Maximum Likelihood Estimation and Optimal Coordinates

P. Spurek(✉) and J. Tabor

Faculty of Mathematics and Computer Science,
Jagiellonian University, Łojasiewicza 6, 30-348 Kraków, Poland
przemyslaw.spurek@ii.uj.edu.pl

Abstract. We show that the MLE (maximum likelihood estimation) in the class of Gaussian densities can be understood as the search for the best coordinate system which "optimally" underlines the internal structure of the data. This allows in particular to the search for the optimal coordinate system when the origin is fixed in a given point.

Keywords: Maximum likelihood estimation · Cross-entropy · Gaussian distribution

1 Introduction

MLE (maximum likelihood estimation) is one the most important estimation methods in statistics [4,11]. In data engineering it plays the crucial role in particular in EM clustering [15], in information theory it can be "identified" with the cross-entropy, which jointly with the Kullback-Leibler divergence plays the basic role in computer science [6]. In this paper we discuss the MLE in the case when the considered density is Gaussian with the center belonging to a given set. We were inspired by the ideas presented by [5] and consider estimations in various subclasses of normal densities.

One of the crucial question in data analysis is how to choose the best coordinate system and define distance which "optimally" underlines the internal structure of the data [3,8,12,17,18,20]. A similar role is played by Mahalanobis distance in discrimination analysis [9]. In general, we first need to decide if we *allow or not the translation of the origin of coordinate system.* Next we usually consider one of the following:

– *no change in coordinates;*
– *possibly different change of scale separately in each coordinate;*
– *arbitrary coordinates.*

P. Spurek—The paper was supported by the National Centre of Science (Poland) Grant No. 2013/09/N/ST6/01178.
J. Tabor— The paper was supported by the National Centre of Science (Poland) Grant No. 2014/13/B/ST6/01792.

J. Świątek and J.M. Tomczak (eds.), *Advances in Systems Science*, Advances in Intelligent Systems and Computing 539, DOI 10.1007/978-3-319-48944-5_1

It occurs that the value of likelihood function, in the case when we restrict to the Gaussian densities, can be naturally interpreted as the measure of the fitness of the given coordinate system to the data. Thus in the paper we search for those coordinates in the above situations which best describe (with respect to MLE) the given dataset $\mathcal{Y} \subset \mathbb{R}^N$.

At the end of the introduction let us mention that our results can be also used in various density estimation and clustering problems which use Gaussian models [1,5], in particular in the case when we consider the model consisting of Gaussians with centers satisfying certain constraints.

2 Entropy and Gaussian Random Variables

Let X be a random variable with density f_X. The differential entropy

$$H(X) := \int -\ln(f_X(y)) f_X(y) dy \tag{1}$$

tells us what is the asymptotic expected amount of memory needed to code X [6], and thus the differential code-length optimized for X is given by $-\ln(f_X(x))$.

If we want to code Y (a continuous variable with density g_Y) with the code optimized for X we obtain the *cross-entropy* which was presented in [6,10] (we follow the notation from [16]):

$$H^{\times}(Y\|X) := \int g_Y(y) \cdot (-\ln f_X(y)) dy, \tag{2}$$

If A is a linear operator, then $H^{\times}(AY\|AX) = H^{\times}(Y\|X) + \ln|\det(A)|$. Since we consider X only from its density f_X point of view, we will commonly use the notation

$$H^{\times}(Y\|f_X) := \int g_Y(y) \cdot (-\ln f_X(y)) dy. \tag{3}$$

Roughly speaking, $H^{\times}(Y\|f)$ denotes (asymptotically) the memory needed to code random variable Y with the code optimized for the density f. In the case of given dataset $\mathcal{Y} \subset \mathbb{R}^N$ we interpret $\mathcal{Y}$ as an uniform discreet variable Y on $\mathcal{Y}$. Consequently, our formula is reduced to

$$H^{\times}(\mathcal{Y}\|f) := H^{\times}(Y\|f) = -\frac{1}{|\mathcal{Y}|} \sum_{x \in \mathcal{Y}} \ln(f(x)), \tag{4}$$

where $|\mathcal{Y}|$ denote cardinality of the set $\mathcal{Y}$.

In our investigations we are interested in (best) coding for Y by densities chosen from a set of densities $\mathcal{F}$, and thus we will need the following definition.

Definition 1. *By the* cross-entropy of Y *with respect to a family of coding densities* $\mathcal{F}$ *we understand*

$$H^{\times}(Y\|\mathcal{F}) := \inf_{f \in \mathcal{F}} H^{\times}(Y\|f). \tag{5}$$

Observe that the search for the density f with minimal cross entropy leads exactly to the maximum likelihood estimation. Thus in general the calculation of $H^{\times}(Y\|\mathcal{F})$ is nontrivial, as it is equivalent to finding ML estimator.

As is the case in many statistical or data-information problems, the basic role in our investigations is played by the Gaussian densities. We recall that the normal variable with mean m and a covariance matrix Σ has the density $\mathcal{N}_{(\mathrm{m},\Sigma)}(x) = \frac{1}{\sqrt{(2\pi)^N \det(\Sigma)}} e^{(-\frac{1}{2}\|x-\mathrm{m}\|^2_\Sigma)}$, where $\|x-\mathrm{m}\|_\Sigma$ is the Mahalanobis norm $\|x-\mathrm{m}\|^2_\Sigma := (x-\mathrm{m})^T\Sigma^{-1}(x-\mathrm{m})$, see [13]. The differential entropy of Gaussian distribution is given by

$$H(\mathcal{N}_{(\mathrm{m},\Sigma)}) = \frac{N}{2}\ln(2\pi e) + \frac{1}{2}\ln(\det(\Sigma)).$$

From now on, if not otherwise specified, we assume that all the considered random variables have finite second moments and that they have values in $\mathbb{R}^N$. For a random variable Y, by $\mathrm{m}_Y = E(Y)$ we denote its mean, and by Σ_Y its covariance matrix, that is $\Sigma_Y = E((Y-\mathrm{m}_Y)\cdot(Y-\mathrm{m}_Y)^T)$.

We will need the following result, which says that the cross-entropy of an arbitrary random variable Y versus normal can be computed just from the knowledge of covariance and mean of Y.

Theorem 1 *([4], Theorem 5.59). Let Y be a random variable with finite covariance matrix. Then for arbitrary* m *and positive-definite covariance matrix Σ we have*

$$H^{\times}(Y\|\mathcal{N}_{(\mathrm{m},\Sigma)}) = \tfrac{N}{2}\ln(2\pi) + \tfrac{1}{2}\|\mathrm{m}-\mathrm{m}_Y\|^2_\Sigma + \tfrac{1}{2}\mathrm{tr}(\Sigma^{-1}\Sigma_Y) + \tfrac{1}{2}\ln(\det(\Sigma)). \tag{6}$$

Remark 1. *Suppose that we are given a data set $\mathcal{Y}$. Then we usually understand the data as a sample realization of a random variable Y. Consequently as an estimator for the mean of Y we use the mean $\mathrm{m}_{\mathcal{Y}} = \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}} y$ of the data $\mathcal{Y}$ and as the covariance we use the ML estimator $\frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}(y-\mathrm{m}_{\mathcal{Y}})(y-\mathrm{m}_{\mathcal{Y}})^T$.*

As a direct corollary we obtain the formula for the optimal choice of origin.

Proposition 1. *Let Y be a random variable and Σ be a fixed covariance matrix. Let M be a nonempty closed subset of $\mathbb{R}^N$. From all normal coding densities $\mathcal{N}_{(\mathrm{m},\Sigma)}$, where $\mathrm{m}\in M$, the minimal cross-entropy is realized for that $\mathrm{m}\in M$ which minimizes $\|\mathrm{m}-\mathrm{m}_Y\|_\Sigma$, and equals*

$$\inf_{\mathrm{m}\in M} H^{\times}(Y\|\mathcal{N}_{(\mathrm{m},\Sigma)}) = \tfrac{1}{2}\big(d^2_\Sigma(\mathrm{m}_Y;M) + \mathrm{tr}(\Sigma^{-1}\Sigma_Y) + \ln(\det(\Sigma)) + N\ln(2\pi)\big),$$

where d_Σ is a Mahalanobis distance.

Consequently, if $M=\mathbb{R}^N$ the minimum is realized for $\mathrm{m}=\mathrm{m}_Y$ and equals

$$\inf_{\mathrm{m}\in\mathbb{R}^N} H^{\times}(Y\|\mathcal{N}_{(\mathrm{m},\Sigma)}) = \tfrac{1}{2}\left(\mathrm{tr}(\Sigma^{-1}\Sigma_Y) + \ln(\det(\Sigma)) + N\ln(2\pi)\right).$$

It occurs that our basic MLE problem, in the case when we restrict to the Gaussian densities, can be naturally interpreted as search for the optimal rescaling (optimal choice of coordinate system).

Remark 2. *Let us start from one dimensional space. In such a case, if we allow the translation of the origin of coordinate system, we usually apply the standarization/normalization:* $s : Y \to (Y - \mathrm{m}_Y)/\sigma_Y$. *In the multivariate case the normalization is given by the transformation* $s : X \to \Sigma_Y^{-1/2}(Y - \mathrm{m}_Y)$. *Then we obtain that the coordinates are uncorrelated, and the covariance matrix is identity. Taking the distance between the transformation of points* x, y*:*

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

we arrive naturally at the Mahalanobis distance $\|x-y\|_\Sigma^2 = (x-y)^T \Sigma^{-1}(x-y)$. *If we do not allow the translation of the origin, we usually only scale each coordinate by dividing it by its mean (then the unit-scale plays the normalizing role, as the mean of each coordinate is one), arriving in the case when the mean is one.*

To study the question what is the optimal procedure, we need the criterion to compare different coordinate systems. Suppose that we are given a basis $\mathrm{v} = (v_1, \ldots, v_N)$ of $\mathbb{R}^N$ and an origin of coordinate system m. Then by $\mathcal{N}_{[\mathrm{m},\mathrm{v}]}$ we denote the "normalized" Gaussian density with respect to the basis v with center at m, that is

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

Then as a measure of fitness of the coordinate system $[\mathrm{m}, \mathrm{v}]$ we understand the cross-entropy $H^\times(Y \| \mathcal{N}_{[\mathrm{m},\mathrm{v}]})$.

3 Rescaling

Let us first consider the question how we should uniformly rescale the classical coordinates to optimally "fit" the data. *Assume that we have fixed an origin of the coordinate system* at 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 $s \to H^\times(Y \| \mathcal{N}_{(\mathrm{m},s\mathrm{I})})$ attains minimum. Since

$$H^\times(Y \| \mathcal{N}_{(\mathrm{m},s\mathrm{I})}) = \tfrac{1}{2}([\mathrm{tr}(\Sigma_Y) + \|\mathrm{m} - \mathrm{m}_Y\|^2]s^{-1} + N\ln(s) + N\ln(2\pi)), \quad (7)$$

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

$$\frac{N}{2}(\ln[\mathrm{tr}(\Sigma_Y) + \|\mathrm{m} - \mathrm{m}_Y\|^2] + \ln(2\pi e/N))$$

for $s = [\mathrm{tr}(\Sigma_Y) + \|\mathrm{m} - \mathrm{m}_Y\|^2]/N$. Thus we have arrived at the following theorem.

Theorem 2. *Let* Y *be a random variable with invertible covariance matrix and* m *be fixed. Then the minimum of* $H^\times(Y \| \{\mathcal{N}_{(\mathrm{m},s\mathrm{I})}\}_{s>0})$ *is realized for* $s = (\mathrm{tr}(\Sigma_Y) + \|\mathrm{m} - \mathrm{m}_Y\|^2)/N$, *and equals*

$$H^\times(Y \| \{\mathcal{N}_{(\mathrm{m},s\mathrm{I})}\}_{s>0}) = \tfrac{N}{2}(\ln[\mathrm{tr}(\Sigma_Y) + \|\mathrm{m} - \mathrm{m}_Y\|^2] + \ln(2\pi e/N)).$$

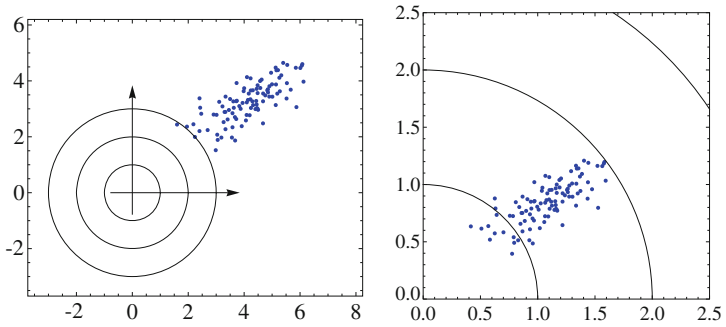

Fig. 1. The original data set with optimal coordinate system (the new "optimal" basis is marked by the bold arrows) in the case of the family $\{\mathcal{N}_{(\mathrm{m},s\mathrm{I})}\}_{s>0}$ (left figure). The data in the new basis (right figure).

Example 1. *Let $\mathcal{Y}$ be a realization of the normal random variable Y with $\mathrm{m}_Y = [3,4]^T$ and $\Sigma = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 0.6 \end{bmatrix}$ and let $\mathrm{m} = [0,0]^T$. In Fig. 1 we present a sample $\mathcal{Y}$ with the coordinate system (marked by bold black segments) obtained by the Theorem 2 and data in the new basis.*

Observe that the above minimum depends only on the trace of the covariance matrix of Y and the Euclidean distance of m from m_Y. If we allow the change of the origin, we have to clearly put the origin it at m_Y:

Corollary 1. *Let Y be a random variable with invertible covariance matrix. Then $H^{\times}(Y\|\{\mathcal{N}_{(\mathrm{m},s\mathrm{I})}\}_{s>0,\mathrm{m}\in\mathbb{R}^N})$ is realized for $\mathrm{m} = \mathrm{m}_Y$, $s = \frac{1}{N}\mathrm{tr}(\Sigma_Y)$, and equals*

$$H^{\times}(Y\|\{\mathcal{N}_{(\mathrm{m},s\mathrm{I})}\}_{s>0,\mathrm{m}\in\mathbb{R}^N}) = \tfrac{N}{2}(\ln(\mathrm{tr}(\Sigma_Y)) + \ln(\tfrac{2\pi e}{N})).$$

Corollary 2. *Let $\mathcal{Y} = (y_1,\ldots,y_n)$ be a given data-set. Assume that we want to move the origin to* m*, and uniformly rescale the coordinates. Then*

$$s \to (s-\mathrm{m})/\sqrt{\tfrac{1}{N}(\mathrm{tr}(\Sigma_{\mathcal{Y}}) + \|\mathrm{m}-\mathrm{m}_Y\|^2)}$$

is the optimal rescaling, where $\Sigma_{\mathcal{Y}}$ is a covariance of $\mathcal{Y}$. If we additionally allow the change of the origin, we should put $\mathrm{m} = \mathrm{m}_S$ and consequently the rescaling takes the form $s \to (s-\mathrm{m}_S)/\sqrt{\mathrm{tr}(\Sigma_{\mathcal{Y}})/N}$.

Applying the above we obtain that in the one dimensional case the rescaling takes the form $s \to (s-\mathrm{m}_{\mathcal{Y}})/\sigma_{\mathcal{Y}}$ (if we allow change of origin), and $s \to s/\sqrt{\mathrm{m}_{\mathcal{Y}}^2+\sigma_{\mathcal{Y}}^2}$ (if we fix the origin at zero).

Example 2. *Let $\mathcal{Y}$ be a realization of the normal random variable Y from Example 1. In Fig. 2 we present a sample $\mathcal{Y}$ with the coordinate system obtained by the Corollary 1 and data in the new basis.*

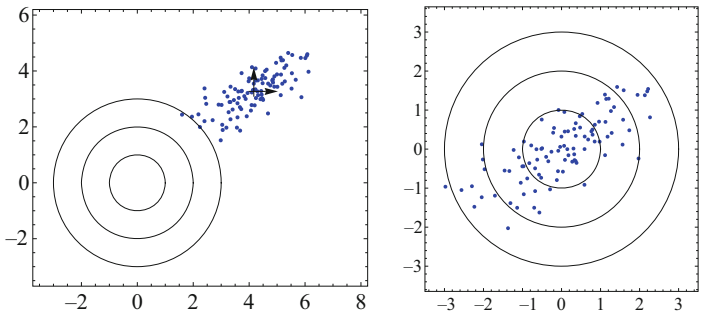

Fig. 2. The original data set with optimal coordinate system in the case of the family $\{\mathcal{N}_{(\mathrm{m},s\mathrm{I})}\}_{s>0,\mathrm{m}\in\mathbb{R}^N}$ (left figure). The data in the new basis (figure on the right).

Now we consider the case when we allow to rescale each coordinate Y_i of $Y = (Y_1, \ldots, Y_N)$ separately. For simplicity we consider the case $N = 2$. Consider the splitting $\mathbb{R}^N = \mathbb{R}^{N_1} \times \mathbb{R}^{N_2}$. For densities f_1 and f_2 on $\mathbb{R}^{N_1}$ and $\mathbb{R}^{N_2}$, respectively, we define the product density $f_1 \otimes f_2$ on $\mathbb{R}^N = \mathbb{R}^{N_1} \times \mathbb{R}^{N_2}$ by the formula

$$(f_1 \otimes f_2)(x_1, x_2) := f_1(x_1) \cdot f_2(x_2),$$

for $(x_1, x_2) \in \mathbb{R}^{N_1} \times \mathbb{R}^{N_2}$. Given density families $\mathcal{F}_1$ and $\mathcal{F}_2$, we put $\mathcal{F}_1 \otimes \mathcal{F}_2 := \{f_1 \otimes f_2 : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$. Let $Y : (\Omega, \mu) \to \mathbb{R}^{N_1} \times \mathbb{R}^{N_2}$ be a random variable and let $Y_1 : \Omega \to \mathbb{R}^{N_1}$ and $Y_2 : \Omega \to \mathbb{R}^{N_2}$ denote the first and second coordinate of Y (observe that in general Y_1 and Y_2 are not independent random variables). On can easily observe that

Proposition 2. *Let $\mathcal{F}_1$ and $\mathcal{F}_2$ denote coding density families in $\mathbb{R}^{N_1}$ and $\mathbb{R}^{N_2}$, respectively, and let $Y : \Omega \to \mathbb{R}^{N_1} \times \mathbb{R}^{N_2}$ be a random variable. Then*

$$H^{\times}(Y \| \mathcal{F}_1 \otimes \mathcal{F}_2) = H^{\times}(Y_1 \| \mathcal{F}_1) + H^{\times}(Y_2 \| \mathcal{F}_2).$$

The above result means that if we allow to rescale coordinates, we can treat them as separate random variables. Thus we obtain the following theorem.

Theorem 3. *Let $\mathcal{Y}$ be a data set, and let $\mathcal{Y}_k$ denote the set containing its k-th coordinate. Then the optimal rescaling for each k-th coordinate is given by*

$$\mathcal{Y}_k \ni s \to (s - \mathrm{m}_{\mathcal{Y}_k})/\sigma_{\mathcal{Y}_k} \ (if\ we\ allow\ change\ of\ origin),$$
$$\mathcal{Y}_k \ni s \to s/\sqrt{\mathrm{m}_{\mathcal{Y}}^2 + \sigma_{\mathcal{Y}}^2} \ (if\ we\ fix\ the\ origin\ at\ zero).$$

Example 3. *Let $\mathcal{Y}$ be a realization of the normal random variable Y from Example 1. In Fig. 4 we present a sample $\mathcal{Y}$ and the coordinate system obtained by the Theorem 3 (if we fix the origin at zero) and data in the new basis. In Fig. 3 we present a sample $\mathcal{Y}$ and the coordinate system obtained by the Theorem 3 (when we allow change of origin) and data in the new basis.*

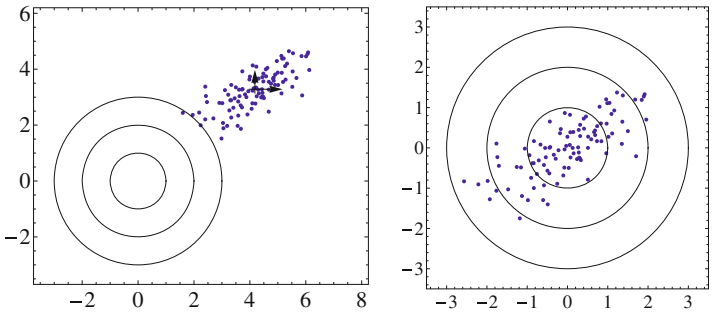

Fig. 3. The original data set with optimal coordinate system in the case of separated random variable when we allow change of origin (figure on the left) and the data in the new basis (figure on the right).

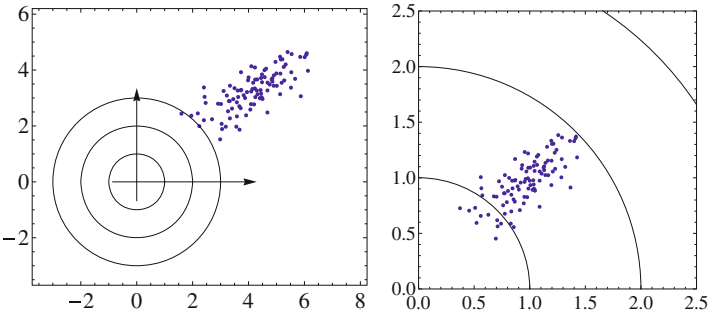

Fig. 4. The original data set with optimal coordinate system in the case of separated random variable when we do not allow change of origin (left hand side illustration) and data in the new basis (right hand side illustration).

4 Main Result

We find the optimal coordinate system in the general case by applying an approach similar to that from [19]. To do so, we need a simple consequence of the famous von Neuman trace inequality. Next we discuss the optimal rescaling if we move the coordinate to the mean of the data.

In most of our further results the following proposition will play an important role. In its proof we will use the well-known von Neumann trace inequality described by [7,14]:

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.

We also need Sherman-Morrison formula [2]:

Theorem [Sherman-Morrison formula]. *Suppose A is an invertible square matrix and u, v are column vectors. Suppose furthermore that $1 + v^T A^{-1} u \neq 0$. Then the Sherman-Morrison formula states that*

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1} u v^T A^{-1}}{1 + v^T A^{-1} u}.$$

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 3. *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*

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

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

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

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 $E = \lambda_N \mathrm{I} - A$, $F = B$. 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)

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

Since

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

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

Now we proceed to the main result of the paper. Let $M \subset \mathbb{R}^N$ then by $\mathcal{G}_M$ we denote the set of Gaussians with mean $\mathrm{m} \in M$.

Theorem 4. *Let $\mathrm{m} \in \mathbb{R}^N$ be fixed and let $\mathcal{G}_{\{\mathrm{m}\}}$ denote the set of Gaussians with mean m. Then $H^{\times}(Y \| \mathcal{G}_{\{\mathrm{m}\}})$ equals*

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

and is attained for $\Sigma = \Sigma_Y + (\mathrm{m} - \mathrm{m}_Y)(\mathrm{m} - \mathrm{m}_Y)^T$.

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

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

Then A is a symmetric positive matrix. Contrary 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

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

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

$$v^T Av = \mathrm{tr}(v^T Av) = \mathrm{tr}(A \cdot (vv^T)).$$

We fix an orthonormal basis 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

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

and then (12) equals $N + \ln(1 + \|\mathrm{m} - \mathrm{m}_Y\|^2_{\Sigma_Y})$, while the formula for A minimizing it is given by $A = \mathrm{I} - \frac{vv^T}{1+\|v\|^2}$. Consequently then the minimal value of (10) is

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

and by (11) and Sherman-Morrison formula is attained for

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

Example 4. *Let $\mathcal{Y}$ be a realization of the normal random variable Y from Example 1 and let* m *be fixed. In Fig. 5 is presented a sample $\mathcal{Y}$ and the coordinate system obtained by the Theorem 4 and data in the new basis.*

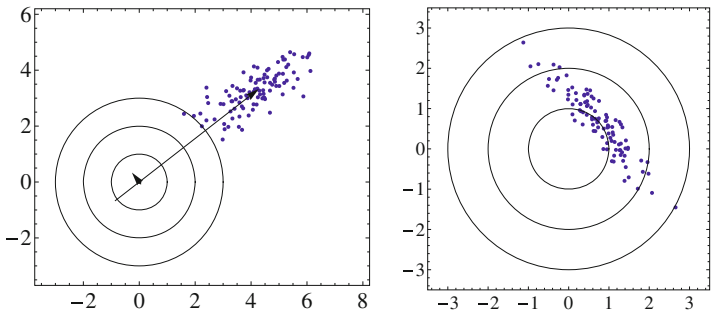

Fig. 5. The original data set with optimal coordinate system in the case of the family $\mathcal{G}_{\{(0,0)\}}$ (left hand side illustration) and data in the new basis (right hand side illustration).

5 Conclusion

In the paper we show that the MLE in the class of Gaussian densities can be understood equivalently as the search for the coordinates which best describe given dataset $\mathcal{Y} \subset \mathbb{R}^N$. The main result of the paper presents the formula of the optimal coordinate system in the case when the mean of the Gaussian density satisfies certain constrains.

Our work can be used in density estimation and clustering algorithms which use different Gaussian models.

References

1. Banfield, J.D., Raftery, A.E.: Model-based gaussian and non-gaussian clustering. Biometrics **49**(3), 803–821 (1993)
2. Bartlett, M.S.: An inverse matrix adjustment arising in discriminant analysis. Ann. Math. Stat. **22**(1), 107–111 (1951)
3. Borg, I., Groenen, P.: Modern multidimensional scaling: Theory and applications. Springer, Heidelberg (2005)
4. Van den Bos, A.: Parameter Estimation for Scientists and Engineers. Wiley Online Library, New York (2007)
5. Celeux, G., Govaert, G.: Gaussian parsimonious clustering models. Pattern Recogn. **28**(5), 781–793 (1995)
6. Cover, T., Thomas, J., Wiley, J., et al.: Elements of Information Theory. Wiley Online Library, New York (1991)
7. Grigorieff, R.: A note on von neumanns trace inequality. Math. Nachr. **151**, 327–328 (1991)
8. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco (2011)
9. Krishnaiah, P.: Handbook of Statistics. North-Holland, New York (1988)
10. Kullback, S.: Information Theory and Statistics. Dover Publications, New York (1997)

11. Lehmann, E., Casella, G.: Theory of Point Estimation. Springer, New York (1998)
12. De Maesschalck, R., Jouan-Rimbaud, D., Massart, D.: The mahalanobis distance. Chemometr. Intell. Lab. Syst. **50**(1), 1–18 (2000)
13. Mahalanobis, P.C.: On the generalised distance in statistics. Proc. Nat. Inst. Sci. **2**, 49–55 (1936)
14. Mirsky, L.: A trace inequality of john von neumann. Monatshefte für mathematik **79**(4), 303–306 (1975)
15. Ng, S., Krishnan, T., McLachlan, G.: The em algorithm. In: Gentle, J.E., Härdle, W.K., Mori, Y. (eds.) Handbook of Computational Statistics Concepts and Methods. Springer Handbooks of Computational Statistics, pp. 139–172. Springer, Heidelberg (2004)
16. Nielsen, F., Nock, R.: Sided and symmetrized bregman centroids. IEEE Trans. Inf. Theory **55**(6), 2882–2904 (2009)
17. Raykov, T., Marcoulides, G.: An Introduction to Applied Multivariate Analysis. Routledge, London (2008)
18. Rencher, A.: Methods of Multivariate Analysis. Wiley Online Library, New York (1995)
19. Theobald, C.: An inequality with application to multivariate analysis. Biometrika **62**(2), 461–466 (1975)
20. Timm, N.: Applied Multivariate Analysis. Springer, New York (2002)