infty$. Therefore, $Z^{D,\#Y}(Z^{D,\#Y})^{\mathsf{T}}$ tends to a $D \times D$ matrix proportional to $\alpha^D(\alpha^D)^{\mathsf{T}}$. Here, the noise has been averaged away in each class y, which is a second advantage. Therefore, in the sequel, we shall refer to the inter-class PCA of [1] simply as PCA.

We have the following spectral characterization of the asymptotic PCA:

Proposition 11. *Asymptotically, PCA has only one principal direction, namely the vector* α^D.

Proof. By Proposition 10, the PCA matrix tends asymptotically to $\alpha^D(\alpha^D)^\mathsf{T}$. This $D \times D$ matrix has rank one, because all its columns are multiple of α^D. Since

$$(\alpha^D(\alpha^D)^\mathsf{T})\alpha^D = \alpha^D((\alpha^D)^\mathsf{T}\alpha^D) = \left\|\alpha^D\right\|_2^2 \times \alpha^D,$$

α^D is the eigenvector with corresponding nonzero eigenvalue $= \|\alpha^D\|_2^2$. □

Notice that the uniqueness of the eigenvector for PCA holds in our model (2). However, Proposition 11 would not hold if e.g., the noise were correlated to the signal.

Remark 1. The classical PCA has the same eigenvector α^D if the noise is *isotropic*, i.e., white and of same variance in every dimension.

The paper [1] presents an optimization procedure to find the eigenelements.

Proposition 12. *The asymptotic signal-to-noise ratio after projection using PCA is equal to* $\frac{\left\|\alpha^D\right\|_2^4}{(\alpha^D)^\mathsf{T}\Sigma\alpha^D}$.

Proof. After projection on the (asymptotic) eigenvector α^D, the leakage becomes: $(\alpha^D)^\mathsf{T}\alpha^D Y_q(k^*)+(\alpha^D)^\mathsf{T}N_q^D$. The projected signal is $((\alpha^D)^\mathsf{T}\alpha^D)Y_q(k^*)$. The projected noise is $(\alpha^D)^\mathsf{T}N_q^D$, which remains centered. Its variance is equal to the expectation of its square:

$$\begin{aligned}\mathrm{Var}((\alpha^D)^\mathsf{T}N_q^D) &= \mathbb{E}\left((\alpha^D)^\mathsf{T}N_q^D\right)^2 = \mathbb{E}\left((\alpha^D)^\mathsf{T}N_q^D(N_q^D)^\mathsf{T}\alpha^D\right)\\ &= (\alpha^D)^\mathsf{T}\mathbb{E}\left(N_q^D(N_q^D)^\mathsf{T}\right)\alpha^D = (\alpha^D)^\mathsf{T}\Sigma\alpha^D.\end{aligned}$$

Therefore,

$$\mathrm{SNR}_{\mathrm{PCA}} = \frac{\mathrm{Var}(((\alpha^D)^\mathsf{T}\alpha^D)Y_q(k^*))}{\mathrm{Var}((\alpha^D)^\mathsf{T}N_q^D)} = \frac{\mathrm{Var}(\left\|\alpha^D\right\|_2^2 Y_q(k^*))}{(\alpha^D)^\mathsf{T}\Sigma\alpha^D} = \frac{\left\|\alpha^D\right\|_2^4}{(\alpha^D)^\mathsf{T}\Sigma\alpha^D}.$$ □

Example 13. For white noise (Sect. 3.1)

$$\mathrm{SNR}_{\mathrm{PCA}} = \frac{\left(\sum_{d=1}^D \alpha_d^2\right)^2}{\sum_{d=1}^D \alpha_d^2\sigma_d^2}. \tag{19}$$

Example 14. For autoregressive noise (Sect. 3.2)

$$\mathrm{SNR}_{\mathrm{PCA}} = \frac{\sum_{d=1}^D \alpha_d^2}{\sigma^2}\frac{1}{1+\frac{2}{\sum_{d=1}^D \alpha_d^2}\sum_{d=1}^{D-1}\rho^d\sum_{d'=1}^{D-d}\alpha_{d'}\alpha_{d'+d}}. \tag{20}$$

We can now compare the performance of the asymptotic PCA to the optimal dimensionality reduction.

Theorem 15. *The SNR of the asymptotic PCA is smaller than the SNR of the optimal dimensionality reduction.*

Proof. By assumption the noise covariance matrix is symmetric positive definite, hence there exists a matrix $\Sigma^{1/2}$, which is such that $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$. By Cauchy-Schwarz inequality,

$$\left(\langle \Sigma^{-1/2}\alpha^D \mid \Sigma^{1/2}\alpha^D \rangle\right)^2 \leq \left\|\Sigma^{-1/2}\alpha^D\right\|_2^2 \cdot \left\|\Sigma^{1/2}\alpha^D\right\|_2^2.$$

Therefore, $\mathrm{SNR}_{\mathrm{PCA}} = \frac{((\alpha^D)^\mathsf{T}\alpha^D)^2}{(\alpha^D)^\mathsf{T}\Sigma\alpha^D} \leq (\alpha^D)^\mathsf{T}\Sigma^{-1}\alpha^D = \widetilde{\mathrm{SNR}}$. □

Corollary 16. *The asymptotic PCA has the same SNR as the the optimal dimensionality reduction if and only if α^D is an eigenvector of Σ. In this case, both dimensionality reductions are equivalent.*

Proof. Equality holds in Theorem 15 if and only if there exists a nonzero real number λ such that $\Sigma^{1/2}\alpha^D = \lambda\Sigma^{-1/2}\alpha^D$, i.e., $\Sigma\alpha^D = \lambda\alpha^D$, i.e., α^D is an eigenvector of Σ.

In this case, the optimal protection is on the vector $\Sigma^{-1}\alpha^D = \frac{1}{\lambda}\alpha^D$, which is proportional to the projection vector belonging to the asymptotic PCA. □

Remark 2. Assume white noise (Sect. 3.1) where all values σ_d^2 $(1 \leq d \leq D)$ are different. Then, by Corollary 16, the asymptotic PCA is optimal only if $\alpha^D = (0, 0, \cdots, 0, 1, 0, \cdots, 0)$, which we may consider unrealistic since only one sample out of D would leak secret information.

In contrast, if $\sigma_1 = \cdots = \sigma_D = \sigma$, the covariance matrix has only one eigenvalue, namely $(1, 1, \cdots, 1)$, which has multiplicity D. Thus, for white homoscedastic noise, PCA is asymptotically optimal if and only if $\alpha_1 = \cdots \alpha_D = \alpha$, that is, the SNR is the same for each sample.

Still in the case of white noise, we can lower bound the SNR of the asymptotic PCA:

Lemma 17. *For white noise, the SNR of the asymptotic PCA is not less than the worst SNR among the samples, but can be strictly smaller than the higher SNR among the samples.*

Proof. We have

$$\sum_{d=1}^{D} \alpha_d^2\sigma_d^2 = \sum_{d=1}^{D} \frac{\sigma_d^2}{\alpha_d^2}\alpha_d^4 \leq \left(\max_{d=1}^{D} \frac{\sigma_d^2}{\alpha_d^2}\right) \sum_{d=1}^{D} \alpha_d^4.$$

Since $\left(\max_{d=1}^{D} \frac{\sigma_d^2}{\alpha_d^2}\right)^{-1} = \min_{d=1}^{D} \frac{\alpha_d^2}{\sigma_d^2} = \min_{d=1}^{D} \mathrm{SNR}_d$, the expression of the SNR of the asymptotic PCA given by Eq. (19) is such that

$$\mathrm{SNR}_{\mathrm{PCA}} = \frac{\left(\sum_{d=1}^{D} \alpha_d^2\right)^2}{\sum_{d=1}^{D} \alpha_d^2\sigma_d^2} \geq \frac{\left(\sum_{d=1}^{D} \alpha_d^2\right)^2}{\sum_{d=1}^{D} \alpha_d^4} \min_{d=1}^{D} \mathrm{SNR}_d \geq \min_{d=1}^{D} \mathrm{SNR}_d \tag{21}$$

where we have used Cauchy-Schwarz inequality $\sum_{d=1}^{D} \alpha_d^2\alpha_d^2 \leq \left(\sum_{d=1}^{D} \alpha_d^2\right)^2$.

Conversely, we can give an example for which $\mathrm{SNR}_{\mathrm{PCA}} < \max_{d=1}^{D} \frac{\alpha_d^2}{\sigma_d^2}$. Take $D = 2$, $\alpha_1 = \alpha_2 = 1$, $\sigma_1 = 1$ and $\sigma_2 = 10$. Then $\mathrm{SNR}_{\mathrm{PCA}} = 4/(1+10^2) = 4/101$, which is strictly smaller than $\alpha_1^2/\sigma_1^2 = 1$. □

4.2 Linear Discriminant Analysis (LDA)

LDA has been introduced in side-channel analysis in [31]. With respect to inter-class PCA, it computes the eigenvectors of the matrix $S_w^{-1} S_b$, where:

- S_w is the *within-class scatter matrix*, asymptotically equal to Σ, and
- S_b is the *between-class scatter matrix*, equal to $\alpha^D (\alpha^D)^\mathsf{T}$.

We have the following spectral characterization of the asymptotic LDA:

Proposition 18. *Asymptotically, LDA has only one principal direction, namely the vector* $\Sigma^{-1}\alpha^D$.

Proof. The matrix $S_w^{-1} S_b = \Sigma^{-1}\alpha^D(\alpha^D)^\mathsf{T}$ has rank one. Indeed, $\alpha^D(\alpha^D)^\mathsf{T}$ has rank one, and multiplying by an invertible matrix (namely Σ^{-1}) keeps the rank unchanged. Since

$$(\Sigma^{-1}\alpha^D(\alpha^D)^\mathsf{T})\Sigma^{-1}\alpha^D = \Sigma^{-1}\alpha^D((\alpha^D)^\mathsf{T}\Sigma^{-1}\alpha^D) = \left((\alpha^D)^\mathsf{T}\Sigma^{-1}\alpha^D\right) \times \Sigma^{-1}\alpha^D,$$

$\Sigma^{-1}\alpha^D$ is the unique eigenvector with corresponding eigenvalue $(\alpha^D)^\mathsf{T}\Sigma^{-1}\alpha^D > 0$. This eigenvalue is equal to the SNR of the asymptotic LDA. □

By Corollary 4, the SNR of the asymptotic LDA is equal to the SNR of the optimal dimensionality reduction, denoted by $\widetilde{\mathrm{SNR}}$. In fact, we have the following.

Theorem 19. *The asymptotic LDA computes exactly the optimal dimensionality reduction.*

Proof. Compare Theorem 1 with Proposition 18: in both cases, the projection vector is collinear with $\Sigma^{-1}\alpha^D$. □

4.3 Numerical Comparison Between Asymptotic PCA and LDA

Numerical comparison between asymptotic PCA and LDA is given in Fig. 2(a) and (b), for $D = 6$ samples. The noise is chosen autoregressive, with $\sigma = 1$ and different values for ρ (Sect. 3.2). The vector α^D is chosen equal to $(1, 1, 1, 1, 1, 1)^\mathsf{T}$ in Fig. 2(a) and to $\sqrt{6.0/6.4} \cdot (1.0, 1.1, 1.2, 1.3, 0.9, 0.5)^\mathsf{T}$ in Fig. 2(b), such that $\widetilde{\mathrm{SNR}} = 6$ when $\rho = 0$. The SNR of the asymptotic LDA is that of the optimal dimensionality reduction (cf. Corollary 4), and that of the asymptotic PCA can be found in Example 14. The first case (Fig. 2(a)) fits the situation depicted in Corollary 9. The asymptotic PCA and LDA are almost similar. Besides, when $\rho \to 1^-$, both SNRs tend to 1 (recall Eqs. (20) and (16)). But, when the SNR

varies over the D samples (Fig. 2(b)), the asymptotic LDA can be significantly better than the asymptotic PCA. The sample-wise extremal SNRs ($\mathrm{SNR}_d = \alpha_d^2/\sigma^2$) are also represented: the SNR of the PCA can be smaller than the largest SNR, namely $\max_{1\leq d\leq D} \mathrm{SNR}_d$, (recall Lemma 17), which is not the case of the SNR of the LDA. Actually, the SNR of LDA increases to infinity because $\widetilde{\mathrm{SNR}} \approx 0.164/(1-\rho)$ when $\rho \to 1^-$ (see Eq. (15)).

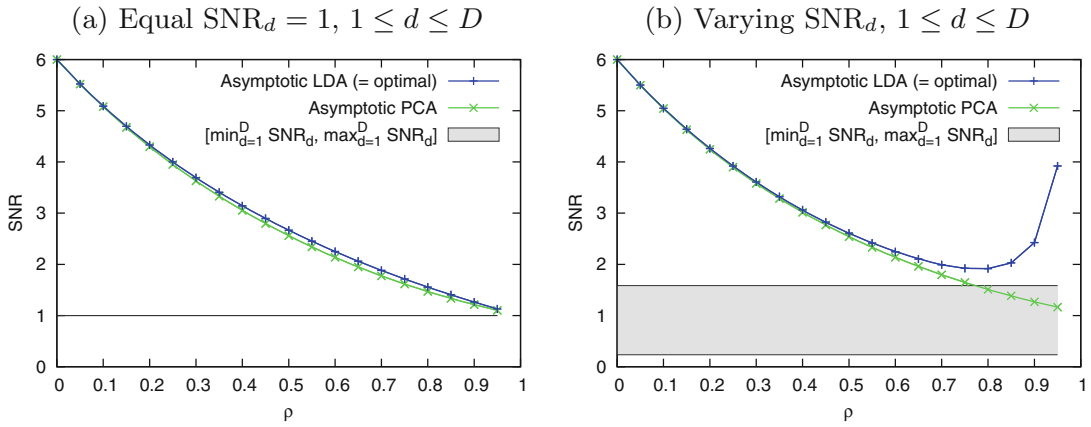

Fig. 2. Comparison of the SNR of asymptotic LDA (optimal) and of asymptotic PCA

5 Practical Validation

In this section, we investigate real traces. Experiments are carried out on the DPA CONTEST V2 [34] traces. One clock cycle lasts $D = 200$ samples. As traces are captured from a hardware implementation of an AES, we consider the Hamming distance leakage model (in accordance with most attacks reported on the analyzed device [8], namely a SASEBO-GII board with a Xilinx XC5VLX30 FPGA [28]). In the sequel, we focus on the Hamming distance between the byte 0 of the last round and that of the cipher text. That is, the function ϕ in Eq. (1) is a normalized Hamming weight; precisely, $\phi : z \in \mathbb{F}_2^n \mapsto \frac{2}{\sqrt{n}}\left(w_H(z) - \frac{n}{2}\right)$, where $n = 8$, because AES is a byte-oriented block cipher. In addition, we emphasize that our model (Eq. (2)) is indeed suitable to leakage dimensionality reduction within one clock period.

5.1 Precharacterization of the Model Parameters α^D and Σ

In order to characterize the model, we need to recover the column matrix α^D and the $D \times D$ covariance matrix Σ of the noise.

Proposition 20. *The parameters of the model (2) which minimize the fitting error are given by*

$$\hat{\alpha}^D = \frac{X^{D,Q}(Y^Q)^\mathsf{T}}{Y^Q(Y^Q)^\mathsf{T}}.$$

Proof. The goal (minimizing the fitting error) is similar to that of the optimal distinguisher, namely maximize the probability of $p_{N^{D,Q}}(X^{D,Q} - \alpha^D Y^Q)$ (Eq. (6)). But in the context of characterization, the correct key is known. Therefore, we wish to minimize in α^D and Σ the following objective function:

$$\text{objective}(\alpha^D, \Sigma) = \sum_{q=1}^{Q} \left\{ \left(x_q^D - \alpha^D y_q(k^*)\right)^\mathsf{T} \Sigma^{-1} \left(x_q^D - \alpha^D y_q(k^*)\right) \right\}, \tag{22}$$

which reminds of Eq. (9) (except that now, the key $k = k^*$ is known). We use the notation $(\hat{\alpha}^D, \hat{\Sigma}) = argmin_{(\alpha^D, \Sigma)}$ objective(α^D, Σ).

We fix Σ and minimize only on α^D. The gradient of objective(α^D, Σ) w.r.t. $(\alpha^D)^\mathsf{T}$ writes:

$$\frac{\partial}{\partial (\alpha^D)^\mathsf{T}} \text{objective}(\alpha^D, \Sigma) = \sum_{q=1}^{Q} -2y_q(k^*) \left(\Sigma^{-1} x_q^D - y_q(k^*) \Sigma^{-1} \alpha^D\right). \tag{23}$$

The objective function is extremal in $\hat{\alpha}^D$ if and only if its derivative is equal to zero at this point. Let Y^Q be an abbreviation for $Y^Q(k^*)$. This condition takes the form of a *normal equation*

$$\hat{\alpha}^D = \frac{\sum_{q=1}^{Q} y_q x_q^D}{\sum_{q=1}^{Q} y_q^2} = \frac{X^{D,Q}(Y^Q)^\mathsf{T}}{Y^Q(Y^Q)^\mathsf{T}}. \tag{24}$$

where the numerator is the inter-covariance matrix of $X^{D,Q}$ and Y^Q, and the denominator is the covariance matrix of Y^Q. □

Interestingly, the most likely value $\hat{\alpha}^D$ of α^D does not depend on the noise covariance matrix. As $N^{D,Q} = X^{D,Q} - \hat{\alpha}^D Y^Q$ has zero mean, the latter can be evaluated on its own as the well-known unbiased estimator of Σ:

$$\hat{\Sigma} = \frac{1}{Q-1} (X^{D,Q} - \hat{\alpha}^D Y^Q)(X^{D,Q} - \hat{\alpha}^D Y^Q)^\mathsf{T}. \tag{25}$$

By plugging Eq. (24) into Eq. (25), one obtains

$$\begin{aligned}
\hat{\Sigma} &= \tfrac{1}{Q-1} \left(X^{D,Q} - X^{D,Q} \frac{(Y^Q)^\mathsf{T} Y^Q}{Y^Q (Y^Q)^\mathsf{T}} \right) \left(X^{D,Q} - X^{D,Q} \frac{(Y^Q)^\mathsf{T} Y^Q}{Y^Q (Y^Q)^\mathsf{T}} \right)^\mathsf{T} \\
&= \tfrac{1}{Q-1} X^{D,Q} \left(I^{Q,Q} - \frac{(Y^Q)^\mathsf{T} Y^Q}{Y^Q (Y^Q)^\mathsf{T}} \right)^2 (X^{D,Q})^\mathsf{T} \qquad (26) \\
&= \tfrac{1}{Q-1} X^{D,Q} \left(I^{Q,Q} - \frac{(Y^Q)^\mathsf{T} Y^Q}{Y^Q (Y^Q)^\mathsf{T}} \right) (X^{D,Q})^\mathsf{T} \qquad (27) \\
&= \tfrac{1}{Q-1} \left(X^{D,Q} (X^{D,Q})^\mathsf{T} - \frac{X^{D,Q} (Y^Q)^\mathsf{T} Y^Q (X^{D,Q})^\mathsf{T}}{Y^Q (Y^Q)^\mathsf{T}} \right).
\end{aligned}$$

In Eq. 26, $I^{Q,Q}$ denotes the $Q \times Q$ identity matrix, and we use in Eq. 27 the fact that the matrix $I^{Q,Q} - (Y^Q)^\mathsf{T} Y^Q / (Y^Q (Y^Q)^\mathsf{T})$ is idempotent, i.e., equal to its square.

Remark 3. We have the following remarkable identity:

$$X^{D,Q}(X^{D,Q})^\mathsf{T} \quad = \quad \hat{\alpha}^D (\hat{\alpha}^D)^\mathsf{T} \, Y^Q (Y^Q)^\mathsf{T} \quad + \quad (Q-1)\hat{\Sigma}.$$

This equation is the non-asymptotic version of Proposition 10.

5.2 Computation of SNRs on the AES Traces from DPA Contest v2 Last Round

The values $\hat{\alpha}^D$ and $\hat{\Sigma}$ are represented in Fig. 3. We obtain:

- $\max_{d=1}^{D} \hat{\alpha}_d^2 / \hat{\Sigma}_{d,d} = 1.69 \cdot 10^{-3}$ (no dimensionality reduction)
- $\mathrm{SNR}_{\mathrm{PCA}} = \frac{((\hat{\alpha}^D)^\mathsf{T} \hat{\alpha}^D)^2}{(\hat{\alpha}^D)^\mathsf{T} \hat{\Sigma} \hat{\alpha}^D} = 1.36 \cdot 10^{-3}$ (PCA)
- $\mathrm{SNR}_{\mathrm{LDA}} = (\hat{\alpha}^D)^\mathsf{T} \hat{\Sigma} \hat{\alpha}^D = 12.78 \cdot 10^{-3}$ (LDA).

Therefore, the LDA has the largest SNR: it is about seven times larger than the maximum sample-wise SNR. The PCA has, in this example, an SNR smaller than the maximum univariable SNR (see Lemma 17).

Interestingly, one can see in Fig. 3 that the noise is locally autoregressive, for instance between samples 107 and 117.

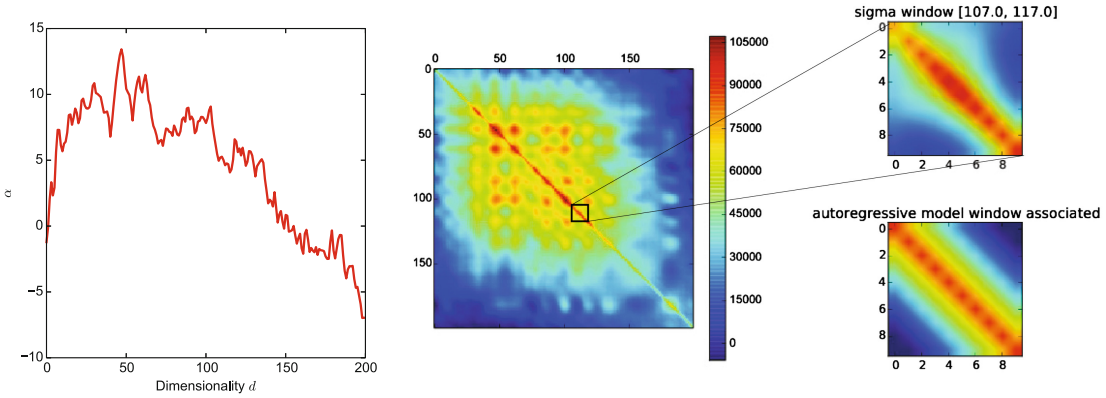

Fig. 3. Estimated $\hat{\alpha}^D$ (*left*) and $\hat{\Sigma}$ (*right*), for $Q = 10,000$ traces

6 Conclusions and Perspectives

Dimensionality reduction is common practice in side-channel analysis. This preprocessing technique has many virtues, such as an elegant multivariate description of the leakages, the concentration of information which reduces the required

number of measurements to extract the key, and the increase of computational efficiency. Nonetheless, as any processing, dimensionality reduction can only reduce some information.

Using a mathematical approach, we have shown that dimensionality reduction is actually part of the optimal attack. This proves rigorously that dimensionality reduction can be achieved without loss in terms of attack success probability in extracting a secret key. As it turns out, the optimal dimensionality reduction consists in a linear projection of the trace samples.

We have also shown that the linear discriminant analysis asymptotically achieves the same projection, and hence becomes optimal as the number of traces increases. When the various samples are weakly correlated, we have found that PCA is nearly equivalent to the optimal dimensionality reduction and to LDA. Thus, in realistic contexts, state-of-the-art dimensionality reduction techniques are actually close to the optimal method.

Finally, we show how to estimate the model parameters (modulation vector α^D and noise covariance matrix Σ), and compute them on a real traces. An SNR gain factor of 7 can be obtained with respect to sample-wise SNR, which stresses the practical interest of dimensionality reduction.

As a perspective, we note that it should also be possible to obtain similar results when the noise is non-Gaussian (e.g., uniform). It is also desirable to compare dimensionality reduction based on linear projections to machine-learning techniques which are also multidimensional, such as SVM, random forests, K-means, etc.

Acknowledgements. The authors would like to thank François-Xavier Standaert for interesting discussions, and also gratefully acknowledge the constructive comments of the reviewers which helped improve the clarity of the paper.

References

1. Archambeau, C., Peeters, E., Standaert, F.-X., Quisquater, J.-J.: Template attacks in principal subspaces. In: Goubin, L., Matsui, M. (eds.) CHES 2006. LNCS, vol. 4249, pp. 1–14. Springer, Heidelberg (2006)
2. Batina, L., Hogenboom, J., van Woudenberg, J.G.J.: Getting more from PCA: first results of using principal component analysis for extensive power analysis. In: Dunkelman, O. (ed.) CT-RSA 2012. LNCS, vol. 7178, pp. 383–397. Springer, Heidelberg (2012)
3. Bhasin, S., Danger, J.-L., Guilley, S., Najm, Z.: Side-channel leakage and trace compression using normalized inter-class variance. In: Proceedings of the Third Workshop on Hardware and Architectural Support for Security and Privacy, HASP 2014, pp. 7:1–7:9. ACM, New York (2014)
4. Brier, E., Clavier, C., Olivier, F.: Correlation power analysis with a leakage model. In: Joye, M., Quisquater, J.-J. (eds.) CHES 2004. LNCS, vol. 3156, pp. 16–29. Springer, Heidelberg (2004)

5. Bruneau, N., Danger, J.-L., Guilley, S., Heuser, A., Teglia, Y.: Boosting higher-order correlation attacks by dimensionality reduction. In: Chakraborty, R.S., Matyas, V., Schaumont, P. (eds.) SPACE 2014. LNCS, vol. 8804, pp. 183–200. Springer, Heidelberg (2014)
6. Chari, S., Rao, J.R., Rohatgi, P.: Template attacks. In: Kaliski, B.S., Koç, K., Paar, C. (eds.) CHES 2002. LNCS, vol. 2523, pp. 13–28. Springer, Heidelberg (2002)
7. Choudary, O., Kuhn, M.G.: Efficient template attacks. In: Francillon, A., Rohatgi, P. (eds.) CARDIS 2013. LNCS, vol. 8419, pp. 253–270. Springer, Heidelberg (2014)
8. Clavier, C., Danger, J.-L., Duc, G., Elaabid, M.A., Gérard, B., Guilley, S., Heuser, A., Kasper, M., Li, Y., Lomné, V., Nakatsu, D., Ohta, K., Sakiyama, K., Sauvage, L., Schindler, W., Stöttinger, M., Veyrat-Charvillon, N., Walle, M., Wurcker, A.: Practical improvements of side-channel attacks on AES: feedback from the 2nd DPA contest. J. Cryptogr. Eng. **4**, 1–16 (2014)
9. Cover, T.M., Thomas, J.A.: Elements of Information Theory, 2nd edn. Wiley-Interscience, New York (2006). ISBN-10: ISBN-10: 0471241954, ISBN-13: 978–0471241959
10. Danger, J.-L., Debande, N., Guilley, S., Souissi, Y.: High-order timing attacks. In: Proceedings of the First Workshop on Cryptography and Security in Computing Systems, CS2 2014, pp. 7–12. ACM, New York (2014)
11. Debande, N., Souissi, Y., Elaabid, M.A., Guilley, S., Danger, J.-L.: Wavelet transform based pre-processing for side channel analysis. In: 45th Annual IEEE/ACM International Symposium on Microarchitecture, MICRO 2012, Workshops Proceedings, Vancouver, BC, Canada, 1–5 December 2012, pp. 32–38. IEEE Computer Society (2012)
12. Durvaux, F., Standaert, F.-X., Veyrat-Charvillon, N., Mairy, J.-B., Deville, Y.: Efficient selection of time samples for higher-order DPA with projection pursuits. In: Mangard, S., Poschmann, A.Y. (eds.) COSADE 2015. LNCS, vol. 9064, pp. 34–50. Springer, Heidelberg (2015). http://eprint.iacr.org/2014/412
13. Fei, Y., Luo, Q., Ding, A.A.: A statistical model for DPA with novel algorithmic confusion analysis. In: Prouff, E., Schaumont, P. (eds.) [25], pp. 233–250
14. Gierlichs, B., Lemke-Rust, K., Paar, C.: Templates vs. stochastic methods. In: Goubin, L., Matsui, M. (eds.) CHES 2006. LNCS, vol. 4249, pp. 15–29. Springer, Heidelberg (2006)
15. Guilley, S., Chaudhuri, S., Sauvage, L., Hoogvorst, P., Pacalet, R., Bertoni, G.M.: Security evaluation of WDDL and SecLib countermeasures against power attacks. IEEE Trans. Comput. **57**(11), 1482–1497 (2008)
16. Hajra, S., Mukhopadhyay, D.: Multivariate leakage model for improving nonprofiling DPA on noisy power traces. In: Lin, D., Xu, S., Yung, M. (eds.) Inscrypt 2013. LNCS, vol. 8567, pp. 325–342. Springer, Heidelberg (2014)
17. Hajra, S., Mukhopadhyay, D.: SNR to success rate: reaching the limit of nonprofiling DPA. Cryptology ePrint Archive, Report 2013/865 (2013). http://eprint.iacr.org/2013/865/
18. Hajra, S., Mukhopadhyay, D.: On the optimal pre-processing for non-profiling differential power analysis. In: Prouff, E. (ed.) COSADE 2014. LNCS, vol. 8622, pp. 161–178. Springer, Heidelberg (2014)
19. Hajra, S., Mukhopadhyay, D.: Reaching the limit of nonprofiling DPA. IEEE Trans. CAD Integr. Circ. Syst. **34**(6), 915–927 (2015)
20. Heuser, A., Rioul, O., Guilley, S.: Good is not good enough. In: Batina, L., Robshaw, M. (eds.) CHES 2014. LNCS, vol. 8731, pp. 55–74. Springer, Heidelberg (2014)

21. Karsmakers, P., Gierlichs, B., Pelckmans, K., De Cock, K., Suykens, J., Preneel, B., De Moor, B.: Side channel attacks on cryptographic devices as a classification problem. COSIC technical report (2009)
22. Kocher, P.C., Jaffe, J., Jun, B.: Differential power analysis. In: Wiener, M. (ed.) CRYPTO 1999. LNCS, vol. 1666, pp. 388–397. Springer, Heidelberg (1999)
23. Oswald, D., Paar, C.: Improving side-channel analysis with optimal linear transforms. In: Mangard, S. (ed.) CARDIS 2012. LNCS, vol. 7771, pp. 219–233. Springer, Heidelberg (2013)
24. Oswald, E., Mangard, S., Herbst, C., Tillich, S.: Practical second-order DPA attacks for masked smart card implementations of block ciphers. In: Pointcheval, D. (ed.) CT-RSA 2006. LNCS, vol. 3860, pp. 192–207. Springer, Heidelberg (2006)
25. Prouff, E., Schaumont, P. (eds.): CHES 2012. LNCS, vol. 7428. Springer, Heidelberg (2012)
26. Renauld, M., Standaert, F.-X., Veyrat-Charvillon, N., Kamel, D., Flandre, D.: A formal study of power variability issues and side-channel attacks for nanoscale devices. In: Paterson, K.G. (ed.) EUROCRYPT 2011. LNCS, vol. 6632, pp. 109–128. Springer, Heidelberg (2011)
27. Reparaz, O., Gierlichs, B., Verbauwhede, I.: Selecting time samples for multivariate DPA attacks. In: Prouff, E., Schaumont, P. (eds.) [25], pp. 155–174
28. Satoh, A.: Side-channel Attack Standard Evaluation Board, SASEBO-GII. Project of the AIST - RCIS (Research Center for Information Security). http://www.rcis.aist.go.jp/special/SASEBO/SASEBO-GII-en.html. Accessed 31 May 2015
29. Souissi, Y., Debande, N., Mekki, S., Guilley, S., Maalaoui, A., Danger, J.-L.: On the optimality of correlation power attack on embedded cryptographic systems. In: Askoxylakis, I., Pöhls, H.C., Posegga, J. (eds.) WISTP 2012. LNCS, vol. 7322, pp. 169–178. Springer, Heidelberg (2012)
30. Souissi, Y., Nassar, M., Guilley, S., Danger, J.-L., Flament, F.: First principal components analysis: a new side channel distinguisher. In: Rhee, K.-H., Nyang, D.H. (eds.) ICISC 2010. LNCS, vol. 6829, pp. 407–419. Springer, Heidelberg (2011)
31. Standaert, F.-X., Archambeau, C.: Using subspace-based template attacks to compare and combine power and electromagnetic information leakages. In: Oswald, E., Rohatgi, P. (eds.) CHES 2008. LNCS, vol. 5154, pp. 411–425. Springer, Heidelberg (2008)
32. Strobel, D., Oswald, D., Richter, B., Schellenberg, F., Paar, C.: Microcontrollers as (in)security devices for pervasive computing applications. Proc. IEEE **102**(8), 1157–1173 (2014)
33. Sugawara, T., Homma, N., Aoki, T., Satoh, A.: Profiling attack using multivariate regression analysis. IEICE Electron. Express **7**(15), 1139–1144 (2010)
34. TELECOM ParisTech. DPA Contest, 2nd edition. http://www.DPAcontest.org/v2/. Accessed 31 May 2015