Less is More
Dimensionality Reduction from a Theoretical Perspective

Nicolas Bruneau, Sylvain Guilley, Annelie Heuser, Damien Marion, and Olivier Rioul
About us...

Nicolas BRUNEAU
is also with

Sylvain GUILLEY
is also with

Annelie HEUSER
is PhD fellow at

Damien MARION
is also with

Olivier RIOUL
is also Prof at
Overview

Introduction
  Motivation
  State-of-the-Art & Contribution
  Notations and Model

Optimal..
  ..distinguisher
  ..dimension reduction

Comparison to..
  ..PCA
  ..LDA

Numerical Comparison

Practical Validation

Conclusion
Overview

Introduction
  Motivation
  State-of-the-Art & Contribution
Notations and Model

Optimal..
  ..distinguisher
  ..dimension reduction

Comparison to..
  ..PCA
  ..LDA

Numerical Comparison
Practical Validation
Conclusion
Motivation

large number of samples/ points of interest
Motivation

Problem (profiled and non-profiled side-channel distinguisher)

How to reduce dimensionality of multi-dimensional measurements?
Motivation

Problem (*profiled* and *non-profiled* side-channel distinguisher)

How to reduce dimensionality of multi-dimensional measurements?

Wish list

- simplification of the problem
- concentration of the information (to distinguish using fewer traces)
- improvement of the computational speed
State-of-the-Art I

Selection of points of interest

- manual selection of educated guesses [Oswald et al., 2006]
- automated techniques: sum-of-square differences (SOSD) and t-test (SOST) [Gierlichs et al., 2006]
- wavelet transforms [Debande et al., 2012]
State-of-the-Art I

Selection of points of interest

- Manual selection of educated guesses [Oswald et al., 2006]
- Automated techniques: sum-of-square differences (SOSD) and t-test (SOST) [Gierlichs et al., 2006]
- Wavelet transforms [Debande et al., 2012]

Leakage detection metrics

- ANOVA (e.g. [Choudary and Kuhn, 2013, Danger et al., 2014]) or [Bhasin et al., 2014] (Normalized Inter-Class Variance (NICV))
State-of-the-Art II

Principal Component Analysis

- compact templates in [Archambeau et al., 2006]
- reduce traces in [Batina et al., 2012]
- eigenvalues as a security metric [Guilley et al., 2008]
- eigenvalues as a distinguisher [Souissi et al., 2010]
State-of-the-Art II

Principal Component Analysis

- compact templates in [Archambeau et al., 2006]
- reduce traces in [Batina et al., 2012]
- eigenvalues as a security metric [Guilley et al., 2008]
- eigenvalues as a distinguisher [Souissi et al., 2010]

Easily and accurately computed with no divisions involved.

Maximizing inter-class variance, but not intra-class variance.
State-of-the-Art II

Linear Discriminant Analysis

- improved alternative
- takes inter-class variance and intra-class variance into account
- empirical comparisons [Standaert and Archambeau, 2008, Renauld et al., 2011, Strobel et al., 2014]


not easily and accurately computed with no divisions involved

maximizing inter-class variance and intra-class variance
Linear Discriminant Analysis

- improved alternative
- takes inter-class variance and intra-class variance into account
- empirical comparisons [Standaert and Archambeau, 2008, Renauld et al., 2011, Strobel et al., 2014]

But..

- advantages due to the statistical tools, their implementation, data set ...
- no clear rationale to prefer one method!
Contribution

- dimensional reduction in SCA from a theoretical viewpoint
- assuming attacker has full knowledge of the leakage
- derivation of the optimal dimensionality reduction

“Less is more”

Advantages of dimensionality reduction can come with no impact on the attack success probability!

- comparison to PCA and LDA: theoretically and practically
Notations

- unknown secret key $k^*$, key byte hypothesis $k$
- $D$ different samples, $d = 1, \ldots, D$
- $Q$ different traces/queries, $q = 1, \ldots, Q$
- matrix notation $M^{D,Q}$ ($D$ rows, $Q$ columns)
- leakage function $\varphi$
- sensitive variable: $Y_q(k) = \varphi(T_q \oplus k)$ (normalized variance $\forall q$)
Model

- trace
  \[ X_{d,q} = \alpha_d Y_q(k^*) + N_{d,q} \]

- traces
  \[ X^{D,Q} = \alpha^{D,Q} Y^Q(k^*) + N^{D,Q} \]

- noise: zero-mean Gaussian distribution, covariance \( \Sigma \)
  independent of \( q \) but can be correlated among \( d \)
Overview

Introduction
Motivation
State-of-the-Art & Contribution
Notations and Model

Optimal..
..distinguisher
..dimension reduction

Comparison to..
..PCA
..LDA

Numerical Comparison

Practical Validation

Conclusion
Optimal distinguisher

Data processing theorem [Cover and Thomas, 2006]

Any preprocessing like dimensionality reduction can only decrease information.

- optimal means optimizing the success rate
- known leakage model: optimal attack $\Rightarrow$ template attack
- maximum likelihood principle
Optimal distinguisher

Data processing theorem [Cover and Thomas, 2006]

Any preprocessing like dimensionality reduction can only decrease information.

- optimal means optimizing the success rate
- known leakage model: optimal attack $\Rightarrow$ template attack
- maximum likelihood principle

Given:
- $Q$ traces of dimensionality $D$ in a matrix $x^D,Q$
- for each trace $x^D_q$: a plaintext/ciphertext $t_q$
Optimal distinguisher

\[ D(x^D, Q, t^Q) = \arg \max_k p(x^D, Q | t^Q, k^* = k) \]

\[ = \arg \max_k p_{N^D, Q} (x^D, Q - \alpha^D y^Q(k)) \]

\[ = \arg \max_k \prod_{q=1}^{Q} p_{N^D_q} (x^D_q - \alpha^D y_q(k)) \]

where

\[ p_{N^D_q}(z^D) = \frac{1}{\sqrt{(2\pi)^D | \det \Sigma|}} \exp\left( -\frac{1}{2} (z^D)^T \Sigma^{-1} z^D \right). \]
Optimal dimension reduction

Theorem

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:

$$\tilde{x}_q = (\alpha^D)^T \Sigma^{-1} x^D_q$$

$(q = 1, \ldots, Q)$. 
Optimal dimension reduction

Theorem

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:

$$\tilde{x}_q = (\alpha^D)^T \Sigma^{-1} x_q^D \quad (q = 1, \ldots, Q).$$

**scalar** = column $D \cdot D \times D \cdot \text{row } D$
Proof I

- taking the logarithm, the optimal distinguisher $D(x^D, t^Q)$ rewrites

$$D(x^D, t^Q) = \arg \min_k \sum_{q=1}^{Q} (x^D_q - \alpha^D y_q(k))^T \Sigma^{-1} (x^D_q - \alpha^D y_q(k)).$$
Proof I

- taking the logarithm, the optimal distinguisher $D(x^D, Q, t^Q)$ rewrites

$$D(x^D, Q, t^Q) = \arg \min_k \sum_{q=1}^{Q} (x^D_q - \alpha^D y_q(k))^T \Sigma^{-1} (x^D_q - \alpha^D y_q(k)).$$

- expansion gives

$$\underbrace{(x^D_q)^T \Sigma^{-1} x^D_q}_{\text{cst. } C \text{ independent of } k} - 2(\alpha^D)^T y_q(k) \Sigma^{-1} x^D_q + (y_q(k))^2 (\alpha^D)^T \Sigma^{-1} \alpha^D$$

$$= C - 2y_q(k)[(\alpha^D)^T \Sigma^{-1} x^D_q] + (y_q(k))^2 [(\alpha^D)^T \Sigma^{-1} \alpha^D]$$

$$= [(\alpha^D)^T \Sigma^{-1} \alpha^D] \left( y_q(k) - \frac{(\alpha^D)^T \Sigma^{-1} x^D_q}{(\alpha^D)^T \Sigma^{-1} \alpha^D} \right)^2 + C'.$$
Proof II

so, for $D(x^D, Q, t^Q)$ we obtain

$$D(x^D, Q, t^Q) = \arg\min_k \sum_{q=1}^{Q} \left( y_q(k) - \frac{\alpha^T D \Sigma^{-1} x_q^D}{(\alpha^T D \Sigma^{-1} \alpha^D)} \right)^2 \left[ (\alpha^T D \Sigma^{-1} \alpha^D) \right]$$

$$= \arg\min_k \sum_{q=1}^{Q} \frac{\tilde{x}_q - y_q(k))^2}{\tilde{\sigma}^2},$$

where

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

Optimal dimension reduction

Optimal distinguisher can be computed either:
- on multivariate traces $x_q^D$, with a noise covariance matrix $\Sigma$
- on monovariate traces $\tilde{x}_q$, with scalar noise of variance $\tilde{\sigma}^2$. 
Optimal dimension reduction

Optimal distinguisher can be computed either:
- on multivariate traces $x_q^D$, with a noise covariance matrix $\Sigma$
- on monovariate traces $\tilde{x}_q$, with scalar noise of variance $\tilde{\sigma}^2$.

- optimal dimensionality reduction does not depend on the distribution of $Y^D(k)$
- also not on the confusion coefficient [Fei et al., 2012]
- only on the signal weights $\alpha^D$ and on the noise covariance $\Sigma$
Corollary

After optimal dimensionality reduction, the signal-noise-ratio is given by

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

Examples

- **white noise:**
  \[ \hat{\text{SNR}} = \sum_{d=1}^{D} \text{SNR}_d \]

- **autoregressive noise**
  (confirmed on dpacontest v2)
Overview

Introduction
Motivation
State-of-the-Art & Contribution
Notations and Model
Optimal..
  ..distinguisher
  ..dimension reduction
Comparison to..
  ..PCA
  ..LDA
Numerical Comparison
Practical Validation
Conclusion
Comparison to PCA

Classical PCA

- centered data $M_{d,q} = X_{d,q} - \frac{1}{Q} \sum_{q'=1}^{Q} X_{d,q'} \ (1 \leq q \leq Q, 1 \leq d \leq D)$
- directions of PCA: eigenvectors of $M_{D,Q}^{D,Q} (M_{D,Q}^{D,Q})^T$
- drawback: depends both on data and noise

Inter-class PCA [Archambeau et al., 2006]

- centered column
- $\sum_{1 \leq q \leq Q} Y_q = y \sum_{1 \leq q \leq Q} Y_q = y X_D q$
- takes into account the sensitive variable $Y$
- noise is averaged away
Comparison to PCA

Classical PCA

- centered data \( M_{d,q} = X_{d,q} - \frac{1}{Q} \sum_{q'=1}^{Q} X_{d,q'} \) \((1 \leq q \leq Q, 1 \leq d \leq D)\)
- directions of PCA: eigenvectors of \( M_{D,Q} (M_{D,Q})^T \)
- drawback: depends both on data and noise

Inter-class PCA [Archambeau et al., 2006]

- centered column \( \sum_{1 \leq q \leq Q} \frac{1}{Y_q=y} \sum_{1 \leq q \leq Q} X_q^D \)
- takes into account the sensitive variable Y
- noise is averaged away
Comparison to PCA

For classical PCA

Asymptotically as $Q \rightarrow +\infty$,

$$
\frac{1}{Q} M_{D,Q}^D (M_{D,Q}^D)^T \rightarrow \alpha_D^D (\alpha_D^D)^T + \Sigma.
$$

Eigenvectors?
Comparison to PCA

For classical PCA

Asymptotically as \( Q \rightarrow +\infty \),

\[
\frac{1}{Q} M^{D,Q} (M^{D,Q})^T \rightarrow \alpha^D (\alpha^D)^T + \Sigma.
\]

Eigenvectors?

Proposition

Asymptotically, Inter-class PCA has only one principal direction, namely the vector \( \alpha^D \).
Comparison to PCA

**Proposition**

The asymptotic SNR after projection using Inter-class PCA is equal to

\[
\frac{\|\alpha^D\|^4}{(\alpha^D)^T \Sigma \alpha^D}.
\]
Proposition

The asymptotic SNR after projection using Inter-class PCA is equal to
\[ \frac{\|\alpha^D\|^4}{(\alpha^D)^T \Sigma \alpha^D}. \]

Theorem

The SNR of the asymptotic Inter-class PCA is smaller than the SNR of the optimal dimensionality reduction.
Comparison to PCA

**Proposition**

The asymptotic SNR after projection using Inter-class PCA is equal to

\[ \frac{\| \alpha^D \|^4_2}{(\alpha^D)^T \Sigma \alpha^D}. \]

**Theorem**

The SNR of the asymptotic Inter-class PCA is smaller than the SNR of the optimal dimensionality reduction.

**Corollary**

The asymptotic Inter-class PCA has the same SNR as the optimal dimensionality reduction if and only if \( \alpha^D \) is an eigenvector of \( \Sigma \). In this case, both dimensionality reductions are equivalent.
Comparison to LDA

- Computes the eigenvectors of $S_w^{-1}S_b$
- $S_w$ is the *intra-class scatter matrix*, asymptotically equal to $\Sigma$
- $S_b$ is the *inter-class scatter matrix*, equal to $\alpha^D (\alpha^D)^T$.

**Proposition**

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

- Computes the eigenvectors of $S_w^{-1}S_b$
- $S_w$ is the *intra-class scatter matrix*, asymptotically equal to $\Sigma$
- $S_b$ is the *inter-class scatter matrix*, equal to $\alpha^D (\alpha^D)^\top$.

**Proposition**

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

**Theorem**

The asymptotic LDA computes exactly the optimal dimensionality reduction.
Asymptotic PCA and LDA

- \( D = 6 \) for autoregressive noise with \( \sigma = 1 \) and different \( \rho \)

(a) Equal \( \text{SNR}_d = 1 \), \( 1 \leq d \leq D \)

\[
\alpha^D = (1, 1, 1, 1, 1, 1)^T
\]

(b) Varying \( \text{SNR}_d \), \( 1 \leq d \leq D \)

\[
\alpha^D = \sqrt{6.0/6.4} \cdot (1.0, 1.1, 1.2, 1.3, 0.9, 0.5)^T
\]
Overview

Introduction
Motivation
State-of-the-Art & Contribution
Notations and Model

Optimal..
..distinguisher
..dimension reduction

Comparison to..
..PCA
..LDA

Numerical Comparison

Practical Validation

Conclusion
Practical Validation

- **DPA CONTEST v2**, one clock cycle $D = 200$
- normalized Hamming weight
- precharacterization of the model parameter $\alpha^D$ and $\Sigma$ (details in the paper)

\[
\max_{d=1}^D \frac{\hat{\alpha}_d^2}{\hat{\Sigma}_{d,d}} = 1.69 \cdot 10^{-3} 
\]  
(no dimensionality reduction)

- $\text{SNR}_{\text{PCA}} = \frac{(\hat{\alpha}^D)^T \hat{\alpha}^D}{\hat{\alpha}^D)^T \hat{\Sigma} \hat{\alpha}^D} = 1.36 \cdot 10^{-3}$  
(PCA)

- $\text{SNR}_{\text{LDA}} = (\hat{\alpha}^D)^T \hat{\Sigma} \hat{\alpha}^D = 12.78 \cdot 10^{-3}$  
(LDA)
Overview

Introduction
  Motivation
  State-of-the-Art & Contribution
Notations and Model
Optimal...
  ..distinguisher
  ..dimension reduction
Comparison to...
  ..PCA
  ..LDA
Numerical Comparison
Practical Validation
Conclusion
Conclusion and Perspectives

Optimal dimension reduction...

- is part of the optimal attack
- can be achieved *without* losing success probability

LDA asymptotically achieves the same projection as optimal when weakly correlated ($\Sigma$ is identity matrix). PCA is nearly equivalent to optimal/LDA. *Extend to non-Gaussian noise*.

Comparison to machine-learning techniques.
Conclusion and Perspectives

Optimal dimension reduction...

- is part of the optimal attack
- can be achieved \textit{without} losing success probability
  
- LDA asymptotically achieves the same projection as optimal
- when weakly correlated (\(\Sigma\) is identity matrix)
  
  PCA is nearly equivalent to optimal/ LDA
Conclusion and Perspectives

Optimal dimension reduction...

- is part of the optimal attack
- can be achieved \textit{without} losing success probability

- LDA asymptotically achieves the same projection as optimal
- when weakly correlated (\( \Sigma \) is identity matrix) PCA is nearly equivalent to optimal/ LDA

★ extend to non-Gaussian noise
★ comparison to machine-learning techniques
Thank you!
Template Attacks in Principal Subspaces.
In CHES, volume 4249 of LNCS, pages 1–14. Springer.
Yokohama, Japan.

Getting more from PCA: first results of using principal component analysis for extensive power analysis.
Side-channel Leakage and Trace Compression Using Normalized Inter-class Variance.

Efficient template attacks.
References III

Elements of Information Theory.
Wiley-Interscience.

High-order timing attacks.

Wavelet transform based pre-processing for side channel analysis.
References IV

A Statistical Model for DPA with Novel Algorithmic Confusion Analysis.

Templates vs. Stochastic Methods.
In CHES, volume 4249 of LNCS, pages 15–29. Springer.
Yokohama, Japan.

Security Evaluation of WDDL and SecLib Countermeasures against Power Attacks.
References V

Practical Second-Order DPA Attacks for Masked Smart Card Implementations of
Block Ciphers.

A formal study of power variability issues and side-channel attacks for nanoscale
devices.
In Paterson, K. G., editor, Advances in Cryptology - EUROCRYPT 2011 - 30th
Annual International Conference on the Theory and Applications of Cryptographic
References VI

First Principal Components Analysis: A New Side Channel Distinguisher.
In Rhee, K. H. and Nyang, D., editors, ICISC, volume 6829 of Lecture Notes in Computer Science, pages 407–419. Springer.

Using Subspace-Based Template Attacks to Compare and Combine Power and Electromagnetic Information Leakages.
Washington, D.C., USA.
Microcontrollers as (in)security devices for pervasive computing applications.