Optimal Rescaling and the Mahalanobis Distance

Przemysaw & Spurek* and Jacek Tabor

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

(*v4.4 released October 2008*)

One of the basic problems in data analysis lies in choosing the optimal rescaling (change of coordinate system) to study properties of a given data-set Y. The classical Mahalanobis approach has its basis in the classical normalization/rescaling formula $Y \ni y \to \Sigma_Y^{-1/2} \cdot (y - \mathrm{m}_Y)$, where m_Y denotes the mean of Y and Σ_Y the covariance matrix .

Based on the cross-entropy we generalize this approach and define the parameter which measures the fit of a given affine rescaling of Y compared to the Mahalanobis one. This allows in particular to find an optimal change of coordinate system which satisfies some additional conditions. In particular we show that in the case when we put origin of coordinate system in m the optimal choice is given by the transformation $Y \ni y \to \Sigma_Y^{-1/2} \cdot (y - \mathrm{m}_Y)$, where

$$\Sigma = \Sigma_Y (\Sigma_Y - \frac{(\mathrm{m} - \mathrm{m}_Y)(\mathrm{m} - \mathrm{m}_Y)^T}{1 + \|\mathrm{m} - \mathrm{m}_Y\|_{\Sigma_Y}^2})^{-1} \Sigma_Y.$$

Keywords: Maximum Likelihood Estimation; MLE; Cross Entropy; Optimal Coordinates; Mahalanobis distance

1. Introduction

One of the crucial problem in statistics, compression, discrimination analysis and in general in data mining is how to choose the optimal linear coordinate system and define distance which "optimally" underlines the internal structure of the data [1–11]. The typical answer is the Mahalanobis distance [2, 12–14], which is strictly connected with PCA [15–17]. In some other cases we choose just a weighted Euclidean distance. Our approach uses a method based on MLE (Maximum Likelihood Estimation) [18–20].

To explain it more precisely consider the data-set $Y \subset \mathbb{R}^N$. If we allow the translation of the origin of coordinate system, in one dimensional case we usually apply the normalization: $s : Y \ni y \to \sigma_Y^{-1}(y - \mathrm{m}_Y)$, which in the multivariate case is replaced by

$$s : Y \ni y \to \Sigma_Y^{-1/2}(y - \mathrm{m}_Y),$$

where m_Y denotes the mean and Σ_Y the covariance matrix of Y. Then we obtain that the coordinates are uncorrelated, and the covariance matrix equals to identity.

---

*Corresponding author. Email: przemyslaw.spurek@ii.uj.edu.pl

Taking the distance between the transformation of points x, y:

$$\|sx - sy\|^2 = (sx - sy)^T(sx - sy) = (x - y)^T \Sigma_Y^{-1}(x - y)$$

we arrive naturally at the definition of the Mahalanobis distance:

$$\|x - y\|_\Sigma^2 := (x - y)^T \Sigma^{-1}(x - y).$$

If we do not allow the translation of the origin away from zero, which in some cases is natural, we usually only scale/normalize each coordinate by dividing it by its mean. This approach usually has good results if the standard deviation is small in comparison to the mean. In the opposite case, when the mean is small, dividing by it may "unnaturally" widen the variable under consideration.

The main results of the paper provide the optimal change of coordinate system which satisfies some additional conditions. In particular we show that in the case when we put origin of coordinate system in m the optimal choice is given by the mapping $Y \ni y \to \Sigma_Y^{-1/2} \cdot (y - \mathrm{m}_Y)$, where

$$\Sigma = \Sigma_Y(\Sigma_Y - \frac{(\mathrm{m} - \mathrm{m}_Y)(\mathrm{m} - \mathrm{m}_Y)^T}{1 + \|\mathrm{m} - \mathrm{m}_Y\|_{\Sigma_Y}^2})^{-1}\Sigma_Y.$$
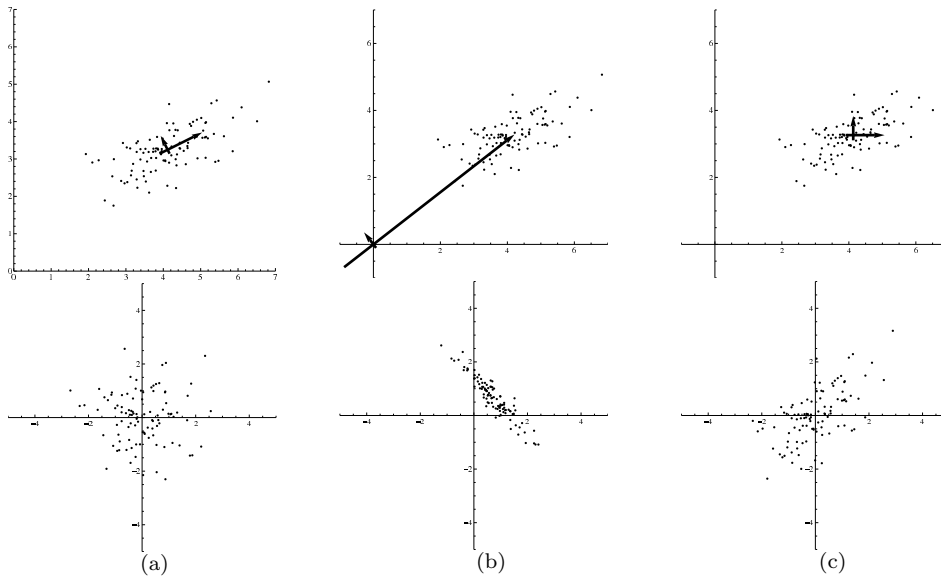

(a)   (b)   (c)

Figure 1.   Optimal coordinate systems and the data rewritten in new bases in various situations: a) optimal Mahalanobis base; b) optimal base in the center in $[0,0]$ c) optimal base in the case of optimal rescaling of each coordinates;

Example 1.1 We illustrate our results on a two-dimensional data drawn from the normal distribution with mean $\mathrm{m} = [3,4]^T$ and covariance $\Sigma = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 0.6 \end{bmatrix}$. In the Figure 1(a) we present the base which represents the best (Mahalanobis) rescaling and the data in new coordinates. In 1(b) we present optimal base in the case we do not allow to move the origin of the coordinate system from zero and the data in new base. In the Figure 1(c) we present optimal rescaling in the case when we the change of the origin, but we restrict to rescale the coordinates separately and the data in new base.

To explain more precisely what we mean by optimal coordinates (in the given class of coordinate systems) we need to introduce the function which measures the "match" of a given coordinate system to the data. Suppose that we are given a base $\mathrm{v} = (v_1, \ldots, v_N)$ of $\mathbb{R}^N$ and we put an origin of coordinate system at $\mathrm{m} \in \mathbb{R}^N$. Then by $\mathcal{N}_{[\mathrm{m},\mathrm{v}]}$ we denote the "normalized" Gaussian density with respect to the base v with center at m. In other words

$$\mathcal{N}_{[\mathrm{m},\mathrm{v}]}(m + x_1 v_1 + \ldots + x_N v_N) = \frac{1}{(2\pi)^{N/2}} e^{-(x_1^2 + \ldots + x_N^2)/2}.$$

Observe that $\mathcal{N}_{[\mathrm{m},\mathrm{v}]} = \mathcal{N}_{(\mathrm{m},(vv^T)^{-1})}$, where by $\mathcal{N}_{(\mathrm{m},\Sigma)}$ we denote the normal density with mean m and covariance Σ. Then we can measure the "match/fit" of the coordinate system by the cross-entropy

$$H^{\times}(Y \| \mathcal{N}_{[\mathrm{m},\mathrm{v}]}) \tag{1}$$

of Y with respect to $\mathcal{N}_{[\mathrm{m},\mathrm{v}]}$. One can easily observe that the minimum in (1) is attained by the Mahalanobis coordinate system, where the mean and covariance is that of Y, and thus we obtain a confirmation that the commonly used procedure is reasonable. This allows to define the match of data Y with respect to the coordinate system (or more precisely the corresponding Gaussian density) by the formula

$$\mathrm{M}(Y \| \mathcal{N}_{(\mathrm{m},\Sigma)}) := H^{\times}(Y \| \mathcal{N}_{(\mathrm{m},\Sigma)}) - H^{\times}(Y \| \mathcal{N}_Y) \geq 0, \tag{2}$$

where $\mathcal{N}_Y$ denotes the normal density with mean and covariance of Y. From the well-known formulas [21, 22] concerning relative entropy and cross-entropy:

$$
\begin{aligned}
H^{\times}(Y \| \mathcal{N}_{(\mathrm{m},\Sigma)}) &= \tfrac{N}{2} \ln(2\pi) + \tfrac{1}{2} \| \mathrm{m} - \mathrm{m}_Y \|_{\Sigma}^2 + \tfrac{1}{2}\mathrm{tr}(\Sigma^{-1}\Sigma_Y) + \tfrac{1}{2} \ln \det\Sigma, \\
H^{\times}(Y \| \mathcal{N}_Y) &= \tfrac{N}{2} \ln(2\pi e) + \tfrac{1}{2} \ln \det\Sigma_Y,
\end{aligned}
\tag{3}
$$

we can easily deduce the formula for our match function for the Gaussian distribution $\mathcal{N}(\mathrm{m}, \Sigma)$ in $\mathbb{R}^N$:

$$\mathrm{M}(Y \| \mathcal{N}_{(\mathrm{m},\Sigma)}) = \frac{1}{2} \left(\| \mathrm{m} - \mathrm{m}_Y \|_{\Sigma}^2 + \mathrm{tr}(\Sigma^{-1}\Sigma_Y) - \ln \det(\Sigma^{-1}\Sigma_Y) - N \right). \tag{4}$$

We can now consider coordinate systems identified with respective subclasses of Gaussian distributions. Consequently for a family $\mathcal{F}$ of Gaussian distributions we put

$$\mathrm{M}(Y \| \mathcal{F}) := \inf_{f \in \mathcal{F}} \mathrm{M}(Y \| f) = \inf_{f \in \mathcal{F}} H^{\times}(Y \| f) - H^{\times}(Y \| \mathcal{N}_Y).$$

Clearly, to find f which minimizes the above it is equivalent to find f which minimizes the cross-entropy $H^{\times}(Y \| \mathcal{F}) := \inf_{f \in \mathcal{F}} H^{\times}(Y \| f)$. Since cross-entropy and log-likelihood functions differ only by sign, we arrive at the typical MLE [18–20] problem.

As $\mathcal{F}$ we consider the following typical subfamilies of all Gaussians $\mathcal{G}$:

$\mathcal{G}_\mathrm{m}$ –   Gaussian densities with mean at m;

$\mathcal{G}_{s\mathrm{I}}$ –   Gaussian densities with covariance proportional to identity;

$\mathcal{G}_{\mathrm{m},s\mathrm{I}}$ –   Gaussian densities with mean at m and covariance proportional to identity;

$\mathcal{G}_\mathrm{diag}$ –   Gaussian densities with diagonal covariance;

$\mathcal{G}_\mathrm{m,diag}$ –   Gaussian densities with mean at m and diagonal covariance.

Observe for example that the use of family $\mathcal{G}_\mathrm{m,diag}$ means that we consider the coordinate systems which have origin at m and which have axes parallel to the original (cartesian) ones. Moreover, if we have found the optimal Gaussian $\mathcal{N}_{(\mathrm{m},\Sigma)}$ in our class of densities (identified with coordinate systems), we can transform the data Y into those coordinates by the affine transformation $Y \ni y \to \Sigma^{-1/2}(y - \mathrm{m}) \in \mathbb{R}^N$.

2.   Rescaling

Assume that we have fixed the origin of the coordinate system at the point m and that we want to find how we should (uniformly) rescale the coordinates to optimally fit the data. This means that we search for s such that $s \to \mathrm{M}(Y\|\mathcal{N}_{(\mathrm{m}_Y,s\mathrm{I})})$ attains minimum. Since

$$H^\times(Y\|\mathcal{N}_{(\mathrm{m}_Y,s\mathrm{I})}) = \frac{1}{2}\big((\mathrm{tr}(\Sigma_Y) + \|\mathrm{m} - \mathrm{m}_Y\|^2)s^{-1} + N\ln s + N\ln(2\pi)\big), \quad (5)$$

by the trivial calculations we obtain that the above function attains its minimum

$$\frac{N}{2}\big(\ln\frac{\mathrm{tr}(\Sigma_Y) + \|\mathrm{m} - \mathrm{m}_Y\|^2}{N} + \ln(2\pi e)\big) \tag{6}$$

for

$$s = \frac{\mathrm{tr}(\Sigma_Y) + \|\mathrm{m} - \mathrm{m}_Y\|^2}{N}. \tag{7}$$

Thus applying (3) to (2) we have arrived at the following theorem.

THEOREM 2.1 *Let Y be a data-set with invertible covariance matrix and* m *be fixed. Then* $\mathrm{M}(Y\|\mathcal{G}_{\mathrm{m},s\mathrm{I}})$ *is minimized for* $s = (\mathrm{tr}(\Sigma_Y) + \|\mathrm{m} - \mathrm{m}_Y\|^2)/N$, *and equals*

$$\mathrm{M}(Y\|\mathcal{G}_{\mathrm{m},s\mathrm{I}}) = \frac{N}{2}\ln\frac{\mathrm{tr}(\Sigma_Y) + \|\mathrm{m} - \mathrm{m}_Y\|^2}{N} - \frac{1}{2}\ln\det\Sigma_Y.$$

If we allow in above Theorem the change of the origin, to minimize the value of M we have to clearly put m at m_Y:

COROLLARY 2.2 *Let Y be a data-set with invertible covariance matrix. Then* $\mathrm{M}(Y\|\mathcal{G}_{s\mathrm{I}})$ *is minimized for* $\mathrm{m} = \mathrm{m}_Y$, $s = \frac{1}{N}\mathrm{tr}(\Sigma_Y)$, *and equals*

$$\mathrm{M}(Y\|\mathcal{G}_{s\mathrm{I}}) = \frac{N}{2}\ln(\mathrm{tr}(\Sigma_Y)/N) - \frac{1}{2}\ln\det\Sigma_Y.$$

Remark 1 Assume that we want to move the origin to m, and uniformly rescale.

Then Theorem 2.1 implies that

$$y \to (y - \mathrm{m})/\sqrt{\frac{1}{N}(\mathrm{tr}(\Sigma_Y) + \|\mathrm{m} - \mathrm{m}_Y\|^2)}$$

is the optimal rescaling of this type. Thus in the case of univariate data, the optimal rescaling when we do not change the origin of the coordinate system is given by $y \to y/\sqrt{\sigma_Y^2 + \mathrm{m}_Y^2} = y/\sqrt{E(Y^2)}$.

We consider the case when we allow to rescale each coordinate Y_i of $Y = (Y_1, \ldots, Y_N)$ separately. Then the optimal change of coordinates of this form is given by optimal rescaling on each coordinate.

COROLLARY 2.3 *Let Y be a data-set with invertible covariance matrix and m be fixed. Then $\mathrm{M}(Y\|\mathcal{G}_{\mathrm{m,diag}})$ is minimized for $s_i = (\Sigma_Y)_{ii} + |\mathrm{m}_i - (\mathrm{m}_Y)_i|^2$, and equals*

$$\mathrm{M}(Y\|\mathcal{G}_{\mathrm{m,diag}}) = \frac{1}{2}\sum_{i=1}^{N}\ln\left((\Sigma_Y)_{ii} + |\mathrm{m}_i - (\mathrm{m}_Y)_i|^2\right) - \frac{1}{2}\ln\det\Sigma_Y.$$

Proof For $\mathcal{N}_{(\mathrm{m,diag}(s_1,\ldots,s_n))} \in \mathcal{G}_{\mathrm{m,diag}}$ we have

$$\mathrm{M}(Y\|\mathcal{N}_{(\mathrm{m,diag}(s_1,\ldots,s_n))}) = H^{\times}(Y\|\mathcal{N}_{(\mathrm{m,diag}(s_1,\ldots,s_n))}) - H^{\times}(Y\|\mathcal{N}_Y)$$
$$= \sum_{i=1}^{N} H^{\times}(Y_i\|\mathcal{N}_{(\mathrm{m}_i,s_i)}) - H^{\times}(Y\|\mathcal{N}_Y).$$

By applying (6) to the univariate data Y_i and (3) we obtain that the minimum of $\mathrm{M}(Y\|\mathcal{N}_{(\mathrm{m,diag}(s_1,\ldots,s_n))})$ is realized for $s_i = (\Sigma_Y)_{ii} + |\mathrm{m}_i - (\mathrm{m}_Y)_i|^2)$ and equals

$$\frac{1}{2}\sum_{i=1}^{N}\left(\ln((\Sigma_Y)_{ii} + |\mathrm{m}_i - (\mathrm{m}_Y)_i|^2) + \ln(2\pi e)\right) - \frac{1}{2}\left(N\ln(2\pi e) + \ln\det\Sigma_Y\right)$$

$$= \frac{1}{2}\sum_{i=1}^{N}\ln((\Sigma_Y)_{ii} + |\mathrm{m}_i - (\mathrm{m}_Y)_i|^2) - \frac{1}{2}\ln\det\Sigma_Y.$$

∎

If we additionally allow the change of the origin, we should put $\mathrm{m} = \mathrm{m}_Y$ and the rescaling takes the form $y \to (y - \mathrm{m}_Y)/\sqrt{\mathrm{tr}(\Sigma_Y)/N}$.

Now we focus our attention on the problem how to find the optimal coordinate system in the general case. To do so we first need to present a simple consequence of the famous von Neuman trace inequality [23, 24].

Theorem [von Neumann trace inequality]. *Let E, F be complex $N \times N$ matrices. Then*

$$|\mathrm{tr}(EF)| \leq \sum_{i=1}^{N} s_i(E) \cdot s_i(F), \tag{8}$$

where $s_i(D)$ denote the ordered (decreasingly) singular values of matrix D.

Let us recall that for the symmetric positive matrix its eigenvalues coincide with singular values. Given $\lambda_1, \ldots, \lambda_N \in \mathbb{R}$ by $S_{\lambda_1, \ldots, \lambda_N}$ we denote the set of all symmetric matrices with eigenvalues $\lambda_1, \ldots, \lambda_N$.

PROPOSITION 2.4 *Let B be a symmetric nonnegative matrix with eigenvalues $\beta_1 \geq \ldots \geq \beta_N \geq 0$. Let $0 \leq \lambda_1 \leq \ldots \leq \lambda_N$ be fixed. Then*

$$\min_{A \in S_{\lambda_1, \ldots, \lambda_N}} \mathrm{tr}(AB) = \sum_i \lambda_i \beta_i.$$

Proof Let e_i denote the orthogonal basis build from the eigenvectors of B, and let operator $\bar{A}$ be defined in this base by $\bar{A}(e_i) = \lambda_i e_i$. Then trivially

$$\min_{A \in S_{\lambda_1, \ldots, \lambda_N}} \mathrm{tr}(AB) \leq \mathrm{tr}(\bar{A}B) = \sum_i \lambda_i \beta_i.$$

To prove the inverse inequality we will use the von Neumann trace inequality. Let $A \in S_{\lambda_1, \ldots, \lambda_N}$ be arbitrary. We apply the inequality (8) for $E = \lambda_N I - A$, $F = B$. Since E and F are symmetric nonnegatively defined matrices, their eigenvalues $\lambda_N - \lambda_i$ and β_i coincide with singular values, and therefore by (8)

$$\mathrm{tr}((\lambda_N I - A)B) \leq \sum_i (\lambda_N - \lambda_i)\beta_i = \lambda_N \sum_i \beta_i - \sum_i \lambda_i \beta_i. \tag{9}$$

Since

$$\mathrm{tr}((\lambda_N I - A)B) = \lambda_N \sum_i \beta_i - \mathrm{tr}(AB),$$

from inequality (9) we obtain that $\mathrm{tr}(AB) \geq \sum_i \lambda_i \beta_i$.   ∎

Now we proceed to the main result of the paper.

THEOREM 2.5 *Let Y be a data-set and $\mathrm{m} \in \mathbb{R}^N$ be fixed. Then*

$$\mathrm{M}(Y\|\mathcal{G}_{\mathrm{m}}) = \frac{1}{2}\ln(1 + \|\mathrm{m} - \mathrm{m}_Y\|_{\Sigma_Y}^2)$$

and is attained for the density $\mathcal{N}(\mathrm{m}, \Sigma) \in \mathcal{G}_{\mathrm{m}}$, where

$$\Sigma = \Sigma_Y (\Sigma_Y - \frac{(\mathrm{m} - \mathrm{m}_Y)(\mathrm{m} - \mathrm{m}_Y)^T}{1 + \|\mathrm{m} - \mathrm{m}_Y\|_{\Sigma_Y}^2})^{-1}\Sigma_Y.$$

Proof Let us first observe that by applying substitution

$$A = \Sigma_Y^{1/2}\Sigma^{-1}\Sigma_Y^{1/2}, v = \Sigma_Y^{-1/2}(\mathrm{m} - \mathrm{m}_Y),$$

we obtain

$$
\begin{aligned}
H^{\times}(Y\|\mathcal{N}_{(\mathrm{m}, \Sigma)}) &= \tfrac{1}{2}\left(\mathrm{tr}(\Sigma^{-1}\Sigma_Y) + \|\mathrm{m} - \mathrm{m}_Y\|_{\Sigma}^2 + \ln\det\Sigma + N\ln(2\pi)\right) \\
&= \tfrac{1}{2}\big(\mathrm{tr}(\Sigma^{-1}\Sigma_Y) + (\mathrm{m} - \mathrm{m}_Y)^T\Sigma^{-1}(\mathrm{m} - \mathrm{m}_Y) - \ln\det\Sigma^{-1}\Sigma_Y \\
&\qquad + \ln\det\Sigma_Y + N\ln(2\pi)\big) \\
&= \tfrac{1}{2}\left(\mathrm{tr}(A) + v^T A v - \ln\det A + \ln\det\Sigma_Y + N\ln(2\pi)\right).
\end{aligned}
\tag{10}
$$

Observe that A is then a symmetric positive matrix, and that given a symmetric positive matrix A we can uniquely determine Σ by the formula

$$\Sigma = \Sigma_Y^{1/2} A^{-1} \Sigma_Y^{1/2}. \tag{11}$$

Thus finding minimum of (10) reduces to finding a symmetric positive matrix A which minimize the value of

$$\operatorname{tr}(A) + v^T A v - \ln \det A. \tag{12}$$

Let us first consider $A \in S_{\lambda_1,\dots,\lambda_N}$, where $0 < \lambda_1 \leq \dots \leq \lambda_N$ are fixed. Our aim is to minimize

$$v^T A v = \operatorname{tr}(v^T A v) = \operatorname{tr}(A \cdot (v v^T)).$$

We fix an orthonormal base such that $v/\|v\|$ is its first element, and then by applying von Neumann trace formula we obtain that the above minimizes when v is the eigenvector of A corresponding to λ_1, and thus the minimum equals

$$\lambda_1 \|v\|^2.$$

Consequently we arrive at the minimization problem

$$\lambda_1(1 + \|v\|^2) + \sum_{i>1} \lambda_i - \sum_i \ln \lambda_i.$$

Now one can easily verify that the minimum of the above is realized for

$$\lambda_1 = 1/(1 + \|v\|^2), \lambda_i = 1 \text{ for } i > 1.$$

and then (12) equals

$$N + \ln(1 + \|\mathrm{m} - \mathrm{m}_Y\|_{\Sigma_Y}^2),$$

while the formula for A minimizing it is given by

$$A = \mathrm{I} - \frac{v v^T}{1 + \|v\|^2}.$$

Consequently then the value of (10) is

$$\frac{1}{2} \left(\ln(1 + \|\mathrm{m} - \mathrm{m}_Y\|_{\Sigma_Y}^2) + \ln |\Sigma_Y| + N \ln(2\pi e) \right).$$

and is attained by (11) for

$$\Sigma = \Sigma_Y^{1/2} (I - \frac{\Sigma_Y^{-1/2}(\mathrm{m} - \mathrm{m}_Y)(\mathrm{m} - \mathrm{m}_Y)^T \Sigma_Y^{-1/2}}{1 + \|\mathrm{m} - \mathrm{m}_Y\|_{\Sigma_Y}^2})^{-1} \Sigma_Y^{1/2}$$

$$= \Sigma_Y (\Sigma_Y - \frac{(\mathrm{m} - \mathrm{m}_Y)(\mathrm{m} - \mathrm{m}_Y)^T}{1 + \|\mathrm{m} - \mathrm{m}_Y\|_{\Sigma_Y}^2})^{-1} \Sigma_Y.$$

∎

By the above theorem we get the formula for $H^{\times}(Y\|\mathcal{G}_{\mathrm{m}})$.

COROLLARY 2.6 *Let* $\mathrm{m} \in \mathbb{R}^N$ *be fixed. Then*

$$H^{\times}(Y\|\mathcal{G}_{\mathrm{m}}) = \frac{1}{2}\left(\ln(1 + \|\mathrm{m} - \mathrm{m}_Y\|_{\Sigma_Y}^2) + \ln|\Sigma_Y| + N\ln(2\pi e)\right),$$

and is attained for the density $\mathcal{N}(\mathrm{m}, \Sigma)$, *where* $\Sigma = \Sigma_Y(\Sigma_Y - \frac{(\mathrm{m}-\mathrm{m}_Y)(\mathrm{m}-\mathrm{m}_Y)^T}{1+\|\mathrm{m}-\mathrm{m}_Y\|_{\Sigma_Y}^2})^{-1}\Sigma_Y$.

$\mathcal{F}$	$M^{\times}(Y\|\mathcal{F})$
$\mathcal{G}$	0
$\mathcal{G}_{\mathrm{m}}$	$\frac{1}{2}\left(\ln(1 + \|\mathrm{m} - \mathrm{m}_Y\|_{\Sigma_Y}^2) + \ln|\Sigma_Y| + N\ln(2\pi e)\right)$
$\mathcal{G}_{sI}$	$\frac{N}{2}\ln\left(\mathrm{tr}\frac{\Sigma_Y}{N}\right) - \frac{1}{2}\ln\det\Sigma_Y$
$\mathcal{G}_{\mathrm{m},sI}$	$\frac{N}{2}\ln\left(\frac{\mathrm{tr}(\Sigma_Y)+\|\mathrm{m}-\mathrm{m}_Y\|^2}{N}\right) - \frac{1}{2}\ln\left(\det\Sigma_Y\right)$
$\mathcal{G}_{\mathrm{diag}}$	$\frac{N}{2}\ln\left(\frac{\mathrm{tr}(\Sigma_Y)}{N}\right) - \frac{1}{2}\ln\det\Sigma_Y$
$\mathcal{G}_{\mathrm{m},\mathrm{diag}}$	$\frac{1}{2}\sum_{i=1}^{N}\ln\left((\Sigma_Y)_{ii} + |\mathrm{m}_i - (\mathrm{m}_Y)_i|^2\right) - \frac{1}{2}\ln\left(\det\Sigma_Y\right)$

Table 1.   Table of $M^{\times}(Y\|\mathcal{F})$ with respect to Gaussian subfamilies.

At the end of this article we illustrate our results on the data generated from the classical Lena picture.

Example 2.7 Let us consider the classical Lena picture from The USC-SIPI Image Database (http://sipi.usc.edu/database/). First, we interpret photo as a dataset, as in the JPG compression. We do this by dividing it into 8 by 8 pixels, where each pixel is described (in RGB) by using 3 parameters. Consequently each of the pieces is represented as a vector from $\mathbb{R}^{192}$. By this operation we obtain dataset Y from $\mathbb{R}^{192}$. In Table 2 we present values of $M(Y\|\mathcal{G}_{\mathrm{m},\mathcal{F}})$ for various $\mathcal{F}$. Moreover we consider the case when the origin is at $\mathrm{m} = \left[\frac{1}{2}, \ldots, \frac{1}{2}\right]$, similarly like in JPG format.

	$M(Y\|\mathcal{G}_{\mathrm{m}})$	$M(Y\|\mathcal{G}_{\mathrm{m},\mathrm{diag}})$	$M(Y\|\mathcal{G}_{\mathrm{m},sI})$
$\mathrm{m} = \mathrm{m}_Y$	0	672.9	679.168
$\mathrm{m} = \left[\frac{1}{2}, \ldots, \frac{1}{2}\right]$	0.520298	717.217	704.124
$\mathrm{m} = [0, \ldots, 0]$	2.73125	883.227	1031

Table 2.   Values of $M(Y\|\mathcal{F})$ for different $\mathcal{F}$ and data from Example 2.7.

As a conclusion we see that the change of the origin from the mean to the point $\left[\frac{1}{2}, \ldots, \frac{1}{2}\right]$ does not "cost" us much, compared to the case when we restrict to the class of $\mathcal{G}_{\mathrm{diag}}$. Moreover we see that the fixed center in $\left[\frac{1}{2}, \ldots, \frac{1}{2}\right]$ is better then $[0, \ldots, 0]$ which shows that JPG approach is reasonable.

3.   Conclusion

In this paper we present the method of determining the optimal coordinate systems for the dataset $Y \in \mathbb{R}^N$ which meets an commonly encountered conditions. We interpret the optimal Gaussian density (in the sense of cross–entropy) as an optimal transformation of a data. Consequently we obtain the measure of the optimality of the given coordinates system (represented by Gaussian family $\mathcal{F}$)

$$\mathrm{M}(Y\|\mathcal{F}) = \inf_{f \in \mathcal{F}} H^{\times}(Y\|f) - H^{\times}(Y\|\mathcal{N}_Y).$$

We obtain estimations in various subclasses of normal densities. The results we present in Tab. 1

References

[1] I. Borg and P. J. Groenen, *Modern multidimensional scaling: Theory and applications.* Springer Verlag, 2005.

[2] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart, "The mahalanobis distance," *Chemometrics and Intelligent Laboratory Systems*, vol. 50, no. 1, pp. 1–18, 2000.

[3] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques.* Morgan kaufmann, 2006.

[4] T. Jayalakshmi and A. Santhakumaran, "Statistical normalization and back propagation for classification," *International Journal of Computer Theory and Engineering*, vol. 3, no. 1, pp. 1793–8201, 2011.

[5] T. Raykov and G. A. Marcoulides, *An introduction to applied multivariate analysis.* Psychology Press, 2008.

[6] A. C. Rencher and W. F. Christensen, *Methods of multivariate analysis*, vol. 709. Wiley, 2012.

[7] N. H. Timm, *Applied multivariate analysis.* Springer Verlag, 2002.

[8] P. R. Krishnaiah and L. N. Kanal, "Classification, pattern recognition, and reduction of dimensionality, volume 2 of handbook of statistics," *North4Holland Amsterdam*, 1982.

[9] M. P. McAssey, "An empirical goodness-of-fit test for multivariate distributions," *Journal of Applied Statistics*, no. ahead-of-print, pp. 1–12, 2013.

[10] E. J. Bedrick, "Graphical modelling and the mahalanobis distance," *Journal of Applied Statistics*, vol. 32, no. 9, pp. 959–967, 2005.

[11] W. Krzanowski, "Non-parametric estimation of distance between groups," *Journal of Applied Statistics*, vol. 30, no. 7, pp. 743–750, 2003.

[12] A. F. Mitchell and W. J. Krzanowski, "The mahalanobis distance and elliptic distributions," *Biometrika*, vol. 72, no. 2, pp. 464–467, 1985.

[13] H. Holgersson and P. S. Karlsson, "Three estimators of the mahalanobis distance in high-dimensional data," *Journal of Applied Statistics*, vol. 39, no. 12, pp. 2713–2720, 2012.

[14] P. C. Mahalanobis, "On the generalized distance in statistics," in *Proceedings of the national institute of sciences of India*, vol. 2, pp. 49–55, New Delhi, 1936.

[15] I. Jolliffe, *Principal component analysis.* Wiley Online Library, 2005.

[16] I. T. Jolliffe, "Rotation of principal components: choice of normalization constraints," *Journal of Applied Statistics*, vol. 22, no. 1, pp. 29–35, 1995.

[17] P. Pack, I. Jolliffe, and B. Morgan, "Influential observations in principal com-

ponent analysis: A case study," *Journal of Applied Statistics*, vol. 15, no. 1, pp. 39–52, 1988.

[18] E. L. Lehmann and G. Casella, *Theory of point estimation*, vol. 31. Springer, 1998.

[19] A. Van den Bos, *Parameter estimation for scientists and engineers*. Wiley-Interscience, 2007.

[20] R. Cheng and N. Amin, "Maximum likelihood estimation of parameters in the inverse gaussian distribution, with unknown origin," *Technometrics*, vol. 23, no. 3, pp. 257–263, 1981.

[21] F. Nielsen and R. Nock, "Clustering multivariate normal distributions," in *Emerging Trends in Visual Computing*, pp. 164–174, Springer, 2009.

[22] W. D. Penny, "Kullback-liebler divergences of normal, gamma, dirichlet and wishart densities," *Wellcome Department of Cognitive Neurology*, 2001.

[23] R. D. Grigorieff, "A note on von neumanns trace inequality," *Math. Nachr*, vol. 151, pp. 327–328, 1991.

[24] L. Mirsky, "A trace inequality of john von neumann," *Monatshefte für Mathematik*, vol. 79, no. 4, pp. 303–306, 1975.