Improving the Accuracy of Principal Component Analysis by the Maximum Entropy Method

Guihong Wan
Department of Computer Science
The University of Texas at Dallas
Richardson, Texas 75083
Guihong.Wan@utdallas.edu

Crystal Maung
7 Next
7-Eleven Inc.
Irving, Texas 75063
Crystal.Maung@7-11.com

Haim Schweitzer
Department of Computer Science
The University of Texas at Dallas
Richardson, Texas 75083
HSchweitzer@utdallas.edu

***Abstract*—Classical Principal Component Analysis (PCA) approximates data in terms of projections on a small number of orthogonal vectors. There are simple procedures to efficiently compute various functions of the data from the PCA approximation. The most important function is arguably the Euclidean distance between data items. This can be used, for example, to solve the approximate nearest neighbor problem. We use random variables to model the inherent uncertainty in such approximations, and apply the Maximum Entropy Method to infer the underlying probability distribution. We propose using the expected values of distances between these random variables as improved estimates of the distance. We show experimentally that in most cases results obtained by our method are more accurate than what is obtained by the classical approach. This improves the accuracy of a classical technique that have been used with little change for over 100 years.**

***Index Terms*—Principal Component Analysis (PCA), Dimension Reduction, Low Rank Matrix Representation, Maximum Entropy Method, Euclidean distance, Rayleigh Quotient**

I. Introduction

We consider the standard representation of numerical data as a large matrix of numeric values. Let n be the number of data items in the dataset, and let m be the size of each item. The data can be viewed as a matrix of size $m \times n$, as illustrated in Fig. 1. In many practical situations both m and n are very large. For example, datasets containing genome data may have m in the thousands and n in the millions [1]. In such cases even simple tasks, such as searching the data for a particular item become computationally expensive.

A standard approach to address this "curse of dimensionality" is dimension reduction, reducing the dimension of each data item from m to k, where $k < m$. For a review of dimension reduction techniques see, e.g., [2]. The most common approach is the Principal Component Analysis (PCA), known for over 100 years. For references see, for example [3]–[6]. The uncentered variant can be described as follows. Let A be the data matrix of size $m \times n$; define the $m \times m$ matrix B by: $B = AA^T$. Let V be an $m \times k$ matrix whose columns are the k eigenvectors of B corresponding to the k largest eigenvalues. The columns of V are orthogonal, and their span gives the best possible approximation of rank k to the column space of A. Let a_i be the ith column of A. The following approximations hold:

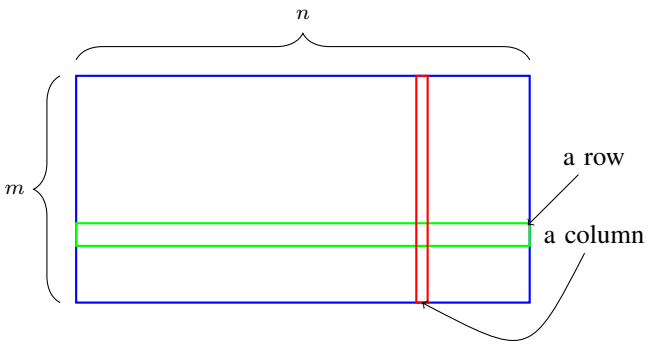

Fig. 1: The view of data as a matrix. There are n data items, and each one is of size m. A data item is a column of an $m \times n$ matrix.

$$A \approx VW, \quad a_i \approx V w_i \tag{1}$$

Here W is $k \times n$, representing A in the reduced dimension. In particular, the ith column of A is the vector a_i, and it is represented by w_i, the ith column of W. The matrix W or any specific column w_i can be computed by:

$$W = V^T A, \quad w_i = V^T a_i \tag{2}$$

The centered variant of the PCA is the same as the uncentered PCA with an initial centering of each column. The centering is performed by mean subtraction. See, e.g., [7]. The PCA enables fast computations of many data related functions. The low dimension also helps with the visualization and the interpretation of the data.

We proceed to describe how to use the representation in (1) to approximate the Euclidean distance between data items. Recall that the squared Euclidean distance between two vectors x and y is given by:

$$\text{distance}^2(x, y) = \|x - y\|^2$$

It is computed as the sum of m squared coordinates. Thus, the cost of computing this distance is $O(m)$. Now suppose that

both x and y are from the same dataset with a known PCA, as given by Equation (1). For clarity we take $x = a_i$ and $y = a_j$. Then because of the orthogonality of V we have:

$$\begin{aligned}\text{distance}^2(a_i, a_j) &= \|a_i - a_j\|^2 \\ &\approx \|V(w_i - w_j)\|^2 = \|w_i - w_j\|^2 \quad (3)\end{aligned}$$

This is a classical approximation formula, known for over 100 years. See, e.g., [2], [4]. It shows that the approximate value of $\|a_i - a_j\|^2$ can be computed in $O(k)$ instead of the exact computation which takes $O(m)$.

Another common situation where the PCA leads to significant improvements in the running time is the following. Suppose the vector x is not necessarily a column of A, and one has to calculate the n squared distances $d_i^2 = \|x - a_i\|^2$ for $i = 1, \ldots, n$. (These are the distances between x and all the columns of A.) This situation occurs, for example, in calculating the nearest neighbor of x among the columns of A (e.g., [8]), or in the computation of multi-dimensional scaling (e.g., [9]). The direct approach requires computing n distances which takes $O(mn)$. If the PCA of A is known, the approximate n distances can be computed in $O(km + kn)$ by the following algorithm:

$$w_x = V^T x, \quad d_j^2 \approx \|w_x - w_j\|^2 \ \text{ for } j = 1, \ldots n \quad (4)$$

Our contributions

Our main result is formulas that improve the quality of the approximations in (3) and (4). Clearly, the approximation in (3), and sometimes also the approximation in (4), can be improved by increasing k, the rank of the reduced dimension. But this increases the computation cost and reduces the effectiveness of working in a reduced dimension. It also makes the interpretation of the data in the low dimension harder. For example, with $k = 2$ the data can be visualized in a plane. Increasing k to 4 creates a representation that is much harder to visualize.

Our main result is new formulas that improve the accuracy in (3) and (4) without increasing k. Specifically, we propose the approximation formula (5) as an alternative to (3), and the approximation formula (6) as an alternative to (4):

$$\begin{aligned}&\text{distance}^2(a_i, a_j) \approx \\ &\quad \|w_i - w_j\|^2 + \|a_i\|^2 - \|w_i\|^2 + \|a_j\|^2 - \|w_j\|^2 \quad (5)\end{aligned}$$

$$\begin{aligned}&\text{distance}^2(x, a_j) \approx \\ &\quad \|w_x - w_j\|^2 + \|x\|^2 - \|w_x\|^2 + \|a_j\|^2 - \|w_j\|^2 \quad (6) \\ &\text{where} \ \ w_x = V^T x\end{aligned}$$

The following are some observations about the result:

- The formulas (5) and (6) use additional information: the squared norms $\|a_i\|^2$ for each column a_i of A. This information can be pre-computed during the PCA calculation, without significant change to the running time of the PCA.
- The complexity of using (5) to compute the approximate squared distance between a_i and a_j is $O(k)$, the same as the complexity of using (3) to approximate the squared distance between a_i and a_j.
- The complexity of using (6) to compute the approximate squared distances between x and all the columns of A is $O(km + kn)$, the same as the complexity of using (4) to compute the approximate squared distances between x and all the columns of A.
- The new approximations (5) and (6) are not always better than the old approximations (3) and (4). But we claim that "on the average" the new approximations are better. This follows from the derivation of these approximations using the Maximum Entropy Method and was verified by extensive evaluation on real datasets.

The technique that we use to derive the formulas (5) and (6) can also be used for other applications of the PCA besides Euclidean distances. We derive related formulas for accurate computation of the Rayleigh Quotient, an important statistic that indicates the similarity of a vector to a collection of vectors.

Another important contribution of the paper is the method in which the approximations are derived. We model the uncertainty in the PCA representation in terms of random variables with an unknown distribution. We then use the Maximum Entropy Method to determine the most likely distribution. Expected values are then used as the improved estimates. This approach appears to be novel. We are not aware of any previous studies that apply similar approaches to improve deterministic estimates.

Paper organization

The paper is organized as follows. Section II formulates dimension reduction as an approximate estimation of column vectors with unknown quantities. A key idea is to model the unknown quantities as random variables in an unknown probability distribution.

Section III describes the Maximum Entropy Method, a classical method of inferring the most likely probability distribution from partial information about random variables. We apply the Maximum Entropy Method to the random variables of Section II to derive the most likely probability distribution of the PCA estimates. A key theorem proved in this section characterizes the probability density of the unknown quantities.

In Section IV we use the probability density of Section III to compute expected values of several expressions of PCA approximations.

In Section V we apply the results of Section IV to compute estimates to distances between vectors. In Section VI we derive maximum entropy estimates to Rayleigh quotients.

Section VII describes extensive experimental results evaluating our approximation formulas on real data.

II. A PROBABILISTIC SETTING FOR PCA

As discussed in Section I the PCA approximation of the $m \times n$ matrix A is given by Equation (1). In this section we

use a slightly different notation for the same relation. We write the PCA approximation as:

$$A \approx V_1 W_1 \tag{7}$$

Since V_1 has orthogonal columns, it is always possible to extend these columns to an orthogonal basis of $\mathbb{R}^m$. Let V_2 be such an extension then V_1 and V_2 are orthogonal complements. They satisfy the following properties:

$$V_1^T V_1 = I, \quad V_2^T V_2 = I, \quad V_1 V_1^T + V_2 V_2^T = I \tag{8}$$

Using both V_1 and V_2 there is an exact representation of A that can be expressed as follows:

$$A = V_1 W_1 + V_2 W_2, \quad a_i = V_1 w_1^i + V_2 w_2^i \tag{9}$$

where a_i is the ith column of A, w_1^i is the ith column of W_1, and w_2^i is the ith column of W_2. Suppose the PCA of A is given as the matrices V_1 and W_1. Without loss of generality V_2 can be selected as any orthogonal complement of V_1. This means that the only unknown quantities in (9) are the entries of the matrix W_2. The special case of classical PCA is obtained by taking $W_2 = 0$. Instead, we propose to view W_2 as a random matrix, with entries that are random variables. From (9) it follows that if W_2 is a random matrix then A is also a random matrix, and so are the columns a_i and w_2^i. Equation (10) identifies random variables with $\widehat{\ }$ as shown below:

$$\widehat{A} = V_1 W_1 + V_2 \widehat{W_2}, \quad \widehat{a}_i = V_1 w_1^i + V_2 \widehat{w}_2^i \tag{10}$$

We note that some of the matrices in (10) are too big to manipulate explicitly. The size of V_2 is $m \times m-k$, and the size of W_2 is $m-k \times n$. A practical solution should not manipulate these matrices explicitly.

We proceed to show that modeling the unknown W_2 as a random matrix has an advantage over setting it to be 0. Suppose the probability density of W_2 is known. Applying the expectation operator $E\{\}$ to both sides of the first equation in (10) we get:

$$E\{\widehat{A}\} = V_1 W_1 + V_2 E\{\widehat{W_2}\}$$

Thus, Taking $E\{\widehat{A}\}$ as an improved estimate of A we can expect an improved result different from the classical result whenever $E\{\widehat{W_2}\}$ is nonzero. Similarly, using the orthogonality of V_1, V_2 it is easy derive the following relation from (10): $\widehat{A}^T \widehat{A} = W_1^T W_1 + \widehat{W_2}^T \widehat{W_2}$. Taking expectations we see that:

$$E\{\widehat{A}^T \widehat{A}\} = W_1^T W_1 + E\{\widehat{W_2}^T \widehat{W_2}\}$$

Therefore, the improved estimate of $A^T A$ is different from the classical estimate whenever $E\{\widehat{W_2}^T \widehat{W_2}\} \neq 0$. Observe that $E\{\widehat{W_2}^T \widehat{W_2}\} \neq E\{\widehat{W_2}\}^T E\{\widehat{W_2}\}$, so that $E\{\widehat{W_2}^T \widehat{W_2}\}$ may be nonzero even if $E\{\widehat{W_2}\}$ is 0.

In our case the probability density of W_2 is unknown. We use the Maximum Entropy Method to compute the most likely probability distribution under the assumption that the column norms of A are known. It is not surprising that under this probability density $E\{\widehat{W_2}\} = 0$, but we found it surprising that $E\{\widehat{W_2}^T \widehat{W_2}\} \neq 0$.

III. The Maximum Entropy Method

The Maximum Entropy Method is a standard technique for inferring probability distributions. When given constraints that the probability distribution must satisfy, the Maximum Entropy Method asserts that the "most likely distribution" is the distribution with the largest entropy that satisfies the constraints. See, e.g., [10]–[12]. In this paper we use the following special case described in Chapter 14 in Papoulis [11].

Theorem 1: Let $x = (x_1, \ldots, x_n)^T$ be a random vector, where the coordinates x_i are n random variables. Let $R = E\{xx^T\}$ be the correlation matrix associated with x. Suppose R is known. Let Δ be the determinant of R and assume $\Delta \neq 0$. Then according to the Maximum Entropy Method the probability density $f(x)$ and the entropy $H(x)$ are given by:

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{(2\pi)^n \Delta}} e^{-\frac{1}{2} x^t R^{-1} x} \\ H(x) &= \ln \sqrt{(2\pi e)^n \Delta} \end{aligned} \tag{11}$$

See [11] for the proof.

As stated in (11) the entropy of x is determined by Δ, the determinant of the correlation matrix R. If R is only partially known, it can be determined by the Maximum Entropy Method by maximizing the determinant Δ over the unknown quantities. We use this technique to derive the following theorem:

Theorem 2: Let $W = (w_1, \ldots, w_n)$ be a random matrix of dimensions $k \times n$. Suppose $z_i = E\{\|w_i\|^2\}$ is known for $i = 1, \ldots, n$, but nothing else is known about the probability density of W. Then according to the Maximum Entropy Method:

1. All entries of the matrix W have 0 mean:

$$E\{w_{ij}\} = 0, \quad \text{for } i = 1, \ldots, n, \ j = 1, \ldots, k$$

2. The random variable w_{i_1,j_1} is independent of the random variable w_{i_2,j_2} unless $i_1 = i_2$ and $j_1 = j_2$.

3. The expected value of $\|w_{ij}\|^2$ is given by:

$$E\{\|w_{ij}\|^2\} = \frac{z_i}{k}$$

4. The probability density of W is given by:

$$f(W) = \frac{1}{\sqrt{(2\pi)^{kn} \Delta}} e^{s(W)}$$

where:

$$\Delta = \frac{\prod_{i=1}^n z_i^k}{k^{kn}}, \quad s(W) = -\frac{k}{2} \sum_{i,j} \frac{w_{i,j}^2}{z_i}$$

Proof: Let q be the vector of kn random variables, created by concatenating all the columns of W:

$$q = \left(w_{11}, \ldots, w_{1k}, \ldots, w_{n1}, \ldots, w_{nk}\right)^T$$

Let R be the correlation matrix of q: $R = E\{qq^T\}$. Observe that the matrix R is $nk \times nk$. The value of R at row I and

column J is: $R_{IJ} = E\{w_{i_1,j_1} w_{i_2,j_2}\}$ for some i_1, j_1, i_2, j_2. Define: $\nu_{ij} = E\{w_{ij}^2\}$. Then from the definition of z_i:

$$z_i = E\{\|w_i\|^2\} = \sum_j E\{w_{ij}^2\} = \sum_j \nu_{ij} \tag{12}$$

The following information is known about the diagonal elements R_{II}, where the location I in q correspond to the location i, j in W:

$$R_{II} = E\{(w_{ij}\}^2) = \nu_{ij}.$$

From Theorem 1 it follows that the maximum entropy of q is obtained by maximizing Δ, the determinant of R, under the constraints (12). The proof of the theorem follows from this maximization.

According to the Hadamard determinant inequality (see, e.g., [13]), $\Delta \leq \prod_I R_{II}$. Since $\Delta = \prod_I R_{II}$ if all off diagonal elements are 0, it follows that there is a maximum where $E\{w_{i_1,j_1} w_{i_2,j_2}\} = 0$ unless $i_1{=}i_2$ and $j_1{=}j_2$. This proves parts 1 and 2 of the theorem.

To prove part 3 observe that from parts 1,2 it follows that

$$\Delta = \prod_I R_{II} = \prod_{i,j} \nu_{ij} = \prod_i (\prod_j \nu_{ij})$$

Therefore, for each i we need to maximize $\prod_{i,j} \nu_{ij}$ subject to the constraint that $\sum_{j=1}^k \nu_{ij} = z_i$. It can be easily shown (for example using the method of Lagrange multipliers) that the maximizing solution is $\nu_{ij} = \frac{z_i}{k}$. This proves part 3 of the theorem. Part 4 of the theorem follows from Theorem 1 by observing that:

- From part 3:

$$\Delta = \prod_{i=1}^n \prod_{j=1}^k (z_i/k) = \frac{\prod_{i=1}^n z_i^k}{k^{kn}}$$

- Since R is diagonal:

$$q^T R^{-1} q = \sum_I \frac{q_I^2}{R_{II}} = \sum_{i,j} \frac{w_{i,j}^2}{z_i/k}$$

This completes the proof of Theorem 2. ■

IV. Expected values of PCA approximations

In Equation (10) the matrix A is expressed as an expression involving random variables. We apply Theorem 2 to the matrix W_2 and determine its most likely probability density. The expected values of various estimates can then be computed in closed form. The matrix W_2 is of size $m{-}k \times n$, and its ith column, w_2^i, is of size $m{-}k$. The value of z_i in Theorem 2 can be computed as follows:

$$z_i = E\{\|\widehat{w}_2^i\|^2\} = \|a_i\|^2 - \|w_1^i\|^2, \quad i = 1, \ldots, n \tag{13}$$

Applying Theorem 2 this gives the following expected values of expressions related to w_2^i, the ith column of W_2:

$$\begin{aligned} &E\{\widehat{w}_2^i\} = 0, \quad E\{\|\widehat{w}_2^i\|^2\} = z_i \\ &E\{(\widehat{w}_2^i)^T \widehat{w}_2^j\} = 0, \quad E\{\widehat{w}_2^i (\widehat{w}_2^i)^T\} = \frac{z_i}{m-k} I \end{aligned} \tag{14}$$

where $i{\neq}j$, and I is the $m{-}k \times m{-}k$ identity matrix.

From (14) we get the following expected values related to the entire matrix W_2:

$$\begin{aligned} &E\{\widehat{W}_2\} = 0 \\ &E\{\widehat{W}_2^T \widehat{W}_2\} = \begin{pmatrix} z_1 & & & \\ & z_2 & & \\ & & \ddots & \\ & & & z_n \end{pmatrix} \\ &E\{\widehat{W}_2 \widehat{W}_2^T\} = \frac{\sum_{i=1}^n z_i}{m-k} I = \delta I \\ &\text{where } \delta = \sum_{i=1}^n z_i/(m-k) \end{aligned} \tag{15}$$

The corresponding formulas for $\widehat{A} = V_1 W_1 + V_2 \widehat{W}_2$ (as in (10)) are:

$$\begin{aligned} &E\{\widehat{A}\} = V_1 W_1 \\ &E\{\widehat{A}^T \widehat{A}\} = W_1^T W_1 + \text{Diag}(z_1, \ldots z_n) \\ &E\{\widehat{A}\widehat{A}^T\} = V_1 W_1 W_1^T V_1^T + \delta(I - V_1 V_1^T) \end{aligned} \tag{16}$$

The first equation follows from the first formula in (15). The second equation follows by applying expectations to the identity: $\widehat{A}^T \widehat{A} = W_1^T W_1 + \widehat{W}_2^T \widehat{W}_2$. The third equation follows by applying expectations to the identity:

$$\begin{aligned} \widehat{A}\widehat{A}^T =& V_1 W_1 W_1^T V_1^T + V_2 \widehat{W}_2 \widehat{W}_2^T V_2^T \\ &+ V_1 W_1 \widehat{W}_2^T V_2^T + V_2 \widehat{W}_2 W_1^T V_1^T \end{aligned}$$

Taking expectations the last two terms disappear. The final result is obtained by applying (15) to second expression, and using (8) to replace $V_2 V_2^T$ with $I - V_1 V_1^T$.

V. Computing distances with PCA

In this section we assume being given the matrix A with pre-computed PCA expressed as: $A \approx V_1 W_1$. In addition to the PCA we assume that the column norms $\|a_i\|$ are known for all the columns of A. Two cases are analyzed. In the first case the goal is to compute distances between columns of A. In the second case the goal is to compute distances between a vector x unrelated to A and columns of A. In each case we describe three formulas. The first formula that we denote by $d_{\text{classical}}$ is the classical formula. It does not use the additional information of column norms. The second formula that we denote by d_{ent} is obtained from the Maximum Entropy Method. It requires the additional information of column norms. Since d_{ent} works much better than $d_{\text{classical}}$ one may suspect that the reason might be the additional information of column norms. We use this information to derive another distance formula, as a tight lower bound to the true distance that also requires the additional information of column norms. We denote this third distance formula by d_{lower}. Our experimental results show that typically d_{lower} is better than $d_{\text{classical}}$, and d_{ent} is much better than d_{lower}.

A. *Distances between columns of* A

We consider approximating the distance between a_i and a_j, two columns of A. Their PCA representation is:

$$a_i \approx V_1 w_1^i, \quad a_j \approx V_1 w_1^j \tag{17}$$

As discussed in Section I the classical approximation formula for the squared distance between them is:

$$\text{distance}^2(a_i, a_j) \approx d_{\text{classical}}(a_i, a_j) = \|w_1^i - w_1^j\|^2 \tag{18}$$

Theorem 3: Let a_i and a_j be two columns of A with PCA representation as shown in (17). The estimate of the squared distance between them according to the Maximum Entropy Method is:

$$\text{distance}^2(a_i, a_j) \approx$$
$$d_{\text{classical}}(a_i, a_j) + \|a_i\|^2 - \|w_1^i\|^2 + \|a_j\|^2 - \|w_1^j\|^2$$

Proof: The random variable representation in (10) gives:

$$\widehat{a}_i = V_1 w_1^i + V_2 \widehat{w}_2^i, \quad \widehat{a}_j = V_1 w_1^j + V_2 \widehat{w}_2^j$$

Computing the squared Euclidean distance between them as a random variable gives:

$$\begin{aligned}
&\|\widehat{a}_i - \widehat{a}_j\|^2 = \|V_1(w_1^i - w_1^j) + V_2(\widehat{w}_2^i - \widehat{w}_2^j)\|^2 \\
&= \|w_1^i - w_1^j\|^2 + \|\widehat{w}_2^i - \widehat{w}_2^j\|^2 \\
&= \|w_1^i - w_1^j\|^2 + \|\widehat{w}_2^i\|^2 + \|\widehat{w}_2^j\|^2 - 2(\widehat{w}_2^i)^T \widehat{w}_2^j
\end{aligned}$$

Going to expectations and using Equation (14) we see that the expected value of the right most term is 0, and the values of the middle two terms are z_i, z_j. This gives:

$$d_{\text{ent}}(a_i, a_j) = d_{\text{classical}}(a_i, a_j) + z_i + z_j \tag{19}$$

The theorem now follows from (13). ■

Theorem 4: Define d_{lower} as follows:

$$d_{\text{lower}}(a_i, a_j) = d_{\text{classical}}(a_i, a_j) + z_i + z_j - 2\sqrt{z_i z_j}$$
$$\text{where } z_i = \|a_i\|^2 - \|w_1^i\|^2$$

Then:

$$d_{\text{classical}}(a_i, a_j) \le d_{\text{lower}}(a_i, a_j) \le \text{distance}^2(a_i, a_j)$$

Proof: The following relations hold:

$$a.\ \text{distance}^2(a_i, a_j) = \|w_1^i - w_1^j\|^2 + \|w_2^i - w_2^j\|^2$$
$$b.\ \|w_2^i - w_2^j\|^2 \ge (\|w_2^i\| - \|w_2^j\|)^2 = z_i + z_j - 2\sqrt{z_i z_j}$$
$$c.\ d_{\text{classical}} = \|w_1^i - w_1^j\|^2$$

Relation a follows from (9). Relation b follows from the triangle inequality. Relation c is the definition of $d_{\text{classical}}$. Combining relations b and c gives the left inequality in the theorem. Combining relations a and b gives the right inequality in the theorem. ■

This shows that d_{lower} is a lower bound on the true distance. The bound is tight since there is an equality in b when the angle between w_2^i and w_2^j is 0.

In summary, we describe 3 formulas for estimating distances between matrix columns using PCA data:

$$d_{\text{classical}}(a_i, a_j) = \|w_1^i - w_1^j\|^2$$
$$d_{\text{lower}}(a_i, a_j) = d_{\text{classical}}(a_i, a_j) + z_i + z_j - 2\sqrt{z_i z_j}$$
$$d_{\text{ent}}(a_i, a_j) = d_{\text{classical}}(a_i, a_j) + z_i + z_j$$

In these formulas w_1^i is the representation of column a_i in PCA space, and $z_i = \|a_i\|^2 - \|w_i\|^2$. Column norms are used by d_{ent} and d_{lower}. Both $d_{\text{classical}}$ and d_{lower} are lower bounds on the true distance, and d_{lower} is guaranteed to be better than $d_{\text{classical}}$. The promise of d_{ent} is that it was derived from the best probability distribution according to the Maximum Entropy Method. As we show in the experimental section its accuracy is significantly better than the accuracy of d_{lower} and $d_{\text{classical}}$.

B. *Distances between an arbitrary vector and columns of* A

Let x be an arbitrary (m dimensional) vector. Our goal is to approximate efficiently and accurately distances between x and the columns of A. As in Section V-A we assume the availability of the PCA of A, as well as the norms of A columns. We begin by defining the vectors w_1^x and w_2^x as analogous to w_1^i and w_2^i for a column of A:

$$w_1^x = V_1^T x, \quad w_2^x = V_2^T x \tag{20}$$

With this definition most of the analysis in Section V-A applies in this case as well. The only difference in the analysis is that w_2^x can be explicitly calculated, and therefore it is not a random variable. Still, the three distance formulas from Section V-A can be used in this case as well. As in (18) the classical error estimate is given below:

$$\text{distance}^2(x, a_j) \approx d_{\text{classical}}(x, a_j) = \|w_1^x - w_1^j\|^2 \tag{21}$$

This approximation can only be accurate when x projection on V_2 is small.

Theorem 5: Define d_{lower} as follows:

$$d_{\text{lower}}(x, a_j) = d_{\text{classical}}(x, a_j) + z_x + z_j - 2\sqrt{z_x z_j}$$
$$\text{where } z_x = \|x\|^2 - \|w_1^x\|^2,\ z_j = \|a_j\|^2 - \|w_1^j\|^2$$

Then:

$$d_{\text{classical}}(x, a_j) \le d_{\text{lower}}(x, a_j) \le \text{distance}^2(x, a_j)$$

Proof: Identical to the proof of Theorem 4. ■

Theorem 6: The estimate of the squared distance between x and a column of A according to the Maximum Entropy Method is:

$$\text{distance}^2(x, a_j) \approx$$
$$d_{\text{classical}}(x, a_j) + \|x\|^2 - \|w_1^x\|^2 + \|a_j\|^2 - \|w_1^j\|^2$$

Proof: Both x and the random variable a_j can be described in terms of their projections on V_1, V_2.

$$x = V_1 w_1^x + V_2 w_2^x, \quad \widehat{a}_j = V_1 w_1^j + V_2 \widehat{w}_2^j$$

Calculating the squared norm of their difference:

$$\begin{aligned}&\|x-\widehat{a}_j\|^2\\&\quad=\|x\|^2+\|w_1^j\|^2+\|\widehat{w}_2^j\|^2-2(w_1^x)^Tw_1^j-2(w_2^x)^T\widehat{w}_2^j\\&\quad=\|w_1^x-w_1^j\|^2-\|w_1^x\|^2+\|\widehat{w}_2^j\|^2+\|x\|^2-2(w_2^x)^T\widehat{w}_2^j\end{aligned}$$

Going to expectations and using Equation (14) we get:

$$d_{\text{ent}}(x,a_j)=d_{\text{classical}}(x,a_j)+\|x\|^2-\|w_1^x\|^2+z_j$$

The theorem now follows from (13). ■

VI. Rayleigh Quotients

The Rayleigh Quotient (e.g., [14]) is given by the following formula:

$$r(v)=\frac{v^TBv}{\|v\|^2} \tag{22}$$

For a given matrix A we are interested in the two special cases of $B=AA^T$, and $B=A^TA$. In the first case the Rayleigh Quotient gives the sum of squared correlation between v and the columns of A. In the second case it gives the sum of squared correlation between v and the rows of A. Intuitively, the Rayleigh quotients measure the likelihood of the direction of v among the columns/rows of the matrix A.

The challenge we address here is how to estimate these Rayleigh Quotients when given the PCA of A instead of A itself. The classical solution is to replace A with its PCA representation, as given by (7). As in the case of distances the Maximum Entropy Method gives an improved solution.

A. *Column space Rayleigh quotient*

In this section we consider the case in which $B=AA^T$ in (22). For a vector $x\in\mathbb{R}^m$ the exact expression we wish to approximate is:

$$r(x)=\frac{x^TAA^Tx}{\|x\|^2}$$

When A is approximated as in (7) we have:

$$\begin{aligned}&r_{\text{classical}}(x)=\frac{x^TV_1W_1W_1^TV_1^Tx}{\|x\|^2}=\frac{\|W_1^Tw_1^x\|^2}{\|x\|^2}\\&\text{where } w_1^x=V_1^Tx.\end{aligned} \tag{23}$$

For the derivation of the Maximum Entropy solution we use the representation of A as a random matrix in (10).

$$\widehat{r}(x)=\frac{x^T\widehat{A}\widehat{A}^Tx}{\|x\|^2}$$

Taking expectations of both side and using the result of Equation (16) we get:

$$\begin{aligned}r_{\text{ent}}&=\frac{x^TV_1W_1W_1^TV_1^Tx+\delta(\|x\|^2-x^TV_1V_1^Tx)}{\|x\|^2}\\&=\frac{x^TV_1(W_1W_1^T-\delta I)V_1^Tx}{\|x\|^2}+\delta\\&=\frac{\|W_1^Tw_1^x\|^2}{\|x\|^2}+\delta(1-\frac{\|w_1^x\|^2}{\|x\|^2})=r_{\text{classical}}+\delta(1-\frac{\|w_1^x\|^2}{\|x\|^2})\end{aligned}$$

B. *Row space Rayleigh quotient*

In this section we consider the case in which $B=A^TA$ in (22). For a vector $y\in\mathbb{R}^n$ the exact expression we wish to approximate is:

$$r(y)=\frac{y^TA^TAy}{\|y\|^2}$$

When A is approximated as in (7) we have:

$$r_{\text{classical}}(y)=\frac{y^TW_1^TW_1y}{\|y\|^2}=\frac{\|W_1y\|^2}{\|y\|^2} \tag{24}$$

For the derivation of the Maximum Entropy approximation we use the representation of A as a random matrix in (10).

$$\widehat{r}(y)=\frac{y^T\widehat{A}^T\widehat{A}y}{\|y\|^2}$$

Taking expectations of both side and using the result in Equation (16) we get:

$$\begin{aligned}r_{\text{ent}}(y)&=\frac{y^T(W_1^TW_1+\text{Diag}(z_1,\ldots z_n))y}{\|y\|^2}\\&=r_{\text{classical}}(y)+\frac{y^T\text{Diag}(z_1,\ldots z_n)y}{\|y\|^2}\\&=r_{\text{classical}}(y)+\frac{\sum_{i=1}^n z_i(y(i))^2}{\|y\|^2}\end{aligned}$$

In summary we have the following formulas:

$$\begin{aligned}&\text{column space}\quad r_{\text{classical}}(x)=\frac{\|W_1^Tw_1^x\|^2}{\|x\|^2}\\&\text{column space}\quad r_{\text{ent}}(x)=r_{\text{classical}}+\delta(1-\frac{\|w_1^x\|^2}{\|x\|^2})\\&\text{row space}\quad r_{\text{classical}}(y)=\frac{\|W_1y\|^2}{\|y\|^2}\\&\text{row space}\quad r_{\text{ent}}=r_{\text{classical}}+\frac{\sum_{i=1}^n z_i(y(i))^2}{\|y\|^2}\end{aligned} \tag{25}$$

VII. Experimental results

We ran many experiments on various real datasets from the UC Irvine repository. In all cases the formulas derived using the Maximum Entropy Method produced better results than the classical formulas. The improvements were very significant on most datasets. The worst case we found was for the "wdbc" dataset, shown later. Experiments on three datasets are described in detail. They include the "Ionosphere" (size 34×351), the "wdbc" (size 30×569), and the "YearPredictionMDS" (size $90\times515,345$).

To experiment with column distances we measured the distances between all pairs of columns of the data matrix. In each case we computed the difference (in absolute value) between the estimated distance and the true distance. This is done for various k values. For the "YearPredictionMSD" dataset, because of the large number of observations we selected 50 columns at random, and computed the distances between all pairs in the selection. To measure the distances between an arbitrary vector x and columns of A, the vector

TABLE I: Distance(Ionosphere)

		x and columns of A		column distances	
		mean	std	mean	std
k= 1	$\lvert d_{\text{classical}} - d\rvert$	3.905E+01	1.218E+01	1.382E+01	1.075E+01
	$\lvert d_{\text{lower}} - d\rvert$	2.529E+01	1.743E+01	9.650E+00	1.021E+01
	$\lvert d_{\text{ent}} - d\rvert$	**3.645E+00**	**3.976E+00**	**2.161E+00**	**3.585E+00**
k= 3	$\lvert d_{\text{classical}} - d\rvert$	3.474E+01	1.096E+01	9.384E+00	9.432E+00
	$\lvert d_{\text{lower}} - d\rvert$	1.870E+01	1.537E+01	5.856E+00	7.551E+00
	$\lvert d_{\text{ent}} - d\rvert$	**2.716E+00**	**3.327E+00**	**1.171E+00**	**1.967E+00**
k= 5	$\lvert d_{\text{classical}} - d\rvert$	3.169E+01	9.788E+00	7.182E+00	7.927E+00
	$\lvert d_{\text{lower}} - d\rvert$	1.523E+01	1.375E+01	4.134E+00	5.952E+00
	$\lvert d_{\text{ent}} - d\rvert$	**2.351E+00**	**3.027E+00**	**7.608E-01**	**1.467E+00**
k= 10	$\lvert d_{\text{classical}} - d\rvert$	2.496E+01	7.874E+00	4.391E+00	4.993E+00
	$\lvert d_{\text{lower}} - d\rvert$	1.054E+01	9.897E+00	2.483E+00	3.720E+00
	$\lvert d_{\text{ent}} - d\rvert$	**1.789E+00**	**2.383E+00**	**4.655E-01**	**9.251E-01**

TABLE II: Distance(YearPredictionMSD)

		x and columns of A		column distances	
		mean	std	mean	std
k= 20	$\lvert d_{\text{classical}} - d\rvert$	4.453E+05	4.606E+05	8.808E+05	6.820E+05
	$\lvert d_{\text{lower}} - d\rvert$	9.876E+03	5.038E+03	7.094E+05	5.553E+05
	$\lvert d_{\text{ent}} - d\rvert$	**9.739E+02**	**9.650E+02**	**1.226E+05**	**2.111E+05**
k= 40	$\lvert d_{\text{classical}} - d\rvert$	1.034E+05	9.739E+04	2.047E+05	1.463E+05
	$\lvert d_{\text{lower}} - d\rvert$	3.986E+03	1.913E+03	1.701E+05	1.236E+05
	$\lvert d_{\text{ent}} - d\rvert$	**4.451E+02**	**4.330E+02**	**2.998E+04**	**4.735E+04**
k= 50	$\lvert d_{\text{classical}} - d\rvert$	4.914E+04	5.302E+04	9.736E+04	8.039E+04
	$\lvert d_{\text{lower}} - d\rvert$	2.550E+03	1.304E+03	8.003E+04	6.153E+04
	$\lvert d_{\text{ent}} - d\rvert$	**3.218E+02**	**2.932E+02**	**1.632E+04**	**2.559E+04**
k= 60	$\lvert d_{\text{classical}} - d\rvert$	1.691E+04	1.485E+04	3.332E+04	2.300E+04
	$\lvert d_{\text{lower}} - d\rvert$	1.298E+03	6.066E+02	2.875E+04	1.922E+04
	$\lvert d_{\text{ent}} - d\rvert$	**1.831E+02**	**1.601E+02**	**6.158E+03**	**8.067E+03**

x was drawn from Gaussian distribution with mean 0 and variance 1.

Tables I, II, and III show the average error of computing these distances with the various formulas. The left part shows the error mean and standard deviation of the formulas described in Section V-B. The right part shows the error mean and standard deviation of the formulas described in Section V-A. In all cases the mean and the standard deviation of the results computed by the Maximum Entropy Method were the best.

To quantify the advantage of the new formulas over the classical formulas we ran the following set of experiments. For a fixed value of k the formula d_{ent} was applied to the data and its error was measured. We then applied $d_{\text{classical}}$ and d_{lower} to the same data, and increased the value of k until they produced the approximately same error. The results for different datasets are shown in figures 2, 3, and 4. For example, Fig.2 was computed for the Ionosphare dataset. To obtain the same error of d_{ent} with $k = 2$ the formula d_{lower} needs $k = 24$,

TABLE III: Distance(WDBC)

		x and columns of A		column distances	
		mean	std	mean	std
k= 2	$\lvert d_{\text{classical}} - d\rvert$	1.980E+03	1.170E+04	3.522E+03	1.757E+04
	$\lvert d_{\text{lower}} - d\rvert$	3.478E+02	3.174E+02	1.878E+03	4.310E+03
	$\lvert d_{\text{ent}} - d\rvert$	**5.265E+01**	**6.826E+01**	**1.594E+03**	**2.463E+03**
k= 4	$\lvert d_{\text{classical}} - d\rvert$	7.563E+01	7.556E+01	9.894E+01	1.341E+02
	$\lvert d_{\text{lower}} - d\rvert$	5.945E+01	4.286E+01	6.877E+01	1.086E+02
	$\lvert d_{\text{ent}} - d\rvert$	**9.616E+00**	**1.074E+01**	**5.017E+01**	**6.652E+01**
k= 10	$\lvert d_{\text{classical}} - d\rvert$	1.990E+01	6.392E+00	1.044E-01	1.740E-01
	$\lvert d_{\text{lower}} - d\rvert$	1.627E+00	1.306E+00	6.845E-02	1.022E-01
	$\lvert d_{\text{ent}} - d\rvert$	**2.876E-01**	**3.430E-01**	**3.787E-02**	**6.065E-02**
k= 20	$\lvert d_{\text{classical}} - d\rvert$	9.954E+00	4.638E+00	5.085E-04	5.629E-04
	$\lvert d_{\text{lower}} - d\rvert$	8.621E-02	6.075E-02	3.935E-04	3.974E-04
	$\lvert d_{\text{ent}} - d\rvert$	**2.213E-02**	**2.244E-02**	**1.517E-04**	**1.786E-04**

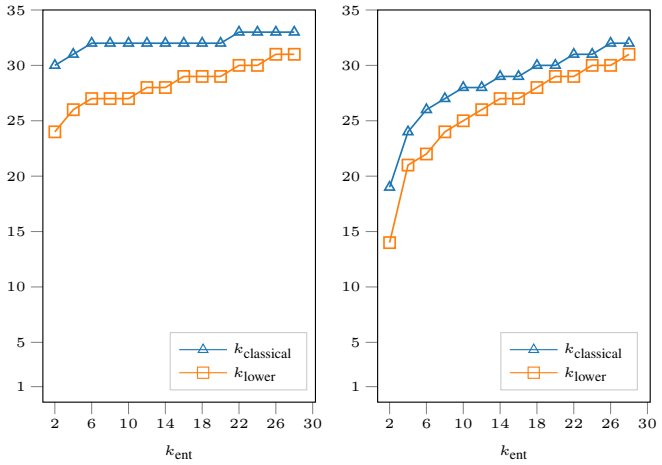

Fig. 2: Comparision of k with a fixed error value. Dataset: Ionosphere. Left panel: Distance between x and the columns of A; Right panel: Column distances.

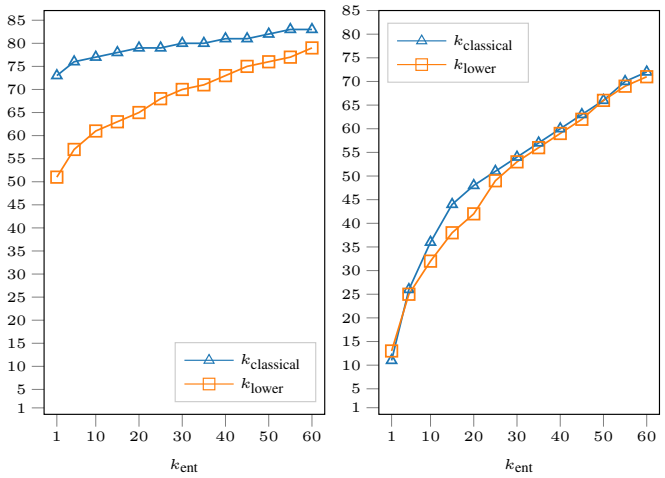

Fig. 3: Comparision of k with a fixed error value. Dataset: YearPredictionMSD. Left panel: Distance between x and the columns of A; Right panel: Column distances.

and the formula $d_{\text{classical}}$ needs $k = 30$. We observe that the advantage of d_{ent} over d_{lower} and $d_{\text{classical}}$ is quite significant. They are not that impressive for the "wdbc" dataset when computing the column distances.

A. *Experiments with Rayleigh Quotients*

Table IV describes the average difference in evaluating the column and row space Rayleigh quotient. Smaller mean and standard deviation of $|r_{\text{ent}} - r|$ indicate better estimates of Rayleigh quotient for the Maximum Entropy Method. The vectors evaluated in this experiment were randomly drawn from Gaussian distribution with mean 0 and variance 1. The plots in 5 show the advantage of the new formulas using the same format as in 2, 3, and 4.

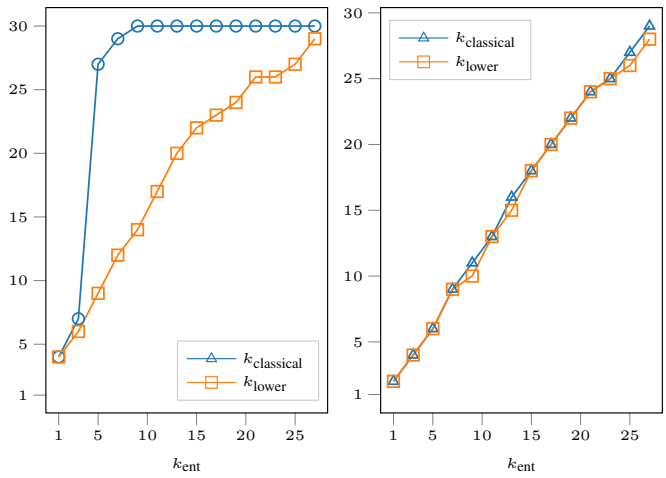

Fig. 4: Comparision of k with a fixed error value. Dataset: wdbc. Left panel: Distance between x and the columns of A; Right panel: Column distances.

TABLE IV: Rayleigh Quotient(Ionosphere)

		Column		Row	
		mean	std	mean	std
k= 2	$\frac{\lvert r_{\text{classical}} - r\rvert}{\lvert r_{\text{ent}} - r\rvert}$	6.151E+01 **1.401E+01**	1.800E+01 **9.829E+00**	5.733E+00 **1.475E+00**	2.098E+00 **1.271E+00**
k= 6	$\frac{\lvert r_{\text{classical}} - r\rvert}{\lvert r_{\text{ent}} - r\rvert}$	3.507E+01 **4.152E+00**	6.625E+00 **2.999E+00**	3.074E+00 **6.486E-01**	8.803E-01 **4.849E-01**
k= 10	$\frac{\lvert r_{\text{classical}} - r\rvert}{\lvert r_{\text{ent}} - r\rvert}$	2.383E+01 **2.125E+00**	3.688E+00 **1.431E+00**	2.370E+00 **6.627E-01**	9.197E-01 **5.844E-01**
k= 14	$\frac{\lvert r_{\text{classical}} - r\rvert}{\lvert r_{\text{ent}} - r\rvert}$	1.423E+01 **1.519E+00**	4.056E+00 **1.103E+00**	1.401E+00 **3.334E-01**	3.911E-01 **2.417E-01**
k= 18	$\frac{\lvert r_{\text{classical}} - r\rvert}{\lvert r_{\text{ent}} - r\rvert}$	9.712E+00 **1.139E+00**	2.699E+00 **8.474E-01**	9.291E-01 **2.149E-01**	3.089E-01 **1.376E-01**
k= 22	$\frac{\lvert r_{\text{classical}} - r\rvert}{\lvert r_{\text{ent}} - r\rvert}$	5.922E+00 **6.551E-01**	2.281E+00 **6.374E-01**	6.001E-01 **1.618E-01**	2.420E-01 **1.417E-01**
k= 26	$\frac{\lvert r_{\text{classical}} - r\rvert}{\lvert r_{\text{ent}} - r\rvert}$	3.007E+00 **2.995E-01**	1.276E+00 **2.960E-01**	3.354E-01 **1.353E-01**	2.059E-01 **1.416E-01**

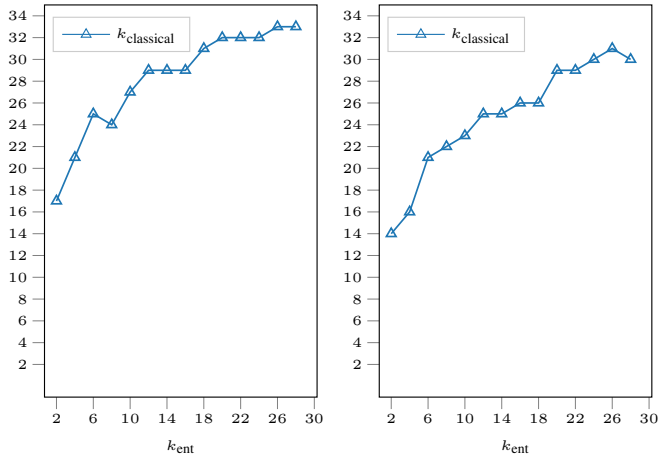

Fig. 5: Comparision of k values with fixed error value. Dataset: Ionosphere. Left panel: Column space Rayleigh Quotient; Right panel: Row space Rayleigh Quotient.

VIII. Concluding Remarks

This paper considers a common situation in which a matrix A is approximated by PCA as: $A = VW$. A nice aspect of this representation is that a lot of the operations that involve matrix data can be performed "in the PCA space", without reconstructing the matrix or any of its columns. The paper discusses two of this cases. The first is computing distances that involve matrix columns, and the second is the computation of Rayleigh quotients.

Our main result is a novel method of modeling the uncertainty in the estimates that one obtains from PCA approximations. The idea is to replace the unknown quantities with random variables. Using information that is typically available during the creation of the matrix W in the above estimation and the Maximum Entropy Method one can determine the likely distribution of these random variables. Thus, evaluating expressions that involve the matrix A become estimates of expected values.

Applying this framework allows us to derive closed form solutions to distances and Rayleigh quotients that appear to be novel. Experimental results show that these new formulas produce a significant improvement in accuracy, when compared to the classical formulas.

References

[1] T. I. H. Consortium, "A haplotype map of the human genome," *Nature*, vol. 437, pp. 1299–1320, 2005.

[2] C. Burges, *Dimension Reduction: A Guided Tour*. Hanover, MA, USA: Now Publishers Inc., January 2010.

[3] V. Gray, *Principal Component Analysis: Methods, Applications and Technology*, ser. Mathematics Research Developments. Nova Science Publishers, Incorporated, 2017.

[4] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. Springer-Verlag, 2002.

[5] B. He, S. Shah, C. Maung, G. Arnold, G. Wan, and H. Schweitzer, "Heuristic search algorithm for dimensionality reduction optimally combining feature selection and feature extraction," in *Proceedings of the 33rd National Conference on Artificial Intelligence (AAAI'19)*. AAAI Press, 2019, p. in press.

[6] S. Shah, B. He, C. Maung, and H. Schweitzer, "Computing robust principal components by A* search," *International Journal on Artificial Intelligence Tools*, vol. 27, no. 7, November 2018.

[7] J. Cadima and I. Jolliffe, "On relationships between uncentred and column-centred principal component analysis," *Pakistan Journal of Statistics*, vol. 25, no. 4, pp. 473–503, 10 2009.

[8] R. Weber, H. J. Schek, and S. Blott, "A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces," in *VLDB '98*, 1998, pp. 194–205.

[9] T. F. Cox and M. A. Cox, *Multidimensional Scaling*. Chapman & Hall, 1994.

[10] E. T. Jaynes, "On the rationale of maximum entropy methods," *Proceedings of IEEE*, vol. 70, no. 9, pp. 939–952, Sep. 1982.

[11] A. Papoulis, *Probability, random Variables, and Stochastic Processes*, 2nd ed. McGraw-Hill, 1984.

[12] Wikipedia contributors, "Principle of maximum entropy — Wikipedia, the free encyclopedia," 2019. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Principle_of_maximum_entropy

[13] M. Rozanski, R. Witula, and E. Hetmaniok, "More subtle versions of the Hadamard inequality," *Linear Algebra and its Applications*, vol. 532, pp. 500–511, Nov. 2017.

[14] G. H. Golub and C. F. Van-Loan, *Matrix Computations*, 4th ed. Johns Hopkins University Press, 2013.