The memory centre

IMUJ PREPRINT 2012/03

P. Spurek

Faculty of Mathematics and Computer Science, Jagiellonian University, Łojasiewicza 6, 30-348 Kraków, Poland

J. Tabor

Faculty of Mathematics and Computer Science, Jagiellonian University, Łojasiewicza 6, 30-348 Kraków, Poland

Abstract

Let $x \in \mathbb{R}$ be given. As we know the, amount of bits needed to binary code x with given accuracy ($h \in \mathbb{R}$) is approximately $\mathrm{m}_h(x) \approx \log_2(\max\{1, |\frac{x}{h}|\})$. We consider the problem where we should translate the origin a so that the mean amount of bits needed to code randomly chosen element from a realization of a random variable X is minimal. In other words, we want to find $a \in \mathbb{R}$ such that

$$\mathbb{R} \ni a \to \mathrm{E}(\mathrm{m}_h(X - a))$$

attains minimum.

We show that under reasonable assumptions, the choice of a does not depend on h asymptotically. Consequently, we reduce the problem to finding minimum of the function

$$\mathbb{R} \ni a \to \int_{\mathbb{R}} \ln(|x - a|) f(x) dx,$$

where f is the density distribution of the random variable X. Moreover, we provide constructive approach for determining a.

Keywords: memory compressing, IRLS, differential entropy, coding, kernel estimation

Email addresses: przemyslaw.spurek@ii.uj.edu.pl (P. Spurek), jacek.tabor@ii.uj.edu.pl (J. Tabor)

1. Introduction

Data compression is usually achieved by assigning short descriptions (codes) to the most frequent outcomes of the data source and necessarily longer descriptions to the less frequent outcomes [1, 3, 7] .

For the convenience of the reader, we shortly present theoretical background of this approach. Let $p = (p_0, \ldots, p_{n-1})$ be a probability distribution for a discrete random variable X. Assume that l_i is the length of the code of x_i for $i = 0, \ldots, n-1$. Then the expected number of bits is given by $\sum_i p_i l_i$. The set of possible codeword with uniquely decodable codes is limited by the Kraft inequality $\sum_i 2^{-l_i} \leq 1$. It is enough to verify that lengths which minimize $\sum_i p_i l_i$ are given by $l_i = \log_2 p_i$. We obtain that minimal amount of information per one element in lossless coding is Shannon entropy [1] defined by

$$H(X) = \sum -p_i \log_2 p_i.$$

By this approach various types of lossless data compression were constructed. An optimal (shortest expected length) prefix code for a given distribution can be constructed by a simple algorithm discovered by Huffman [5].

If we want to consider continuous random variables and code with given maximal error h we arrive at the notion of differential entropy [1]. Let $f\colon \mathbb{R} \to \mathbb{R}$ be a continuous density distribution of the random variable X, and let $l\colon \mathbb{R} \to \mathbb{R}$ be the function of the code length. We divide the domain of f into disjoint intervals of length h. Let $X_h := \frac{1}{h}\lfloor hx \rfloor$ be the discretization of X. The Shannon entropy of X_h can be rewritten as follows

$$\begin{aligned} H(X_h) &\approx \sum -f(x_i)h \log_2 \left(f(x_i)h\right) \\ &= \sum -f(x_i) \log_2 \left(f(x_i)\right) h - \log_2 (h) \sum f(x_i)h \\ &\approx \int -f(x) \log_2 \left(f(x)\right) dx - \log_2 (h) \int f(x)dx \\ &= \int -f(x_i) \log_2 \left(f(x_i)\right) dx - \log_2 (h)\,. \end{aligned} \tag{1}$$

By taking the limit of $H(X_h) + \log_2(h)$ as $h \to 0$, we obtain the definition of the differential entropy[1]

$$H(f) := -\int f(x) \log_2(f(x))dx. \tag{2}$$

[1] Very often $\ln$ is used instead of $\log_2$, also in this article we use this convention.

In this paper, we follow a different approach. Instead of looking for the best type of coding for a given dataset, we use standard binary coding[2] and we search for the optimal center a of the coordinate system so that the mean amount of bits needed to code the dataset is minimal. The main advantage of this idea is that we do not have to fit the type of compression to a dataset. Moreover, codes are given in very simple way. This approach allows to immediately encrypt and decrypt large datasets (we use only one type of code). Clearly, classical binary code is far from being optimal but it is simple and commonly used in practise.

The number of bits needed to code $x \in \mathbb{Z}$, by the classical binary code, is given by

$$\mathrm{m}(x) \approx \log_2(\max\{1, |x|\}).$$

Consequently, the memory needed to code $x \in \mathbb{R}$ with given accuracy h is approximately

$$\mathrm{m}_h(x) = \mathrm{m}\left(\frac{x}{h}\right) \approx \log_2\left(\max\left\{1, \left|\frac{x}{h}\right|\right\}\right).$$

Our aim is to find the place where to put the origin a of the coordinate system so that the mean amount of bits needed to code randomly chosen an element from a sample from probability distribution of X is minimal. In other words, we want to find $a \in \mathbb{R}$ such that

$$\mathrm{E}(\mathrm{m}_h(X - a)) = \int_{\mathbb{R}} \mathrm{m}_h(x - a) f(x) dx$$

attains minimum, where f is the density distribution of the random variable X.

Our paper is arranged as follows. In the next section we show that under reasonable assumptions, the choice of a does not depend on h asymptotically. This reasoning is similar to the derivation of the differential entropy (1).

In the third section, we consider the typical situation when the density distribution of a random variable X is not known. We use a standard kernel method to estimate the density f. Working with the estimation is possible, but from the numerical point of view, complicated and not effective. So in the next section we show reasonable approximation which has better properties.

In the fourth section, we present our main algorithm and in Appendix B we put full implementation.

In the last section, we present how our method works on typical datasets.

[2]In the classical binary code we use one bit for the sign and then the standard binary representation. This code is not prefix so we have to mark ends of words. Similar coding is used in the decimal numeral system.

2. The kernel density estimation

As it was mentioned in the previous section, our aim is to minimize, for a fixed h, the function $a \to \mathrm{E}(\mathrm{m}_h(X-a))$. In this chapter, we show that the choice of a (asymptotically) does not depend on h. In Theorem 2.1, we use a similar reasoning as in the derivation of differential entropy (1) and we show that it is enough to consider the function

$$M_f(a) := \int_{\mathbb{R}} \ln(|x-a|) f(x) dx.$$

Theorem 2.1. *We assume that the random variable X has locally bounded density distribution $f\colon \mathbb{R} \to \mathbb{R}$. If $M_f(a) < \infty$ for $a \in \mathbb{R}$, then*

$$\lim_{h\to 0} \left| \mathrm{E}(\mathrm{m}_h(X-a)) - \frac{1}{\ln(2)} M_f(a) + \log_2(h) \right| = 0 \quad \textit{for } a \in \mathbb{R}.$$

Proof. Let $a \in \mathbb{R}$ be fixed. Then

$$\begin{aligned}
&\mathrm{E}(\mathrm{m}_h(X-a)) = \\
&\int_{\mathbb{R}} \mathrm{m}_h(x-a) f(x) dx = \int_{\mathbb{R}} \log_2 \left(\max\left\{ 1, \frac{|x-a|}{h} \right\} \right) f(x) dx = \\
&= \int_{\mathbb{R}\setminus(a-h,a+h)} \log_2 \left(\frac{|x-a|}{h} \right) f(x) dx = \\
&= \int_{\mathbb{R}\setminus(a-h,a+h)} \log_2 \left(|x-a| \right) f(x) dx + \int_{\mathbb{R}\setminus(a-h,a+h)} \log_2 \left(h \right) f(x) dx = \\
&= \int_{\mathbb{R}} \log_2 \left(|x-a| \right) f(x) dx - \int_{(a-h,a+h)} \log_2 \left(|x-a| \right) f(x) dx + \\
&\int_{\mathbb{R}} \log_2 \left(h \right) f(x) dx - \int_{(a-h,a+h)} \log_2 \left(h \right) f(x) dx.
\end{aligned}$$

Since the function f is a locally bounded density distribution so

$$-\int_{(a-h,a+h)} \log_2 \left(|x-a| \right) f(x) dx - \int_{(a-h,a+h)} \log_2 \left(h \right) f(x) dx = 0.$$

Consequently

$$\lim_{h\to 0} \left| \mathrm{E}(\mathrm{m}_h(X-a)) - \frac{1}{\ln(2)} M_f(a) + \log_2(h) \right| = 0 \quad \text{for } a \in \mathbb{R}.$$

□

As we see, when we increase the accuracy ($h \to 0$) of coding the shape of the function $\mathrm{E}(\mathrm{m}_h(X-a))$ stabilizes (modulo subtraction of $\log_2(h)$).

Example 2.1. *Let f be a uniform density distribution on interval $\left[-\frac{1}{2},\frac{1}{2}\right]$. Then*

$$M_f(a) = \int_{\mathbb{R}} \ln(|x-a|) f(x) dx = \int_{-\frac{1}{2}}^{\frac{1}{2}} \ln(|x-a|) dx.$$

Moreover, since

$$\int \ln(x-a)dx = \ln(x-a)x - \ln(x-a)a - x - a,$$

then we have

$$M_f(a) = \begin{cases} \ln\left(|\frac{1}{2}-a|\right)\left(\frac{1}{2}-a\right) + \ln\left(|\frac{1}{2}+a|\right)\left(\frac{1}{2}+a\right) - 1 & \textit{for } |a| \neq \frac{1}{2}, \\ -1 & \textit{for } |a| = \frac{1}{2}. \end{cases}$$

Function $M_f(a)$ is presented in Fig. 1. It is easy to see that $\min\left(M_f(a)\right) = 0$.

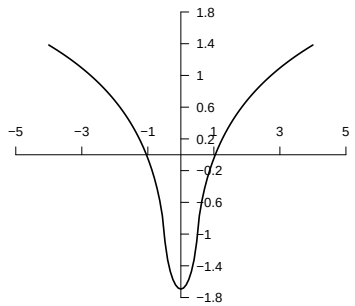

Figure 1: Function $M_f(a)$ constructed for uniform density on $\left[-\frac{1}{2},\frac{1}{2}\right]$.

In a typical situation, we do not know the density distribution f of random variable X. We only have a sample $S = (x_1, \ldots, x_N)$ from X. To approximate f we use kernel method [8]. For the convenience of the reader and to establish

the notation we shortly present this method. The kernel $K\colon \mathbb{R} \to \mathbb{R}$ is simply a function satisfying the following condition

$$\int_{\mathbb{R}} K(x)dx = 1.$$

Usually, K is a symmetric probability density function. The kernel estimator of f with kernel K is defined by

$$\bar{f}(x) := \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right),$$

where h is the window width[3]. Asymptotically optimal (for $N \to \infty$) choice of kernel K in class of symmetric and square-integrable functions is the Epanechnikov kernel [4]

$$K(x) = \begin{cases} \frac{3}{4}(1 - x^2) & \text{for } |x| < 1, \\ 0 & \text{for } |x| \geq 1. \end{cases}$$

Asymptotically optimal choice of window width (under the assumption that density is Gaussian) is given by

$$h \approx 2.35sN^{-\frac{1}{5}}, \quad \text{where} \quad s = \sqrt{\sum_{i=1}^{N} \frac{(x_i - m(S))^2}{N - 1}}.$$

Thus our aim is to minimize the function

$$a \to M_S(a) := \frac{1}{Nh} \sum_{i=1}^{N} \int_{x_i - h}^{x_i + h} \ln(|x - a|) \frac{3}{4} \left(1 - \left(\frac{x - x_i}{h}\right)^2\right) dx.$$

To compute $M_S(a)$, we analyse the function $L\colon \mathbb{R} \to \mathbb{R}$ (see Fig. 2) given by:

$$L\colon \mathbb{R} \ni a \to \frac{3}{4} \int_{-1}^{1} \ln(|x - a|) \left(1 - x^2\right) dx.$$

[3] Also called "the smoothing parameter" or "bandwidth" by some authors.

[4] Often a rescaled (normalized) Epanechnikov kernel is used.

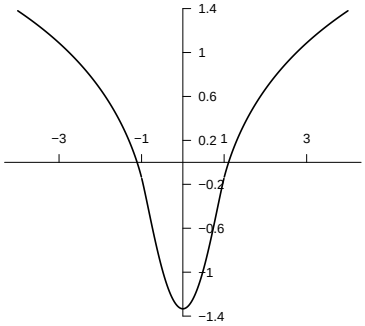

Figure 2: Function L.

Lemma 2.1. *We have*

$$M_S(a) = \ln(h) + \frac{1}{N}\sum_{i=1}^{N} L\left(\frac{x_i - a}{h}\right).$$

Proof. By simple calculations we obtain

$$\begin{aligned}
M_S(a) &= \\
&= \frac{1}{Nh}\sum_{i=1}^{N}\int_{x_i-h}^{x_i+h} \ln(|x-a|)\frac{3}{4}\left(1-\left(\frac{x-x_i}{h}\right)^2\right)dx = \\
&= \left|\begin{array}{c} y = \frac{x-x_i}{h} \\ dy = \frac{dx}{h} \\ x = hy + x_i \end{array}\right| = \frac{1}{N}\sum_{i=1}^{N}\frac{3}{4}\int_{-1}^{1} \ln\left(h\left|y + \frac{x_i - a}{h}\right|\right)\left(1-y^2\right)dy = \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{3}{4}\int_{-1}^{1} \ln\left(\left|y + \frac{x_i - a}{h}\right|\right)\left(1-y^2\right)dy + \ln(h) = \ln(h) + \frac{1}{N}\sum_{i=1}^{N} L\left(\frac{x_i - a}{h}\right).
\end{aligned}$$

□

3. Approximation of the function L

As it was shown in the previous section, the crucial role in our investigation is played by the function L and therefore, in this chapter we study its basic properties. Let us begin with the exact formula for L.

Proposition 3.1. *The function L is given by the following formula*

$$L(a) = \begin{cases} \frac{1}{2}\ln(|1-a^2|) + \left(\frac{3}{4}a - \frac{1}{4}a^3\right)\ln\left(\left|\frac{1+a}{1-a}\right|\right) + \frac{1}{2}a^2 - \frac{4}{3} & \text{for } |a| \neq 1, \\ \ln(2) - \frac{5}{6} & \text{for } |a| = 1. \end{cases}$$

Moreover, L is even, $L(0) = -\frac{4}{3}$ and $\lim_{|a|\to\infty}(L(a) - \ln(|a|)) = 0$.

Proof. We consider the case when $a > 1$ (by similar calculation, we can get this result for all $a \in \mathbb{R}$)

$$\begin{aligned} &\int \ln(x-a)\left(1-x^2\right)dx = \\ &\quad = \begin{vmatrix} \ln(x-a) & 1-x^2 \\ \frac{1}{x-a} & x - \frac{1}{3}x^3 \end{vmatrix} = \left(x - \frac{1}{3}x^3\right)\ln(|x-a|) - \int \frac{x - \frac{1}{3}x^3}{x-a}dx = \\ &\quad = \ln(x-a)\left(-\frac{1}{3}x^3 + x - a + \frac{1}{3}a^3\right) - x + a + \frac{1}{9}x^3 + \frac{1}{6}x^2 a + \frac{1}{3}xa^2 - \frac{11}{18}a^3. \end{aligned}$$

Consequently

$$L(a) = \begin{cases} \frac{1}{2}\ln(|1-a^2|) + \left(\frac{3}{4}a - \frac{1}{4}a^3\right)\ln\left(\left|\frac{1+a}{1-a}\right|\right) + \frac{1}{2}a^2 - \frac{4}{3} & \text{for } |a| \neq 1, \\ \ln(2) - \frac{5}{6} & \text{for } |a| = 1. \end{cases}$$

As a simple corollary, we obtain that L is even and $L(0) = -\frac{4}{3}$. To show the last property, we use the equality

$$\frac{3}{4}\int_{-1}^{1} \ln(|x-a|)\left(1-x^2\right)dx = \frac{3}{4}\int_{-1}^{1} \ln(|x+a|)\left(1-x^2\right)dx.$$

Then for $|a| > 1$, we obtain

$$\begin{aligned} \frac{3}{4}\int_{-1}^{1} \ln(|x-a|)\left(1-x^2\right)dx &= \frac{3}{8}\int_{-1}^{1} \left(\ln(|x-a|) + \ln(|x+a|)\right)\left(1-x^2\right)dx = \\ &= \frac{3}{8}\int_{-1}^{1} \ln\left(a^2\right)\left(1-x^2\right)dx + \frac{3}{8}\int_{-1}^{1} \ln\left(1 - \frac{x^2}{a^2}\right)\left(1-x^2\right)dx = \\ &= \frac{1}{2}\ln(a^2) + \frac{3}{8}\int_{-1}^{1} \ln\left(1 - \frac{x^2}{a^2}\right)\left(1-x^2\right)dx. \end{aligned}$$

Since

$$\frac{3}{8}\int_{-1}^{1}\ln\left(1-\frac{x^2}{a^2}\right)\left(1-x^2\right)dx \in \left[\min_{x\in[-1,1]}\left(\frac{1}{2}\ln\left(1-\frac{x^2}{a^2}\right)\right), \max_{x\in[-1,1]}\left(\frac{1}{2}\ln\left(1-\frac{x^2}{a^2}\right)\right)\right] =$$

$$= \left[\left(\frac{1}{2}\ln\left(1-\frac{1}{a^2}\right)\right), 0\right],$$

we get

$$0 \geq \lim_{a\to\infty}\left(L(a)-\ln(a)\right) \geq \lim_{a\to\infty}\left(\frac{1}{2}\ln\left(1-\frac{1}{a^2}\right)\right) = 0.$$

□

From the numerical point of view, the use of the function L (Fig. 2) is complicated and not effective. The main problem is connected with a possible numerical instability for a close to 1. Moreover, in our algorithm we use the first derivative (more information in the next chapter and Appendix A) by considering the function

$$\mathbb{R}_+ \to L'(\sqrt{a}) = \begin{cases} \frac{3}{8}\ln\left(\left|\frac{\sqrt{a}-1}{\sqrt{a}+1}\right|\right)a + \frac{1}{\sqrt{a}}\ln\left(\left|\frac{\sqrt{a}+1}{\sqrt{a}-1}\right|\right) + 2\sqrt{a} & \text{for } a \neq 1, \\ \frac{3}{4} & \text{for } a = 1. \end{cases}$$

In this case, we have numerical instability for a close to $0, 1$. Thus, instead of L, we use Cauchy M-estimator [10] $\bar{L}\colon \mathbb{R} \to \mathbb{R}$ which is given by

$$\bar{L}(a) := \frac{1}{2}\ln(e^{-\frac{8}{3}} + a^2).$$

The errors caused by the approximation are reasonably small in relation to those connected with kernell estimation[5].

Observation 3.1. *The function $\bar{L}$ is analytic, even, $\bar{L}(0) = L(0)$ and $\lim_{|a|\to\infty}(\bar{L}(a) - L(a)) = 0$.*

Consequently, the problem of finding the optimal (respectively to the needed memory) center of the coordinate system can be well approximated by searching for the global minimum of the function

$$\bar{M}_S(a) := \frac{1}{N}\sum_{i=1}^{N}\bar{L}\left(\frac{a-x_i}{h}\right) \text{ for } a \in \mathbb{R}.$$

[5]As it was said in this method one assume that dataset is realization of Gaussian random variable while usually, in practise, does not have to.

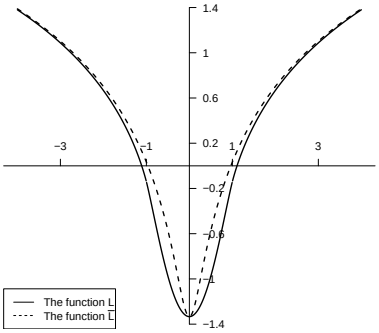

Figure 3: Comparison of functions L and $\bar{L}$.

4. Search for the minimum

In this section, we present a method of finding the minimum of the function $\bar{M}_S(a)$. We use the robust technique which is called M–estimator [4].

Remark 4.1. *For the convince of the reader we shortly present the standard use of the M–estimator's method. Let $\{x_1, \ldots, x_n\}$ be a given dataset. We are looking for the best representation of the points*

$$\min_a \sum_i (x_i - a)^2.$$

Normally we choose barycentre of the data but elements which are fare from the center usually courses undesirable effect. The M–estimators try to reduce the effect of outliers by replacing the squares by another function (in our case Cauchy M-estimator [10])

$$\min_a \sum_i L(x_i - a),$$

where L is a symmetric, positive–definite function with a unique minimum at zero, and is chosen to be less increasing than square. Instead of solving directly this problem, one usual implements an Iterated Reweighted Least Squares method (IRLS) [see Appendix A]. In our case we are looking for

$$\min_a \sum_i L(x_i - a),$$

where L is interpreted as a function which describes the memory needed to code x_i with respect to a.

Our approach based on Iterated Reweighted Least Squares method (IRLS), is similar to that presented in [9] and [6]. For convenience of the reader, we include the basic theory connected with this method in Appendix A.

In our investigations the crucial role is played by the following proposition.

Corollary IRLS (see Appendix A)**.** *Let $f : \mathbb{R}_+ \to \mathbb{R}_+$, $f(0) = 0$ be a given concave and differentiable function and let $S = \{x_i, \ldots, x_N\}$ be a given data-set. We consider the function*

$$F(a) := \sum_i f(|a - x_i|^2) \quad \textit{for } a \in \mathbb{R}.$$

Let $\bar{a} \in \mathbb{R}$ be arbitrarily fixed and let

$$w_i = f'(|x_i - \bar{a}|^2) \quad \textit{for } i = 1 \ldots N \quad \textit{and} \quad \bar{a}_w = \frac{1}{\sum_{i=0}^{N} w_i} \sum_{i=0}^{N} w_i x_i.$$

Then

$$F(\bar{a}_w) \leq F(\bar{a}).$$

Making use of the above corollary, by substitution $\bar{a} \to \bar{a}_w$ in each step we come closer to a local minimum of the function F. It is easy to see, that $\bar{L}$ defined in the previous section, satisfies the assumptions of Corollary IRLS. Let $S = (x_1, \ldots, x_N)$ be a given realization of the random variable X and let

$$h = 2.35 N^{-\frac{1}{5}} \sqrt{\sum_{i=1}^{N} \frac{(x_i - m(S))^2}{N - 1}}.$$

The algorithm (based on IRLS) can be described as follows.

Algorithm.

initialization
stop condition
$\varepsilon > 0$
initial condition
$j = 0$
$a_j = m(S)$
repeat
calculate
$w_i := \left(e^{-\frac{8}{3}} + \frac{1}{h^2}(x_i - a_j)^2\right)^{-1}$ *for* $i = 1, \ldots, N$
$j = j + 1$
$a_j = \frac{1}{\sum_{i=0}^{N} w_i} \sum_{i=0}^{N} w_i x_i$
until $|a_j - a_{j-1}| < \varepsilon$

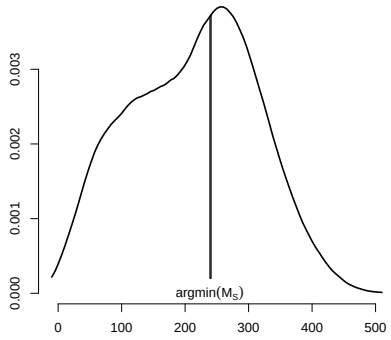

Figure 4: Estimation of the density distribution for the Forex data.

The first initial point can be chosen in many different ways. We usually start from the barycentre of the dataset because, for equal weights, the barycentre is the best approximation of the sample (for full code written in R Project, see Appendix B).

Now we show how our method works in practise.

Example 4.1. *Let S be the sample of the index of USD/EUR from Forex stoke [2]. The density obtain by the kernel method is presented at Fig. 4. As a result*

of the algorithm we obtained $\mathrm{alg\_centre}(S) = 238.4174$*. We compare our result with the global minimum* $\mathrm{argmin}(M_S) = 239.509$ *and the barycentre of data* $m(S) = 212.8004$*:*

$$\min M_S = 4.0477,$$

$$M_S(\mathrm{alg\_centre}(S)) = 4.0478,$$

$$M_S(m(S)) = 4.0768.$$

As we see the difference between $\min M_S$ *and* $M_S(\mathrm{alg\_centre}(S))$ *is small. Moreover, the barycentre gives a good approximation of the memory centre but, as we see in next examples, the difference can be large for not uni–modal densities.*

In the next step we consider a random variables of the form

$$X := p_1 \cdot X_1 + p_2 \cdot X_2$$

where X_1, X_2 are two normal random variables[6] or two uniform random variables[7].

In Table 1, we present comparison of the result of our algorithm and global minimum of the function M_S where S is the realization of random variable X of size 500 with different parameters. As we see, in the second and the third columns the algorithm which uses the function $\bar{L}$, gives a good approximation of the minimum for the function L. It means that the use of $\bar{L}$ from the third section is reasonable and causes minimal errors.

In our cases, we obtained a good approximation of the global minimum. Moreover, the difference between the minimum of the original function and the result of our algorithm is small (see fifth column). Consequently, we see that the barycentre of data sets is a good candidate for initial point.

Clearly (see sixth column), we see that the barycentre of a data is not a good approximation of the memory center, especially in the situation of not uni-modal densities.

5. Appendix A

In the fourth section, we presented a method for finding the minimum of the function $\bar{M}_S(a)$. Our approach based on the Iterated Reweighted Least Squares

[6]The normal random variables with means m and the standard deviation s we denote by $N_{(m,s)}$.

[7]The uniform random variable on the interval $[a, b]$ we denote by $U_{[a,b]}$.

Model $p_1 \cdot X_1 + p_2 \cdot X_2$	a_r	a_m	$M_S(a_r) - M_S(a_m)$	$M_S(m(S)) - M_S(a_m)$
$0.6N_{(-1,1)} + 0.4N_{(1,1)}$	-0.347	-0.379	0.00017	0.00658
$0.4N_{(-6,1)} + 0.6N_{(6,1)}$	5.803	5.676	0.00079	0.55374
$0.4N_{(-1,1)} + 0.6N_{(1,1)}$	0.483	0.416	0.00093	0.01353
$0.3N_{(-6,0.5)} + 0.7N_{(6,1)}$	5.979	5.863	0.00089	0.57296
$0.2N_{(-2,0.5)} + 0.8N_{(3,2)}$	2.721	2.770	0.00021	0.05129
$0.6U_{[-3,-1]} + 0.4U_{[0,1]}$	-1.805	-1.824	0.00014	0.19388
$0.4U_{[-3,-1]} + 0.6U_{[0,1]}$	0.364	0.403	0.00144	0.41139
$0.3U_{[-3,-1]} + 0.7U_{[0,1]}$	0.433	0.447	0.00028	0.44633
$0.2U_{[-2,-1]} + 0.8U_{[1,2]}$	1.403	1.445	0.00265	0.38220
$0.2U_{[-5,-2]} + 0.8U_{[3,4]}$	3.464	3.439	0.00019	0.50817

Table 1: In this table we have following the notation: $a_r = \text{alg\_centre}(S)$ and $a_m = \text{argmin}(M_S)$.

algorithm (IRLS). The method can be applied in statistic and computer since. We have used them to minimize the function $\bar{M_S}(a)$. Similar approach is presented in [9] and [6]. For convenience of the reader, we show the basic theoretical information about IRLS algorithm. The main theorem, related with this method, can be formulated as follows:

Theorem IRLS. *Let $f_i : \mathbb{R}_+ \to \mathbb{R}_+$, $f(0) = 0$ be a set of concave and differentiable function and let $X = (x_i)$ be a given data-set. Let $\bar{a} \in \mathbb{R}$ be fixed. We consider functions*

$$F(a) := \sum_i f_i(|a - x_i|^2) \quad \textit{for } a \in \mathbb{R}$$

and

$$H(a) := \sum_i [f_i(|\bar{a}-x_i|^2) - f_i'(|\bar{a}-x_i|^2)|\bar{a}-x_i|^2] + f_i'(|\bar{a}-x_i|^2)|a-x_i|^2 \quad \textit{for } a \in \mathbb{R}.$$

Then $H(\bar{a}) = F(\bar{a})$ and

$$H \geq F.$$

Proof. Let i be fixed. For simplicity, we denote $f = f_i$.

Let us first observe that, without loss of generality, we may assume that $\bar{x} = 0$ (we can make an obvious substitution $a \to a + \bar{x}$).

Then since all the considered functions are radial, it is sufficient to consider the one dimensional case when $N = 1$. Thus, from now on we assume that $N = 1$ and $\bar{x} = 0$. Since the functions are even, it is sufficient to consider the situation on $\mathbb{R}_+$.

Concluding: we are given $\bar{a} \in \mathbb{R}_+$ and consider functions

$$g : \mathbb{R}_+ \ni a \to f(a^2)$$

and

$$h : \mathbb{R}_+ \ni a \to [f(\bar{a}^2) - f'(\bar{a}^2)\bar{a}^2] + f'(\bar{a}^2)a^2.$$

Clearly, $h(\bar{a}) = g(\bar{a})$. We have to show that $h \geq g$. We consider two cases.

Let us first discuss the situation on the interval $[0, \bar{a}]$. We show that $g - h$ is increasing on this interval, since coincide at $\bar{a}$ this makes the proof completed. Clearly g and h are absolutely continuous functions (since f is concave). Thus, to prove that $g - h$ is increasing on $[0, \bar{a}]$, it is sufficient to show that $g' \geq h'$ a.e. on $[0, \bar{a}]$. But f is concave, and therefore f' is decreasing, which implies that

$$g'(a) = f'(a^2)2a \geq f'(\bar{a}^2)2a = h'(a).$$

So let us consider the situation on the interval $[\bar{a}, \infty)$. We will show that $g - h$ is decreasing on $[\bar{a}, \infty)$. Since $(g - h)(\bar{a}) = 0$, this is enough. Note that

$$(g - h)'(a) = f'(a^2)2a - f'(\bar{a}^2)2a = 2a(f'(a^2) - f'(\bar{a}^2)) \leq 0.$$

Now, the assertion of the theorem is a simple consequence of the previous property. □

Now we can form the most important results:

Corollary IRLS. *Let $f : \mathbb{R}_+ \to \mathbb{R}_+$, $f(0) = 0$ be a given concave and differentiable function and let $S = \{x_i\}_{i=1}^N$ be a given data-set. We consider the function*

$$F(a) = \sum_i f(|a - x_i|^2) \quad \textit{for } a \in \mathbb{R}.$$

Let $\bar{a} \in \mathbb{R}$ be arbitrarily fixed and let

$$w_i = f'(|\bar{a} - x_i|^2) \quad \textit{for } i = 1 \dots N \quad \textit{and} \quad a_w = \frac{1}{\sum_{i=0}^N w_i} \sum_{i=0}^N w_i x_i.$$

Then

$$F(a_w) \leq F(\bar{a}).$$

Proof. Let H be defined like in Theorem IRLS

$$H(a) = \sum_i [f(|\bar{a}-x_i|^2) - f'(|\bar{a}-x_i|^2)|\bar{a}-x_i|^2] + f'(|\bar{a}-x_i|^2)|a-x_i|^2 \text{ for } a \in \mathbb{R}.$$

Moreover by Theorem IRLS we have

$$F(\bar{a}) = H(\bar{a}).$$

Function H is quadratic so the minimum is

$$a_w = \frac{1}{\sum_{i=0}^{N} w_i} \sum_{i=0}^{N} w_i x_i$$

and consequently

$$F(\bar{a}) = H(\bar{a}) \geq H(a_w).$$

□

Making use of the above theorem, we obtain a simple method of finding a better approximation of the minimum. For given $a \in \mathbb{R}$, by taking weighted average a_w (see Corollary IRLS) we find the point which reduces the value of the function. So, to find minimum, we iteratively calculate weighted barycentre of data set.

6. Appendix B

In this section we present source code of our algorithm written in R Project.

```
#The derivative of the function bar L
d_bar_L = function(s){(exp(-8/3)+s)^(-1)}

#The function which describe a single iteration
#S - data p - starting point
iteration = function(p,S){
  suma=0
  weight=0
  v=sd(S) #standard deviation
  h=2.35*v*(length(S))^(-1/5)
  for(s in S){
    weight_s=d_bar_L((abs(s-p)^2)/(h^2))
    suma=suma+weight_s*s
    weight=weight+weight_s
```

```
15    }
      suma/weight
17  }

19  #The function which determine a local minimum
    #S - data
21  #e - epsilon
    #N - maximal number of iterations
23  look_minimum=function(S,e,N){
      ans=mean(S)
25    dist=1
      while(dist>e & N>0){
27      ans_n=iteration(ans,S)
      N=N-1
29    dist=abs(ans_n-ans)
      ans=ans_n
31    }
      ans
33  }
```

In our simulation we use $\varepsilon = 0.001$ and $N = 50$.

[1] T. M. Cover, J. A. Thomas, *Elements of information theory*, Wiley-India, 1999.

[2] Forex Rate, *http://www.forexrate.co.uk/forexhistoricaldata.php*,number of point, time 1 min, date end 10/28/2011.

[3] D. R. Hankerson, G. A. Harris, P. D. Johnson, *Introduction to information theory and data compression*, Chapman & Hall/CRC, 2003.

[4] P. J. Huber, E. M. Ronchetti, *Robust statistics*, John Wiley & Sons, 2009.

[5] D. A. Huffman, *A method for the construction of minimum-redundancy codes*, Proceedings of the IRE, 40, 1098–1101, 1952.

[6] J. Idier, *Convex half-quadratic criteria and interacting auxiliary variables for image restoration*, IEEE Trans. Image Process., 10, 1001–1009, 2001.

[7] D. Salomon, G. Motta, D.C.O.N. Bryant, *Handbook of data compression*, Springer-Verlag New York Inc, 2009.

[8] B. W. Silverman, *Density estimation for statistics and data analysis*, Chapman & Hall/CRC, 1986.

[9] R. Wolke, *Iteratively reweighted least squares. A comparison of several single step algorithms for linear models*, BIT Numerical Mathematics, 32, 506–524, 1992.

[10] Z. Zhang, *Parameter estimation techniques: A tutorial with application to conic fitting*, Image and vision Computing, 59-76, 1997.