Side channel attacks on cryptographic devices as a classification problem

Peter Karsmakers[1,2], Benedikt Gierlichs[3], Kristiaan Pelckmans[2], Katrien De Cock[2],
Johan Suykens[2], Bart Preneel[3], Bart De Moor[2]

[1] K. H. Kempen (Associatie K.U.Leuven), IIBT
Kleinhoefstraat 4, B-2440 Geel, Belgium
[2] K.U.Leuven, ESAT-SCD/SISTA, [3] K.U.Leuven, ESAT-SCD/COSIC
Kasteelpark Arenberg 10, B-3001, Leuven, Belgium
e-mail: name.surname@esat.kuleuven.be

Abstract

In this contribution we examine three data reduction techniques in the context of Template Attacks. The Template Attack is a powerful two-step side channel attack which models an almost omnipotent adversary in the profiling step, but restricts him to a single observation in the classification step. The profiling step requires data reduction due to computational complexity and vast amounts of data. Here we examine the inter class variance, the Spearman correlation coefficient, and principal component analysis. The classification step requires a distinguisher, which we implemented by linear discriminant analysis. Our results lead to the conclusion that PCA in combination with LDA gives the highest classification accuracies on unseen data from the tried linear classifier methods.

1. Introduction

Secure cryptographic algorithms are what is noted as black-box secure, *i.e.* an adversary cannot gather information from observing the inputs and/or outputs of the algorithm. However, in this vision an algorithm is a purely abstract mathematical object.

To satisfy nowadays great demand for instant secure electronic communication, secure embedded devices such as mobile phones and PDAs, and secure financial and identity tokens, *e.g.* banking cards, SIM cards, identity cards, cryptographic algorithms are implemented in electronic devices. In the last decade a whole new class of attacks, not against cryptographic algorithms but against their physical implementations, has received much attention: side channel attacks.

A side channel is formed by the physical realization of a cryptographic algorithm. It exists due to the fact that the electronic device has a certain influence on physical observables in its vicinity. For example: an electronic device emits electromagnetic radiation while processing and dissipates a certain amount of power. Since these physical observables depend on the data words processed by the device which in turn depend on secret information, *e.g.* cryptographic keys, a side channel leaks sensitive information. Side channel attacks aim at exploiting this information leakage to reveal the secret.

The Template Attack [1] is a so called two-step side channel attack. During the first step, an adversary has full access to and control over a training device which he uses to build templates. More precisely he builds a template, *i.e.* a characterization of the typical behavior of the side channel, for a certain set of instructions and/or data words. In the second step, the adversary has access to only a single observation of the side channel and uses the prior built templates to deduce, which instruction respectively data word has been processed by the target device.

The remainder of the paper is organized as follows. In Section 2 we introduce the two steps of our template attack: the classification method and the dimensionality reduction techniques. Section 3 describes the experiments and the classification results. In Section 4 we conclude the paper.

2. Template Attack

In this section we explain how to use Linear Discriminant Analysis (LDA) in the context of template attacks. It is assumed that the secret key information leakage is mainly hidden in the local variability of the mean time series. It is therefore appropriate to work only in a subspace of the original input space. Therefore, we examined three different dimensionality reduction techniques in combination with LDA.

In Section 2.1 we explain Linear Discriminant Analysis. The reduction techniques are discussed in Section 2.2.

2.1. Linear Discriminant Analysis

After introducing some notations, we recall the principles of linear discriminant analysis [3]. Suppose we have a multi-class problem with C classes ($C \geq 2$) with a training set $\{(x_i, y_i)\}_{i=1}^{N} \subset \mathbb{R}^d \times \{1, 2, ..., C\}$ with N samples, where input samples x_i are i.i.d. from an unknown probability distribution over the random vectors $(\mathbf{X}, \mathbf{Y})$. Suppose $f_c(x)$ is the class-conditional density of X in class $Y = c$, denoted as $\Pr(\mathbf{Y} = c|\mathbf{X} = x)$, and let π_c be the prior probability of class c, with $\sum_{c=1}^{C} \pi_c = 1$. A simple application of Bayes theorem gives us

$$\Pr(\mathbf{Y} = c|\mathbf{X} = x) = \frac{f_c(x)\pi_c}{\sum_{l=1}^{C} f_l(x)\pi_l}. \quad (1)$$

In LDA we model each class density as a multivariate gaussian

$$f_c(x) = \frac{1}{(2\pi)^{d/2}|\Sigma_c|^{1/2}} e^{-\frac{1}{2}(x-\mu_c)^T \Sigma_c^{-1}(x-\mu_c)}. \quad (2)$$

The covariance matrix of the different classes is assumed to be equal in LDA, $\Sigma_c = \Sigma, \forall c$. In comparing two classes c and l, it

is sufficient to look at the log-ratio,

$$\begin{aligned}\ln \frac{\Pr(\mathbf{Y}=c|\mathbf{X}=x)}{\Pr(\mathbf{Y}=l|\mathbf{X}=x)} &= \ln \frac{f_c(x)}{f_l(x)} + \ln \frac{\pi_c}{\pi_l} \\ &= \ln \frac{\pi_c}{\pi_l} - \frac{1}{2}(\mu_c+\mu_l)^T \Sigma^{-1}(\mu_c-\mu_l) + \\ & x^T \Sigma^{-1}(\mu_c-\mu_l).\end{aligned} \tag{3}$$

It is seen that this equation is linear in x. The equal covariance matrices cause the normalization factors to cancel, as well as the quadratic part in the exponents. This log-odds function implies that the decision boundary between any two classes c and l is linear. From (3) we obtain the linear discriminant functions

$$\delta_c(x) = x^T \Sigma^{-1} \mu_c - \frac{1}{2} \mu_c^T \Sigma^{-1} \mu_c + \ln \pi_c, \tag{4}$$

for $c = 1, ..., C$.

Using these functions we can define the classification rule

$$\underset{c \in 1,...,C}{\arg\max}\, \delta_c(x). \tag{5}$$

In practice the parameters of the Gaussian distributions are not known and have to be estimated using the training data. The empirical mean, covariance and prior are defined as follows

$$\begin{cases} \hat{\mu}_c = \sum_{l \in \mathcal{D}_c} \frac{x_l}{N_c} \\ \hat{\Sigma} = \sum_{c=1}^{C} \sum_{l \in \mathcal{D}_c} \frac{(x_i - \hat{\mu}_c)(x_i - \mu_c)^T}{N-C} \\ \hat{\pi}_c = \frac{N_c}{N}, \end{cases} \tag{6}$$

where N_c is the number of observations of class c and $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup ... \cup \mathcal{D}_C$, $\mathcal{D}_i \cap \mathcal{D}_j = \varnothing, \forall i \neq j$ and $y_i = c, x_i \in \mathcal{D}_c$.

2.2. Dimensionality reduction

To retain sufficient side channel information from the recording device, which usually has high clock rates, the number of samples d per time series is large. This leads to excessive computational loads and large memory requirements. However, as previously said, the expected number of relevant time samples is limited. We have tried three different dimensionality reduction methods. The first is to select time samples showing the largest difference between the mean time series vectors, the second uses Spearman rank correlation and the third does a dimensionality reduction via principal component analysis.

2.2.1. Mean class variances

A first simple rule proposed by [1] is to select time samples which show the largest difference between the class mean time series vectors.

2.2.2. Spearman correlation

The Spearman rank correlation test investigates the correlation on the basis of the ordinal rank score of two independent variables [5]. The goal is to verify how significantly dependent the scores of the two variables are. This is expressed by Spearman's rank correlation coefficient

$$\rho = 1 - \frac{6 \sum_i t_i^2}{N(N^2-1)}, \tag{7}$$

where t_i is the difference between each rank of corresponding values of x and y.

2.2.3. Principal Component Analysis

A well-known and frequently used technique for dimensionality reduction is linear Principal Components Analysis (PCA) [4]. Suppose one wants to map vectors $x \in \mathbb{R}^d$ into lower dimensional vectors $z \in \mathbb{R}^m$ with $m < n$. One proceeds then by estimating the covariance matrix $\hat{\Sigma}$ of all training data and computes the eigenvalue decomposition

$$\hat{\Sigma} u_i = \lambda_i u_i. \tag{8}$$

By selecting the m largest eigenvalues and the corresponding eigenvectors, one obtains the transformed variables (score variables)

$$z_i = u_i^T (x - \mu), \tag{9}$$

for $i = 1, ..., m$. One has to note, however, that these transformed variables are no longer real physical variables. The error $\sum_{i=m+1}^{d} \lambda_i$ resulting from the dimensionality reduction is determined by the values of the neglected components.

3. Experiments

Our experimental platform is an 8-bit ATmega163 micro controller which performs AES-128 (also known as Rijndael) [2] encryption in software. Our side channel measurements represent the voltage drop over a 50Ω resistor inserted in the chip's ground line. We sample the power dissipation during the first round of AES-128 encryption at a sampling frequency of 200MS/s.

For the profiling step, we stored an AES key k_1 in the device and obtained a set of 20.000 measurements from the encryption of uniformly chosen random plaintexts. For the profiling step, we stored a different key k_2 in the device and obtained a set of 500 measurements from the encryption of uniformly chosen random plaintexts. As intermediate result, our attacks focus on the Sbox output for the first byte of the AES state in the first round, denoted by the random variable $\mathbf{X}$. Accordingly, the voltage drop over the resistor at one specific sampling point is denoted by $\mathbf{Y}$.

Table 1 shows the classification accuracies when using the three different dimensionality reduction techniques as explained before in cooperation with the LDA classifier. For each of these techniques we have to empirically determine the number of selected dimensions (m) (time instants or principal components). In order to tune this m we divided our measurements set in a training set, which consists of 15,000 data points, and validation set, including 5,000 data points, and select the m which gives the highest classification accuracy on the validation set. For the mean class variance dimensionality reduction (Section 2.2.1) we retained the 300 time instants with highest variance within the class means (see Fig. 1). Using Spearman's method (Section 2.2.2) we selected the 1000 time instants with the highest correlation coefficients (see Fig. 2). In Fig. 3 the classification accuracies in function of the number of selected principal components (Section 2.2.3) are shown. The classification accuracies in the figure are those on the validation set. From the figure we see that a dimensionality reduction from 9,000 to 400 seems to produce good results.

4. Conclusions

In this paper we presented LDA in cooperation with three different dimensionality reduction techniques for the task of template attacks. In our experiments PCA in combination

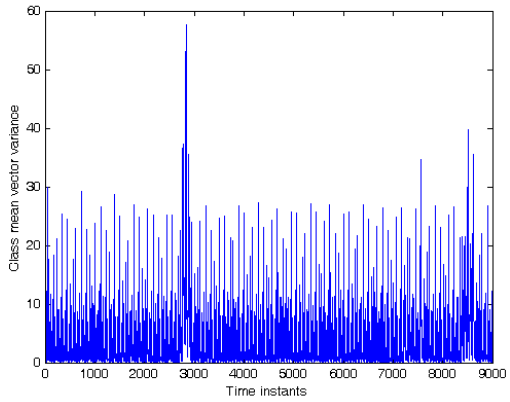

Figure 1: The variance between the class means of each separate time instant.

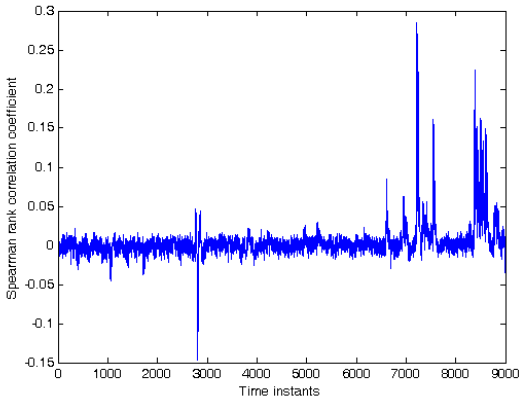

Figure 2: The Spearman rank correlation coefficients of the input vectors and the class labels for each separate time instant.

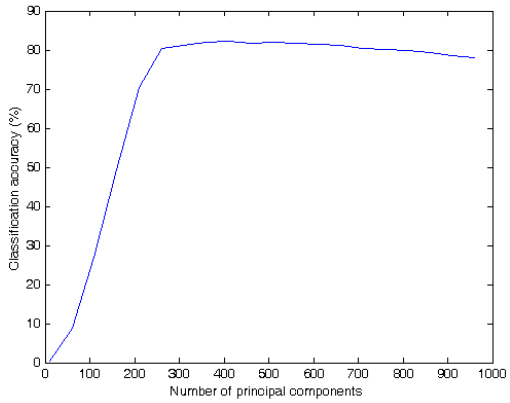

Figure 3: Classification accuracy of LDA on test set of 5,000 time series not included in the training process, which consists of 15,000 input vectors, in function of the number principal components.

	acc (%)	**10-best (%)**
PCA	28.4	74.4
MCVAR	28	73
SPEARMAN	5.8	34.6

Table 1: LDA classification accuracies on the test set of 500 unseen measurements with three different dimensionality reduction techniques. The acronym PCA stands for Principal Component Analysis (Section 2.2.3), MCVAR stands for Mean class variances (Section 2.2.1) and SPEARMAN for Spearman's rank correlation (Section 2.2.2). The column **acc** gives the percentage of correctly classified measurements. The percentages in the **10-best** column are equal to the proportion of measurements for which the correct class was one of the 10 most probable classes.

with LDA gives the highest classification accuracies on unseen data. In the future we will examine the use of Support Vector Machines on side-channel data for template attacks because in many different application areas this technique is known to produce good classification results.

Acknowledgements

Bart De Moor and Bart Preneel are full professors and Johan Suykens is a professor at the Katholieke Universiteit Leuven, Belgium.
Research supported by

- Research Council KUL: GOA AMBioRICS, CoE EF/05/006 Optimization in Engineering, several PhD/postdoc & fellow grants;
- Flemish Government:
  - FWO: PhD/postdoc grants, projects, G.0407.02 (support vector machines), G.0197.02 (power islands), G.0141.03 (Identification and cryptography), G.0491.03 (control for intensive care glycemia), G.0120.03 (QIT), G.0452.04 (new quantum algorithms), G.0499.04 (Statistics), G.0211.05 (Nonlinear), G.0226.06 (cooperative systems and optimization), G.0321.06 (Tensors), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM);
  - IWT: PhD Grants, McKnow-E, Eureka-Flite2
- Belgian Federal Science Policy Office: IUAP P6/04 (Dynamical systems, control and optimization, 2007-2011) ;
- EU: ERNSI;

5. References

[1] S. Chari, J. R. Rao, P. Rohatgi, "Template Attacks", *4th International Workshop on Cryptographic Hardware and Embedded Systems*, vol. 2523, 2002

[2] J. Daemen, V. Rijmen, "Rijndael for AES", *3rd Conference on the Advanced Encryption Standard (AES)*, 5 pages, 2000.

[3] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer, 2001.

[4] I. T. Jolliffe, *Principal Component Analysis*, Springer-Verlag,1986.

[5] E. L. Lehmann, H. J. D'Abrera, *Nonparametrics: Statistical Methods Based on Ranks*, Prentice-Hall, 1998.