Multidimensional Extension of Matsui's Algorithm 2

Miia Hermelin[1], Joo Yeon Cho[1], and Kaisa Nyberg[1,2]

[1] Helsinki University of Technology
[2] Nokia Research Center, Finland

Abstract. Matsui's one-dimensional Alg. 2 can be used for recovering bits of the last round key of a block cipher. In this paper a truly multidimensional extension of Alg. 2 based on established statistical theory is presented. Two possible methods, an optimal method based on the log-likelihood ratio and a χ^2-based goodness-of-fit test are compared in theory and by practical experiments on reduced round Serpent. The theory of advantage by Selçuk is generalised in multiple dimensions and the advantages and data, time and memory complexities for both methods are derived.

1 Introduction

Linear cryptanalysis was introduced by Matsui in [1]. The method uses a one-dimensio- nal linear relation for recovering information about the secret key of a block cipher. Matsui presented two algorithms, Algorithm 1 (Alg. 1) and Algorithm 2 (Alg. 2). While Alg.1 extracts one bit of information about the secret key, Alg. 2 ranks several candidates for a part of the last round key of a block cipher according to a test statistic such that the right key should be ranked highest. Using the recovered last round key, it is then possible to extract one bit of information about the other round keys.

Since then researchers have been puzzled by the question how the linear cryptanalysis method could be enhanced by making use of multiple linear approximations simultaneously. In [2] Kaliski and Robshaw used several linear relations involving the same key bits in an attempt to reduce the data complexities of Matsui's algorithms. Multiple linear relations were also used by Biryukov, et al., [3] for extracting several bits of information about the key in an Alg. 1 type attack. This basic attack was also extended to an Alg. 2 type attack. However, both [2] and [3] depend on theoretical assumptions about the statistical properties of the one-dimensional linear relations that may not hold in the general case as was shown in [4].

The statistical linear distinguisher presented by Baignères, et al., in [5] does not suffer from this limitation. It has also another advantage over the previous approaches [2] and [3]: it is based on a well established statistical theory of log-likelihood ratio, LLR, see also [6]. In [7] it was further shown how to distinguish one known probability distribution from a set of other distributions.

O. Dunkelman (Ed.): FSE 2009, LNCS 5665, pp. 209–227, 2009.

The purpose of this paper is to present two new multidimensional extensions of Matsui's Alg. 2 including an effective ranking method for the key candidates based on Selçuk's concept of advantage [8]. First a straightforward solution for Alg. 2 based on goodness-of-fit test using χ^2-statistic will be presented. We will then discuss a χ^2-based version of Alg. 1 [9] and show that the method of Biryukov, et al., is related to a combination of the χ^2-based Alg. 1 and Alg. 2. We will then present a method based on LLR which actually combines Alg. 1 and Alg. 2 and outperforms the χ^2-based method in theory and practice. In the practical experiments the data, memory and time complexity for achieved advantage is determined and compared with the values given by the theoretical statistical models developed in this paper.

The structure of this paper is as follows: In Sect. 2 the basic statistical theory and notation is given. The advantage and the generalisation of Selçuk's theory is presented in Sect. 3. The multidimensional Alg. 2 is described in Sect. 4 and the different methods based on the two test statistics are described in Sect. 5 and Sect. 6. The time, memory and data complexities of both methods are examined in Sect. 7. The experimental results are given in Sect. 8. Finally, Sect. 9 draws conclusions.

2 Boolean Function and Probability Distribution

We will denote the space of n-dimensional binary vectors by V_n. A function $f : V_n \to V_1$ is called a Boolean function. A function $f : V_n \to V_m$ with $f = (f_1, \ldots, f_m)$, where f_i are Boolean functions is called a vector Boolean function of dimension m. A linear Boolean function from V_n to V_m is represented by an $m \times n$ binary matrix U. The m rows of U are denoted by $u_1, \ldots, u_m$, where each u_i is a binary vector of length n.

The correlation between a Boolean function and zero is

$$c(f) = c(f, 0) = 2^{-n} \left(\#\{\xi \in V_n \,|\, f(\xi) = 0\} - \#\{\xi \in V_n \,|\, f(\xi) \neq 0\}\right)$$

and it is also called the correlation of f.

We say that the vector $p = (p_0, \ldots, p_M)$ is a probability distribution (p.d.) of random variable (r.v.) X and denote $X \sim p$, if $\Pr(X = \eta) = p_\eta$, for all $\eta = 0, \ldots, M$. We will denote the uniform p.d. by θ. Let $f : V_n \to V_m$ and $X \sim \theta$. We call the p.d. p of the r.v. $Y = f(X)$ the p.d. of f.

Let us study some general properties of p.d.'s. Let $p = (p_0, \ldots, p_M)$ and $q = (q_0, \ldots, q_M)$ be some p.d.'s of r.v.'s taking on values in a set with $M + 1$ elements. The Kullback-Leibler distance between p and q is defined as follows:

Definition 1. *The* relative entropy *or* Kullback-Leibler distance *between p and q is*

$$D(p \,||\, q) = \sum_{\eta=0}^{M} p_\eta \log \frac{p_\eta}{q_\eta}, \tag{1}$$

with the conventions $0 \log 0/b = 0$, $b \neq 0$ *and* $b \log b/0 = \infty$.

The following property usually holds for p.d.'s related to any real ciphers, so it will be frequently used throughout this work:

Property 1. *We say that distribution p* is close to *q if $|p_\eta - q_\eta| \ll q_\eta$, for all $\eta = 0, 1, \ldots, M$.*

If p is close to q then we can approximate the Kullback-Leibler-distance between p and q by its Taylor series. We call the first term of the series the capacity of p and q and it is defined as follows:

Definition 2. *The capacity between two p.d.'s p and q is defined by*

$$C(p, q) = \sum_{\eta=0}^{M} \frac{(p_\eta - q_\eta)^2}{q_\eta}. \tag{2}$$

If q is the uniform distribution, then $C(p, q)$ will be denoted by $C(p)$ and called the capacity of p.

The normed normal distribution with mean 0 and variance 1 is denoted by $\mathcal{N}(0, 1)$. Its probability density function (p.d.f.) is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{3}$$

and the cumulative distribution function (c.d.f.) is

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)\, dt\,. \tag{4}$$

The normal distribution with mean μ and variance σ^2 is denoted by $\mathcal{N}(\mu, \sigma^2)$ and its p.d.f. and c.d.f. are ϕ_{μ,σ^2} and Φ_{μ,σ^2}, respectively.

The χ^2_M-distribution with M degrees of freedom has mean M and variance $2M$. The non-central $\chi^2_M(\lambda)$-distribution with M degrees of freedom has mean $\lambda + M$ and variance $2(M + 2\lambda)$. If $M > 30$, we may approximate $\chi^2_M(\lambda) \sim \mathcal{N}(\lambda + M, 2(M + 2\lambda))$ [10].

Let $X_1, \ldots, X_n$ be a sequence independent and identically distributed (i.i.d.) random variables where either $X_i \sim p$, for all $i = 1, \ldots, N$ (corresponding to null hypothesis H_0) or $X_i \sim q \neq p$, for all $i = 1, \ldots, N$ (corresponding to alternate hypothesis H_1) and let $\hat{x}_1, \ldots, \hat{x}_N$ be the empirical data. The hypothesis testing problem is then to determine whether to accept or reject H_0. The Neyman-Pearson lemma [11] states that an optimal statistic for solving this problem, or distinguishing between p and q, is the log-likelihood ratio defined by

$$\mathrm{LLR}(\hat{q}, p, q) = \sum_{\eta=0}^{M} N\hat{q}_\eta \log \frac{p_\eta}{q_\eta}, \tag{5}$$

where $\hat{q} = (\hat{q}_0, \ldots, \hat{q}_M)$ is the empirical p.d. calculated from the data $\hat{x}_1, \ldots, \hat{x}_N$ by

$$\hat{q}_\eta = \frac{1}{N} \#\{i = 1, \ldots, N \mid \hat{x}_i = \eta\}.$$

The distinguisher accepts H_0, that is, outputs p (respectively rejects H_0 or outputs q) if $\mathrm{LLR}(\hat{q}, p, q) \geq \gamma$ ($< \gamma$) where γ is the threshold that depends on the level and the power of the test. If the power and the level of the test are equal (as is often the case) then $\gamma = 0$.

The proof for the following result can be found in [11], see also [5].

Proposition 1. *The LLR-statistic calculated from i.i.d. empirical data $\hat{x}_i$, $i = 1, \ldots, N$ using (5) is asymptotically normal with mean and variance $N\mu_0$ and $N\sigma_0^2$ ($N\mu_1$ and $N\sigma_1^2$, resp.) if the data is drawn from p (q, resp.). The means and variances are given by*

$$\begin{gathered} \mu_0 = D(p \,||\, q) \quad \mu_1 = -D(q \,||\, p) \\ \sigma_0^2 = \sum_{\eta=0}^{M} p_\eta \log^2 \frac{p_\eta}{q_\eta} - \mu_0^2 \quad \sigma_1^2 = \sum_{\eta=0}^{M} q_\eta \log^2 \frac{p_\eta}{q_\eta} - \mu_1^2. \end{gathered} \tag{6}$$

Moreover, if p is close to q, we have

$$\mu_0 \approx -\mu_1 \approx \frac{1}{2} C(p, q) \quad \sigma_0^2 \approx \sigma_1^2 \approx C(p, q). \tag{7}$$

3 Advantage in Key Ranking

In a key recovery attack one is given a set of key candidates, and the problem is to determine which key is the right one. Usually the keys are searched from the set V_n of all 2^n strings of n bits. The algorithm consists of four phases, the *counting phase*, *analysis phase*, *sorting phase* and *searching phase* [12]. In the counting phase one collects data from the cipher, for example, plaintext-ciphertext pairs. In the analysis phase a real-valued statistic T is used in calculating a rank (or "mark" [12]) $T(\kappa)$ for all candidates $\kappa \in V_n$.

In the sorting phase the candidates κ are sorted, i.e., ranked, according to the statistic T. Optimally, the right key, denoted by κ_0, should be at the top of the list. If this is not the case, then one must also run through a search phase, testing the keys in the list until κ_0 is found. The goal of this paper is to find a statistic $T(\kappa)$ that is easy to compute and that is also reliable and efficient in finding the right key.

The time complexity of the search phase, given amount N of data, was measured using a special purpose quantity "gain" in [3]. A similar but more generally applicable concept of "advantage" was introduced by Selçuk in [8], where it was defined as follows:

Definition 3. *We say that a key recovery attack for an n-bit key achieves an advantage of a bits over exhaustive search, if the correct key is ranked among the top $r = 2^{n-a}$ out of all 2^n key candidates.*

Statistical tests for key recovery attacks are based on the Wrong-key Hypothesis [13]. We state it as follows:

Assumption 1 (Wrong-key Hypothesis). *There are two p.d.'s q and q', $q \neq q'$ such that for the right key κ_0, the data is drawn from q and for a wrong key $\kappa \neq \kappa_0$ the data is drawn from $q' \neq q$.*

A real-valued statistic T is computed from q and q', where one of these p.d.'s may be unknown, and the purpose of a statistic T is to distinguish between q and q'. We use D_R to denote the p.d. such that $T(\kappa_0) \sim D_R$. We will assume $D_R = \mathcal{N}(\mu_R, \sigma_R^2)$, with parameters μ_R and σ_R, as this will be the case with all statistics in this paper. Then μ_R and σ_R are determined with the help of linear cryptanalysis. We denote by D_W the p.d. known based on the Wrong-key Hypothesis such that $T(\kappa) \sim D_W$ for all $\kappa \neq \kappa_0$. The p.d.f. and c.d.f. of D_W are denoted by f_W and F_W, respectively.

Ranking the keys κ according to T means rearranging the 2^n r.v.'s $T(\kappa), \kappa \in V_n$, in decreasing order of magnitude. Writing the ordered r.v.'s as $T_0 \geq T_1 \geq \cdots \geq T_M$, we call T_i the ith order statistic. Let us fix the advantage a such that the right key should be among the $r = 2^{n-a}$ highest ranking keys. Hence, the right key should be at least as high as the rth wrong key corresponding to T_r. By Theorem 1. in [8] we get that the r.v. T_r is distributed as

$$\begin{aligned} T_r &\sim \mathcal{N}(\mu_a, \sigma_a^2), \text{ where} \\ \mu_a = F_W^{-1}(1 - 2^{-a}) &\text{ and } \sigma_a \approx \frac{2^{-(n+a)/2}}{f_W(\mu_a)}. \end{aligned} \tag{8}$$

If we now define the success probability P_S of having κ_0 among the r highest ranking keys we have

$$P_S = \Pr(T(\kappa_0) - T_r > 0) = \Phi\left(\frac{\mu_R - \mu_a}{\sqrt{\sigma_R^2 + \sigma_a^2}}\right), \tag{9}$$

since $T(\kappa_0) - T_r \sim \mathcal{N}(\mu_R - \mu_a, \sigma_R^2 + \sigma_a^2)$.

As the data complexity N depends on the parameters $\mu_R - \mu_a$ and $\sigma_R^2 + \sigma_a^2$, we can solve N from (9) as a function of a and vice versa. Hence, (9) describes the trade-off between the data complexity N and the complexity of the search phase.

In a block cipher, the unknown key is divided into a number of round keys not necessarily disjoint or independent. In [3], the keys of the last round (or first and last round) were called the outer keys and the rest of the round keys were called inner keys. The unknown key κ may consist of outer keys, the parity bits of inner keys or both. Traditionally, in Matsui's Alg. 1 key parity bit(s) of the inner keys are searched, whereas in Alg. 2. the main goal is to determine parts of the outer keys.

4 Algorithm 2

4.1 Multidimensional Linear Approximation

Let us study a block cipher with t rounds. Let $x \in V_n$ be the plaintext, $y \in V_n$ the ciphertext, $K \in V_\nu$ the fixed round key data (the inner key) used in all but

the last round and $z = f_t^{-1}(y,k)$, $k \in V_l$, the input to the last round function f_t, obtained from y by decrypting with the last round key data k (outer key). Let $m \leq n$ be an integer. Using m-dimensional linear cryptanalysis one can determine an approximation p of the p.d. of the Boolean function

$$x \mapsto Ux + Wz + VK, \tag{10}$$

which defines an m-dimensional linear approximation, where U and W are $m \times n$ matrices and V is an $m \times \nu$ matrix. A way of obtaining p from the one-dimensional correlations was presented in [4]. The linear mapping V divides the inner key space to 2^m equivalence classes $g = VK \in V_m$. Let the right last round key be denoted by k_0. Denote $M = 2^m - 1$ from now on.

In the counting phase we draw N data pairs $(\hat{x}_i, \hat{y}_i), i = 1, \ldots, N$. In the analysis phase, for each last round key k, we first calculate $\hat{z}_i^k = f_t^{-1}(\hat{y}_i, k), i = 1, \ldots, N$. Then, for each k, we calculate the empirical p.d. $\hat{q}^k = (\hat{q}_0^k, \ldots, \hat{q}_M^k)$, where

$$\hat{q}_\eta^k = \frac{1}{N} \#\{i = 1, \ldots, N \,|\, U\hat{x}_i + W\hat{z}_i^k = \eta\}. \tag{11}$$

If we use the wrong key $k \neq k_0$ to decrypt $\hat{y}_i, i = 1, \ldots, N$, it means we essentially encrypt over one more round and the resulting data will be more uniformly distributed. This heuristics is behind the original Wrong-key Randomisation Hypothesis [14], which in our case means that the data $U\hat{x}_i + W\hat{z}_i^k$, $i = 1, \ldots, N$, $k \neq k_0$ is drawn i.i.d. from the uniform distribution.

When decrypting with the correct key k_0 the data $U\hat{x}_i + W\hat{z}_i^{k_0} + g$, $i = 1, \ldots, N$, where g is an unknown inner key class, is drawn i.i.d. from p. This means that the data $U\hat{x}_i + W\hat{z}_i^{k_0}$, $i = 1, \ldots, N$ is drawn i.i.d. from a fixed permutation of p denoted by p^g. These permuted p.d.'s have the property that $p^g_{\eta \oplus h} = p^{g \oplus h}_\eta$, for all $g, \eta, h \in V_m$, and consequently

$$D(p^g \,||\, \theta) = D(p \,||\, \theta) \text{ and } C(p) = C(p^g) \text{ for all } g \in V_m. \tag{12}$$

Moreover, $D(p \,||\, p^h) = D(p^g \,||\, p^{h \oplus g})$, for all $h, g \in V_m$, from which it follows that

$$\min_{g' \neq g} D(p^g \,||\, p^{g'}) = \min_{h \neq 0} D(p \,||\, p^h), \tag{13}$$

which is a constant value for all $g \in V_m$. We will denote this value by $D_{\min}(p)$ and assume in the sequel that $D_{\min}(p) \neq 0$ without restriction: We can unite the key classes for which the Kullback-Leibler distance is zero. Then we just have $m' < 2^m$ key classes whose Kullback-Leibler distance from each other is non-zero. The corresponding minimum capacity $\min_{h \neq 0} C(p, p^h)$ is denoted by $C_{\min}(p)$.

4.2 Key Ranking in One-Dimensional Alg. 2

Key ranking and advantage in the one-dimensional case, $m = 1$, of Alg. 2 was studied in [8]. We will present it here briefly for completeness. Let $c > 0$ be the correlation of (10) (the calculations are similar if $c < 0$) and let $\hat{c}^k$ be the

empirical correlation calculated from the data. The statistic used in ranking the keys is then $s(k) = |\hat{c}^k|$. The r.v. $\hat{c}^{k_0}$ is binomially distributed with mean $\mu_R = c$ and variance $\sigma_R^2 = (1 - c^2)/N \approx 1/N$. The wrong key r.v.'s $\hat{c}^k$, $k \neq k_0$, are binomially distributed with mean $\mu_W = 0$ (following Assumption 1) and variance $\sigma_W^2 = \sigma_R^2$. Since N is large, we can approximate $s(k_0) \sim \mathcal{N}(\mu_R, \sigma_R^2)$ and $s(k) \sim \mathcal{FN}(\mu_W, \sigma_W^2)$, where $\mathcal{FN}$ is the folded normal distribution, see Appendix A in [8]. Now we can proceed as in [8]. We get that, with given success probability P_S and advantage a, the data complexity is

$$N = \frac{(\Phi^{-1}(P_S) + \Phi^{-1}(1 - 2^{-a-1}))^2}{c^2}. \tag{14}$$

4.3 Different Scenarios in Multiple Dimensions

When considering generalisation of Alg.2 to the case, where multiple linear approximations are used, different approaches are possible. In a previous work by Biryukov, et al., [3], a number of selected one-dimensional linear approximations with high bias are taken into account simultaneously under the assumption that they are statistically independent. As we will show later in Sect. 5.3, the statistic used in [3] is essentially a goodness-of-fit test based on least squares and searches simultaneously the key parts k_0 and g_0 which give the best fit with the theoretically estimated correlations.

The approaches taken in [5] for linear distinguishing and later in [4] for Alg. 1 do not need assumptions about independence of the linear approximations as they are based on the p.d. of the multidimensional linear approximation (10). When using the multidimensional p.d., basically two different standard statistical methods can be used:

- Goodness-of-fit (usually based on χ^2-statistic) and
- Distinguishing of an unknown p.d. from a given set of p.d.'s (usually based on LLR-statistic)

The goodness-of-fit approach is a straightforward generalisation of one-dimensional Alg. 2. It can be used in searching for $\kappa = k$. It measures whether the data is drawn from the uniform (wrong) distribution, or not, by measuring the deviation from the uniform distribution. It ranks highest the key candidate whose empirical distribution is farthest away from the uniform distribution. The statistic does not depend on the inner key class g. Information about p.d. p is required only for measuring the strength of the test. We will study this method in Sect. 5.1. After the right round key k is found, one can use the data derived in Alg. 2 in any form of Alg. 1 for finding the inner key class g. In this manner, the χ^2-approach allows separating between Alg. 1 and Alg. 2.

The LLR-method uses the information about the p.d. related to the inner key class also in Alg. 2. In this sense, it is similar to the method of [3], where the Alg. 1 and Alg. 2 were combined together for finding both the outer and inner round keys. As we noted in Sect. 2, the LLR-statistic is the optimal distinguisher between two known p.d.'s. If we knew the right inner key class g_0, we could simply

use the empirical p.d.'s $\hat{q}^k$ for distinguishing p^{g_0} and the uniform distribution and then choose the k for which this distinguisher is strongest [5]. In practice, the correct inner key class g_0 is unknown when running Alg. 2 for finding the last round key.

Our approach is the following. In [7] it was described how one can use LLR to distinguish one known p.d. from a set of p.d.'s. We will use this distinguisher for distinguishing θ from the given set p^g, $g \in V_m$. In the setting of Alg. 2, we can expect that for the right k_0, it should be possible to clearly conclude that the data $(\hat{x}_i, \hat{y}_i)$, $i = 1, \ldots, N$, yields data $(\hat{x}_i, \hat{z}_i^{k_0})$, $i = 1, \ldots, N$, which follows a p.d. p^g, for some $g \in V_m$, rather than the uniform distribution. On the other hand, for the wrong $k \neq k_0$, the data follows the uniform distribution, rather than any p^g, $g \in V_m$.

To distinguish k_0 from the wrong key candidates we determine, for each round key candidate k, the inner key class g, for which the LLR-statistic is the largest with the given data. The right key k_0 is expected to have g_0 such that the LLR-statistic with this pair (k_0, g_0) is larger than for any other pair $(k, g) \neq (k_0, g_0)$. In this manner, we also recover g_0 in addition to k_0. The LLR-method is studied in Sect. 6.

5 The χ^2-Method

This method separates the Alg. 1. and Alg. 2 such that the latter does not need any information of p. Both methods are interpreted as goodness-of-fit problems, for which the natural choice of ranking statistic is χ^2. We will show how to find the last round key k with Alg. 2 first.

5.1 Algorithm 2 with χ^2

Given empirical p.d. $\hat{q}^k$, we can calculate the χ^2-statistic from the data as

$$S(k) = 2^m N \sum_{\eta=0}^{M} (\hat{q}_\eta^k - 2^{-m})^2, \tag{15}$$

where $M = 2^m - 1$ is the number of degrees of freedom. The statistic can be interpreted as the l_2-distance between the empirical p.d. and the uniform distribution. By Assumption 1, the right round key should produce data that is farthest away from the uniform distribution and we will choose the round key k for which the statistic (15) is largest. Obviously, if $m = 1$, we get the statistic $(\hat{c}^k)^2$.

According to [15] the r.v. $S(k_0)$ is distributed approximately as

$$S(k_0) \sim \chi^2_M(NC(p^{g_0})) = \chi^2_M(NC(p)), \tag{16}$$

because of the symmetry property (12). Hence, we may approximate the distribution by a normal distribution with $\mu_R = M + NC(p)$ and $\sigma_R^2 = 2(M + 2NC(p))$.

The parameters do not depend on g_0 or k_0. For the wrong keys $k \neq k_0$, we obtain by [15] that

$$S(k) \sim \chi^2_M(0) = \chi^2_M, \tag{17}$$

so that $\mu_W = M$ and $\sigma_W^2 = 2M$. The mean and variance in (8) are $\mu_a = \sigma_W b + M = \sqrt{2M}b + M$ and $\sigma_a^2 = 2^{-(l+a)/2}\sigma_W^2/\phi(b) \ll \sigma_0^2$. Now we can solve N from (9) and get that the data complexity is proportional to

$$N_{\chi^2} = \frac{\beta(M, b, P_S)}{C(p)},\ b = \Phi^{-1}(1 - 2^{-a}), \tag{18}$$

where $\beta(M, b, P_S)$ is a parameter that depends on M, b and P_S. Assuming large b, that is, large advantage a and large P_S, we can approximate β by

$$\beta = 2\sqrt{M}b + 4\Phi^{-2}(2P_S - 1). \tag{19}$$

Note that the normal approximation of the wrong-key distribution is valid only when $m > 5$, that is, when the approximation of χ^2-distribution by a normal distribution is valid. It is not possible to perform the theoretical calculations for small m as the χ^2-distribution does not have a simple asymptotic form in that case and we cannot determine f_W and F_W in (8). Since our χ^2-statistic reduces to the square of $s(k)$ that was used by Selçuk, the theoretical distributions differ from our calculations and we get a slightly different formula for the advantage. Despite this difference, the methods are equivalent for $m = 1$.

Keeping the capacity constant, it seems that the data complexity increases exponentially as $2^{m/2}$ as the dimension m of the linear approximation increases and is sufficiently large. Hence, in order to strengthen the attack, the capacity should increase faster than $2^{m/2}$ when the m is increased. This is a very strong condition and it suggests that in applications, only approximations with small m should be used with χ^2-attack. The experimental results for different m presented in Sect. 8 as well as the theoretical curves depicted in Fig. 5(a) suggest that increasing m in the χ^2-method does not necessarily mean improved performance for Alg. 2.

Since $2^{-a} = \Phi(-b) \approx 1/\sqrt{2\pi}e^{-b^2/2}$, we can solve a from (18) as a function of N and we have proved the following theorem that can be used in describing the relationship between the data complexity and the search phase:

Theorem 1. *Suppose the cipher satisfies Assumption 1 where $q' = \theta$ and the p.d.'s p^g, $g \in V_m$ and θ are close to each other. Then the advantage of the χ^2-method using statistic* (15) *is given by*

$$a_{\chi^2} = \frac{(NC(p) - 4\varphi)^2}{4M},\ \varphi = \Phi^{-2}(2P_S - 1),\ M = 2^m - 1, \tag{20}$$

where P_S (> 0.5) is the probability of success, N is the amount of data used in the attack and $C(p)$ and m (≥ 5) are the capacity and the dimension of the linear approximation (10)*, respectively.*

While (20) and (18) depend on the theoretical distribution p, the actual χ^2-statistic (15) is independent of p. Hence, we do not need to know p accurately to realise the attack, we only need to find an approximation (10) that deviates as much as possible from the uniform distribution. On the other hand, if we use time and effort for computing an approximation of the theoretical p.d. and if we may assume that the approximation is accurate, we would also like to exploit this knowledge for finding the right inner key class with Alg. 1. As noted in [9], there are several ways to realising a multidimensional Alg. 1. Next we discuss Alg. 1 as a χ^2-based goodness-of-fit problem.

5.2 Algorithm 1 with χ^2

Suppose that we have obtained an empirical distribution $\hat{q}$ of data that can be used for determining the inner key class g_0 using Alg. 1. For example, we have successfully run Alg. 2 and found the correct last round key k_0 and set $\hat{q} = \hat{q}^{k_0}$.

One approach is to consider Alg. 1 as a goodness-of-fit problem, where one determines, for each g, whether the empirical p.d. $\hat{q}$ follows p^g or not. The χ^2-based ranking statistic is then

$$S_{\text{Alg1}}(g) = N \sum_{\eta=0}^{M} \frac{(\hat{q}_\eta^{k_0} - p_\eta^g)^2}{p_\eta^g}, \tag{21}$$

which should be small for g_0 and large for the wrong inner key classes $g \neq g_0$. In [9] it is shown that the data complexity of finding g_0 with given success probability P_S is

$$N_{\text{Alg 1},\chi^2} = \frac{4m - 4\gamma_S + 2\sqrt{2M(m - \gamma_S)}}{C_{\min}(p)}, \tag{22}$$

where $\gamma_S = \ln(\sqrt{2\pi} \ln P_S^{-1})$.

5.3 Combined Method and Discussion

The sums of squares of correlations used in [3] are closely related to the sums of squares (15) and (21). Indeed, we could define a combined χ^2-statistic B by considering the sum of the statistics from (15) and (21) and setting

$$B(k, g) = \sum_{k' \neq k} S(k) + S_{\text{Alg 1}}(k, g), \tag{23}$$

where $S_{\text{Alg 1}}(k, g)$ is the statistic (21) calculated from the empirical p.d. $\hat{q}^k$, $k \in V_l$. If we approximate the denominators in (21) by 2^{-m} and scaling by $2^m N$ we obtain from $B(k, g)$ the statistic

$$B'(k, g) = \sum_{k' \neq k} ||\hat{q}^{k'} - \theta||_2^2 + ||\hat{q}^k - p^g||_2^2. \tag{24}$$

This statistic is closely related to the one used in [3].

$$\sum_{k' \neq k} ||\hat{c}^{k'}||_2^2 + ||\hat{c}^k - c^g||_2^2. \tag{25}$$

Indeed, if in (25) all correlation vectors $\hat{c}^k$ and c^g contain correlations from all linear approximations then (25) becomes the same as $2^m B'(k, g)$ as can be seen using Parseval's theorem. Initially, in the theoretical derivation of (25) only linearly and statistically independent approximations were included in the correlation vectors. However, in Sect. 3.4 of [3] it was proposed to take into account all linear approximations with strong correlations when forming the statistic (25) in practice. In practical experiments by Collard, et al. [16] this heuristic enhancement was demonstrated to improve the results. In this paper, we have shown how to remove the assumption about independence of the linear approximations and that all linear approximations that have sufficient contribution to the capacity (cf. discussion in Sect. 5.1) can and should be included.

Other possibilities for combining Alg. 1 and Alg. 2 based on χ^2 or its variants are also possible, with different weights on the terms of the sum in (24), for instance. However, the mathematically more straightforward way is to use the pure χ^2-method defined by (15) and (21), as its statistical behaviour is well-known. An even more efficient method can be developed based on LLR as will be shown next.

6 The LLR-Method

This method is also based on the same heuristic as the Wrong-key Hypothesis: For $k \neq k_0$, the distribution of the data should look uniform and for k_0 it should look like p^{g_0}, for some g_0. Hence, for each k, the problem is to distinguish the uniform distribution from the discrete and known set p^g, $g \in V_m$. Let us use the notation $L(k, g) = \mathrm{LLR}(\hat{q}^k, p^g, \theta)$. We propose to use the following ranking statistic

$$L(k) = \max_{g \in V_m} L(k, g). \tag{26}$$

Now k_0 should be the key for which this maximum over g's is the largest and ideally, the maximum should be achieved when $g = g_0$. While the symmetry property (12) allows one to develop statistical theory without knowing g_0, in practice one must search through V_l for k_0 and V_m for g_0 even if we are only interested in determining k_0.

We assume that the p.d.'s p^g and θ are all close to each other. Using Theorem 1 and property (12) we can state Assumption 1 as follows: For the right pair k_0 and g_0

$$L(k_0, g_0) \sim \mathcal{N}(N\mu_R, N\sigma_R^2), \text{ where } \mu_R = \frac{1}{2}C(p) \text{ and } \sigma_R^2 = C(p), \tag{27}$$

and for $k \neq k_0$ and any $g \in V_m$

$$L(k, g) \sim \mathcal{N}(N\mu_W, N\sigma_W^2), \text{ where } \mu_W = -\frac{1}{2}C(p) \text{ and } \sigma_W^2 = C(p). \tag{28}$$

Hence, μ_R, σ_R^2, μ_W and σ_W^2 do not depend on $g \in V_m$. For fixed $k \neq k_0$, the r.v.'s $L(k,g)$ for $k \neq k_0$ are identically normally distributed with mean μ_W and variance σ_W^2. We will assume that they are statistically independent to simplify calculations. In particular, the assumption about statistical independence of $L(k,g)$ for different g does not mean that the linear approximations should be statistically independent. The statistic itself does not depend on this assumption[1]. Moreover, the theoretical results obtained this way are a little more pessimistic that those obtained by empirical tests, as shown in Sect. 8. Hence, these calculations give a theoretical model that can be used in describing how the method behaves especially compared to other methods. Assuming that for each $k \neq k_0$, the r.v.'s $L(k,g)$'s are independent, we obtain that the c.d.f. of their maximum is given by [17]

$$F_W(x) = \Phi_{N\mu_W, N\sigma_W^2}(x)^{M+1} \tag{29}$$

and p.d.f. is

$$f_w(x) = (M+1)\Phi_{N\mu_W, N\sigma_W^2}(x)^M \phi_{\mu_W, \sigma_W^2}(x). \tag{30}$$

Let us fix the advantage a such that $r = 2^{l-a}$. The mean μ_a of the rth wrong key statistic L_r can now be calculated from (8) to be

$$\begin{aligned} \mu_a &= N\mu_W + \sqrt{N}\sigma_W b = -1/2NC(p) + \sqrt{NC(p)}b, \\ b &= \Phi^{-1}(\sqrt[M+1]{1-2^{-a}}), \end{aligned} \tag{31}$$

and the variance is

$$\sigma_a^2 = \frac{2^{-l-a}\sigma_W^2}{(M+1)^2(1-2^{-a})^{2(1-1/(M+1))}\phi^2(b)} \ll \sigma_0^2. \tag{32}$$

Let

$$P_1 = \Pr(L(k_0, g_0) > \max_{g \neq g_0} L(k_0, g)) \tag{33}$$

be the the probability that given k_0, we choose g_0, i.e., the probability of success of Alg. 1. Let

$$P_2 = \Pr(L(k_0) > L_r) \tag{34}$$

be the probability that we rank k_0, paired with *any* $g \in V_m$, among the r highest ranking keys. Finally, let

$$P_{12} = \Pr(L(k_0) > L_r \,|\, L(k_0, g_0) > \max_{g \neq g_0} L(k_0, g)) \tag{35}$$

be the probability that we rank k_0 among the r highest ranking keys provided that we pair g_0 with k_0. Then

$$\begin{aligned} P_2 &= P_{12}P_1 + \Pr(L(k_0) > L_r \,|\, L(k_0) = \mathrm{LLR}(k_0, p^g, \theta), g \neq g_0)(1 - P_1) \\ &\geq P_{12}P_1. \end{aligned} \tag{36}$$

[1] See for example [17] for calculating the c.d.f. of the maximum of dependent and identically distributed r.v.'s, when $M \geq 100$. The theoretical predictions calculated that way are slightly more pessimistic than the ones obtained in Theorem 2.

If we pair k_0 with $g \neq g_0$ then $L(k_0) \geq L(k_0, g_0)$ for a fixed empirical p.d. $\hat{q}^{k_0}$, so that k_0 gets ranked *higher* than by using the correct g_0. Hence, assuming that k_0 gets paired with g_0 only decreases P_2 so the corresponding estimate of the data complexity gets larger. Let N_1, N_2 and N_{12} be the data complexities needed to achieve success probabilities P_1, P_2 and P_{12}, respectively.

We can calculate P_{12} using (27), (28) and (9) to obtain

$$P_{12} = \Phi(\frac{\mu_R - \mu_W - \sigma_w b}{\sigma_R}) = \Phi(\sqrt{N_{12}C(p)} - b),\ b = \Phi^{-1}(\sqrt[M+1]{1 - 2^{-a}}). \quad (37)$$

Hence, the data complexity is proportional to

$$N_{12} = \left(\Phi^{-1}(P_{12}) + b\right)^2 / C(p), \quad (38)$$

which can be used in approximating an upper bound for N_2. We can approximate $\Phi(b) = \sqrt[M+1]{1 - 2^{-a}} \approx 1 - 2^{-m-a}$ such that $a \approx b^2/2 - m$ and we can solve the advantage a as a function of $N_{12} \approx N_2$ from (38). We get the following theorem:

Theorem 2. *Suppose the cipher satisfies Assumption 1 where $q' = \theta$ and the p.d.'s p^g, $g \in V_m$ and θ are close to each other. Then the advantage of the LLR-method for finding the last round key k_0 is given by*

$$a_{\mathrm{LLR}} = (\sqrt{NC(p)} - \Phi^{-1}(P_{12}))^2/2 - m \approx NC(p) - m. \quad (39)$$

Here N is the amount of data used in the attack, P_{12} (> 0.5) is the probability of success and $C(p)$ and m are the capacity and the dimensions of the linear approximation (10)*, respectively.*

Theorem 2 now gives the trade-off between the search phase and the data complexity of the algorithm. With fixed N and capacity $C(p)$, the advantage decreases linearly with m whereas in (20) the logarithm of advantage decreases linearly with m. For fixed m and p, the advantage of the LLR-method seems to be larger than the advantage of the χ^2-method. The experimental comparison of the methods is presented Sect. 8

In [4] it is shown that the data complexity of Alg. 1 for finding the right inner key class g_0 is proportional to

$$N_1 = \frac{16m \ln 2 - 16P_1'}{C(p)}, \quad (40)$$

where $P_1' = \ln(\sqrt{2\pi} \ln P_1^{-1})$. If we want to be certain that we have paired the right inner key class g_0 with k_0, the data complexity is given by

$$N_{\mathrm{LLR}} = \max(N_1, N_2) \propto \frac{m}{C(p)}. \quad (41)$$

The data complexity N_1 is an overestimate for the actual data complexity of Alg. 1 [9] so in practice, N_2 dominates.

7 Algorithms and Complexities

For comparing the two methods, LLR and χ^2, we are interested in the complexities of the first two phases of the Alg. 2 since the sorting and searching phase do not depend on the chosen statistic. The counting phase is done on-line and all the other phases can be done off-line. However, we have not followed this division [12] in our implementation, as we do part of the analysis phase on-line. We will divide the algorithm in two phases as follows: In the *on-line phase*, depicted in Fig. 1, we calculate the empirical p.d.'s for the round key candidates. The marks $S(k)$ for the χ^2-method and $L(k)$ for the LLR-method are then assigned to the keys in the *off-line phase*. The off-line phases for χ^2-method and LLR-method are depicted in Fig. 2 and Fig. 4, respectively. After the keys k are each given the mark, they can be ranked according to the mark. If we wish to recover g_0 with χ^2-method, we also need to store, in addition to the marks, the empirical p.d.'s q^k. Given q^{k_0}, one can use the multidimensional Alg. 1 described in Fig. 3 for finding g_0 off-line. The version of Alg. 1 is based on LLR. Obviously, one could use some other method, e.g. use the χ^2-based ranking statistic (21), which gives similar results in practice even if the LLR-based method is more powerful in theory [9].

initialise $2^l \times 2^m$ counters $F(k, \eta)$, $k = 0, \ldots, 2^l - 1$, $\eta = 0, \ldots, M$;
for $i = 1, \ldots, N$ **do**
  for *candidates* $k = 0, \ldots, 2^l - 1$ **do**
    decrypt the ciphertext partially: $\hat{z}_i^k = f^{-1}(\hat{y}_i, k)$;
    for $j = 1, \ldots, m$ **do**
      calculate bit $\eta_j = u_j \cdot \hat{x}_i \oplus w_j \cdot \hat{z}_i^k$;
    end
    increment counter $F(k, \eta) = \#\{i \,|\, U\hat{x}_i + W\hat{z}_i^k = \eta\}$, where η is the vector $(\eta_1, \ldots, \eta_m)$ interpreted as an integer;
  end
end

Fig. 1. On-line phase of Matsui's Alg. 2 in multiple dimensions

Input: table $F(k, \eta), k = 0, \ldots, 2^l - 1$, $\eta = 0, \ldots, M$;
for $k = 0, \ldots, 2^l - 1$ **do**
  compute $S(k) = \sum_{\eta=0}^{M} (F(k, \eta)/N - 2^{-m})^2$;
  if *wish to recover* g_0 **then**
    store $(S(k), F(k, 0), \ldots, F(k, M))$;
  else
    store $S(k)$;
  end
end

Fig. 2. Off-line phase of Alg. 2 using χ^2-method

Table 1. Data, time and memory complexities of the χ^2- and LLR-method

	On-line			Off-line		
	χ^2 for k_0	χ^2 for k_0, g_0	LLR	χ^2 for k_0	χ^2 for k_0, g_0	LLR
Data	N_{χ^2}	N_{χ^2}	N_{LLR}	–	–	–
Time	$N_{\chi^2} 2^l m$	$N_{\chi^2} 2^l m$	$N_{\mathrm{LLR}} 2^l m$	2^{l+m}	2^{l+m}	2^{l+m}
Memory	2^{l+m}	2^{l+m}	2^{l+m}	2^l	$2^m \max(2^l, 2^m)$	$2^m \max(2^l, 2^m)$

Input: counter values $F(k_0, 0), \ldots, F(k_0, M)$;
compute the theoretical distribution of m-dimensional approximations for each value of 2^m inner key classes and store them in a $2^m \times 2^m$ table $P(g, \eta), g = 0, \ldots, M$, $\eta = 0, \ldots, M$;
for *inner key classes* $g = 0, \ldots, M$ **do**
    calculate $G(g) = \sum_{\eta=0}^{M} F(k_0, \eta) \log P(g, \eta)$;
end
Output: g_0 such that $\max_{g \in V_m} G(g) = G(g_0)$

Fig. 3. Matsui's Alg. 1 in multiple dimensions (using LLR)

Input: table $F(k, \eta)$, $k = 0, \ldots, 2^l - 1$, $\eta = 0, \ldots, M$;
compute the theoretical distribution of m-dimensional approximations for each value of 2^m inner key classes and store them in a $2^m \times 2^m$ table $P(g, \eta)$, $g = 0, \ldots, M$, $\eta = 0, \ldots, M$;
for $k = 0, \ldots, 2^l - 1$ **do**
    for $g = 0, \ldots, M$ **do**
        $L(k, g) = \mathrm{LLR}(\hat{q}^k, p^g, \theta)$, where $\hat{q}^k_\eta = F(k, \eta)/N$;
    end
    store $L(k) = \max_{g \in V_m} L(k, g)$;
end

Fig. 4. Off-line phase of Alg. 2 using LLR-method

The data, time and memory complexities for on-line and off-line phase for both methods are shown in Table 1. Given success probability P_S and advantage a, the data complexity N_{χ^2} is given by (18). If we want to recover g_0 also, then theoretically, data complexity N_1 given by (40) is needed to successfully run Alg. 1 given in Fig. 3. As noted in [9], the theoretical value N_1 is an overestimate and the total data complexity in practice is probably dominated by the data complexity N_{χ^2} of ranking k_0 high enough. Nevertheless, the data complexity of the LLR-method is smaller than the χ^2-method.

Otherwise, the complexities for the LLR-method are mostly the same as for χ^2-method provided that m is not much larger than l which is usually the case. Thus, we recommend using the LLR-method rather than χ^2-method unless there is great uncertainty about the validity of the approximative p.d p of the linear relation (10).

In some situations it may also be advantageous to combine the different methods. For example, one may want to first find, say, r best round keys by χ^2, such that the data complexity N_{χ^2} is given by (18), where the advantage is $a = l - r$. Then one can proceed by applying the LLR-method to the remaining r keys, thus reducing the size of the round key space to be less than 2^l. Other similar variants are possible. Their usefulness depends on the cipher that is being studied.

8 Experiments

The purpose of the experiments was to test the accuracy of the derived statistical models and to demonstrate the better performance of the LLR-based method in practice. Similarly as in previous experiment on multiple linear cryptanalysis, see [16] and [3], the Serpent block cipher was used as a test-bed. The structure of Serpent is described, for example, in [18]. We have searched for a 12-bit part of the fifth round key based on m linear approximations with different m. Each experiment was performed for 16 different keys.

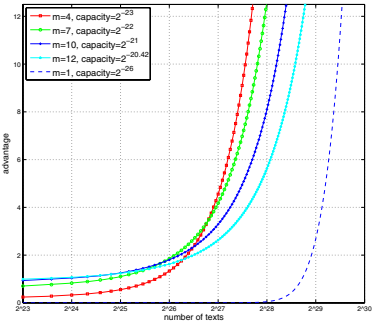
We calculated the capacities for the approximation (10) over 4-round Serpent for different m. Practical experiments were used in confirming that $C_{\min}(p) \approx C(p)$ and especially $C_{\min}(p) \neq 0$. We also saw that $|p_\eta^g - p_\eta^{g'}| < \frac{1}{150} p_\eta^g$, for all g, g' and $\eta \in V_m$. Hence, p^g's can be considered to be close to each other and θ.

The theoretical advantage of the χ^2-method predicted in (20) has been plotted as a function of data complexity in Fig. 5(a). The figure shows that increasing m larger than 4, the attack is weakened. This suggests using $m = 4$ base approximations in the χ^2-attack. Since we should have m at least 5 for the normal approximation of χ^2_M to hold, the theoretical calculations do not necessarily hold for small m. However, the experiments, presented in Fig. 5(b), seem to confirm the theory for $m = 1$ and $m = 4$, too. The most efficient attack is obtained by using $m = 4$ equations. Increasing m (and hence, the time and memory complexities of the attack, see Table 1) actually weakens the attack. The optimal choice of m depends on the cipher. However, the theoretical calculations suggest that using $m \geq 5$ is usually not advantageous.

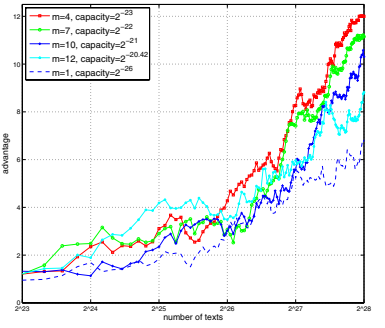
The reason is the χ^2-squared statistic itself: it only measures if the data follows a certain distribution, the uniform distribution in this case. The more approximations we use, the larger the distributions become and the more uncertainty we have about the "fitting" of the data. Small errors in experiments generate large errors in χ^2 as the fluctuations from the relative frequency 2^{-m} become more significant.

The theoretical advantage of the LLR-method (39) is plotted against the data complexity in Fig. 5(c) for different m. The empirical advantages for several different m are shown in Fig. 5(d). Unlike for χ^2 we see that the method can be strengthened by increasing m, until the increase in the capacity $C(p)$ becomes negligible compared to increase in m. For 4-round Serpent, this happens when $m \approx 12$.

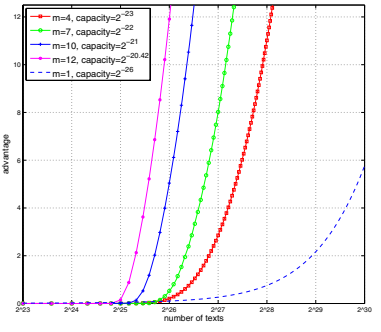
Experimental results presented in Figures 5(d) and 5(b) confirm the theoretical prediction that the LLR-method is more powerful than the χ^2-method. Also



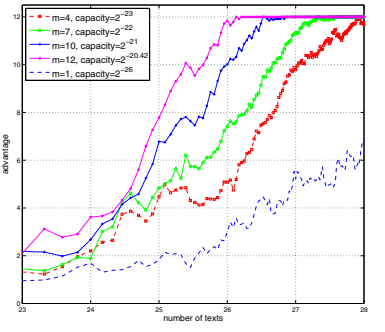
(a) Theoretical advantage for χ^2-method



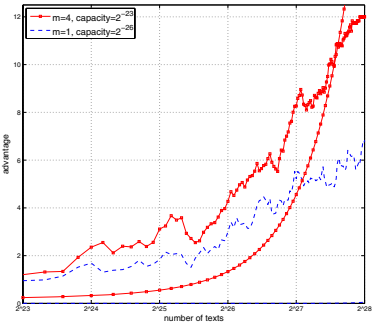
(b) Empirical advantage for χ^2-method



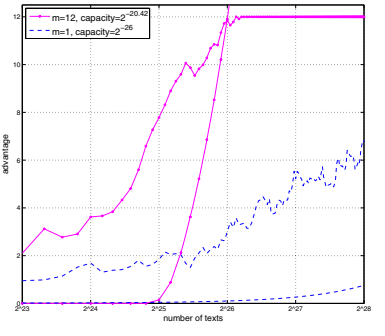
(c) Theoretical advantage for LLR-method



(d) Empirical advantage for LLR-method



(e) Empirical and theoretical advantage for χ^2 for $m = 1$ and $m = 4$



(f) Empirical and theoretical advantage for LLR for $m = 1$ and $m = 12$

Fig. 5. Theoretical and empirical advantages for χ^2- and LLR-method for different m and $P_S = P_{12} = 0.95$

the theoretical and empirical curves seem to agree nicely. For example, the full advantage of 12 bits with $m = 7$ achieved at $\log N = 27.5$ for LLR whereas χ^2-method needs about $\log N = 28$. Moreover, the LLR can be strengthened by increasing m. For $m = 12$, the empirical logarithmic data complexity is about 26.5.

9 Conclusions

There are several approaches of realising Matsui's Alg. 2 using multiple linear approximations. In this paper, methods based on two standard statistics, LLR and χ^2, were studied. Selçuk's theory of advantage describing the trade-off between data complexity and search phase was extended to multiple dimensions. The advantages of the two methods in key ranking were then determined. A description of the multidimensional Alg. 2 for both methods was given so that their performance measured in time, memory and data could be compared.

The χ^2-statistic, based on the classic goodness-of-fit test, was observed to perform poorly for large dimensions m of linear approximation, whereas the LLR-statistic, an optimal statistic for testing two known hypotheses, was shown to improve with the dimension m of the linear approximation much further. In particular, the advantage of using multiple linear approximations instead of just one is significant and of real practical importance if LLR-statistic is used in Alg. 2. In general, it was shown that the LLR-method is usually more advantageous compared to the χ^2-method. As long as there is no significant error, stemming from the linear hull-effect, for example, in determining the approximate p.d. of the multidimensional linear approximation, we recommend to use the LLR-method proposed in this paper rather than the χ^2-method.

Acknowledgements

We would like to thank Christophe de Cannière for insightful discussions and the anonymous referees for comments that helped us to improve the presentation of this paper.

References

1. Matsui, M.: Linear Cryptanalysis Method for DES Cipher. In: Helleseth, T. (ed.) EUROCRYPT 1993. LNCS, vol. 765, pp. 386–397. Springer, Heidelberg (1994)
2. Kaliski Jr., B.S., Robshaw, M.J.B.: Linear Cryptanalysis Using Multiple Approximations. In: Desmedt, Y.G. (ed.) CRYPTO 1994. LNCS, vol. 839, pp. 26–39. Springer, Heidelberg (1994)
3. Biryukov, A., Cannière, C.D., Quisquater, M.: On Multiple Linear Approximations. In: Franklin, M. (ed.) CRYPTO 2004. LNCS, vol. 3152, pp. 1–22. Springer, Heidelberg (2004)
4. Hermelin, M., Nyberg, K., Cho, J.Y.: Multidimensional Linear Cryptanalysis of Reduced Round Serpent. In: Mu, Y., Susilo, W., Seberry, J. (eds.) ACISP 2008. LNCS, vol. 5107, pp. 203–215. Springer, Heidelberg (2008)

5. Baignères, T., Junod, P., Vaudenay, S.: How Far Can We Go Beyond Linear Cryptanalysis? In: Lee, P.J. (ed.) ASIACRYPT 2004. LNCS, vol. 3329, pp. 432–450. Springer, Heidelberg (2004)
6. Junod, P.: On the optimality of linear, differential and sequential distingishers. In: Biham, E. (ed.) EUROCRYPT 2003. LNCS, vol. 2656, pp. 17–32. Springer, Heidelberg (2003)
7. Baignères, T., Vaudenay, S.: The Complexity of Distinguishing Distributions (Invited Talk). In: Safavi-Naini, R. (ed.) ICITS 2008. LNCS, vol. 5155, pp. 210–222. Springer, Heidelberg (2008)
8. Selçuk, A.A.: On probability of success in linear and differential cryptanalysis. Journal of Cryptology 21(1), 131–147 (2008)
9. Hermelin, M., Cho, J.Y., Nyberg, K.: Statistical Tests for Key Recovery Using Multidimensional Extension of Matsui's Algorithm 1. In: EUROCRYPT 2009 - poster session (2009)
10. Cramér, H.: Mathematical Methods of Statistics, 7th edn. Princeton Mathematical Series. Princeton University Press, Princeton (1957)
11. Cover, T.M., Thomas, J.A.: Elements of Information Theory, 2nd edn. Wiley Series in Telecommunications and Signal Processing. Wiley-Interscience, Hoboken (2006)
12. Vaudenay, S.: An experiment on DES statistical cryptanalysis. In: CCS 1996: Proceedings of the 3rd ACM conference on Computer and communications security, pp. 139–147. ACM, New York (1996)
13. Harpes, C., Kramer, G.G., Massey, J.L.: A Generalization of Linear Cryptanalysis and the Applicability of Matsui's Piling-Up Lemma. In: Guillou, L.C., Quisquater, J.-J. (eds.) EUROCRYPT 1995. LNCS, vol. 921, pp. 24–38. Springer, Heidelberg (1995)
14. Junod, P., Vaudenay, S.: Optimal Key Ranking Procedures in a Statistical Cryptanalysis. In: Johansson, T. (ed.) FSE 2003. LNCS, vol. 2887, pp. 235–246. Springer, Heidelberg (2003)
15. Drost, F., Kallenberg, W., Moore, D.S., Oosterhoff, J.: Power Approximations to Multinomial Tests of Fit. Journal of the American Statistican Association 84(405), 130–141 (1989)
16. Collard, B., Standaert, F.X., Quisquater, J.J.: Experiments on the Multiple Linear Cryptanalysis of Reduced Round Serpent. In: Nyberg, K. (ed.) FSE 2008. LNCS, vol. 5086, pp. 382–397. Springer, Heidelberg (2008)
17. David, H.A.: Order Statistics, 1st edn. A Wiley Publication in Applied Statistics. John Wiley & Sons, Inc., Chichester (1970)
18. Biham, E., Dunkelman, O., Keller, N.: Linear Cryptanalysis of Reduced Round Serpent. In: Matsui, M. (ed.) FSE 2001. LNCS, vol. 2355, pp. 219–238. Springer, Heidelberg (2001)