Metric Correction of Similarities Based on Orthogonal Decomposition

Sergey Dvoenko
Institute of Applied Math and Computer Sciences
Tula State University
Tula, Russia
dvsrge@gmail.com

Denis Pshenichny
Institute of Applied Math and Computer Sciences
Tula State University
Tula, Russia
denispshenichy@yandex.ru

Abstract. **Raw data in modern machine learning usually appear as similarities or dissimilarities between members of a limited set. A positive definite similarity matrix represents a limited set of elements immersed in some metric space with dimensionality right up to the matrix size with similarities considered as scalar products. In a case of a nonpositive definite similarity matrix, it needs metric correction of similarities to be considered as scalar products. The known discrete Karhunen-Loeve expansion is usually used to reduce the dimensionality of the similarity matrix by removing eigenvectors corresponded to negative eigenvalues. As a result, a new similarity matrix of the reduced size is calculated to immerse members of a limited set in a reduced space of eigenvectors corresponded only to positive eigenvalues with data dispersion reduced. According to an orthogonal decomposition based metric correction here, it is proposed not to remove, but change negative eigenvalues to become positive ones. As a result, such an optimal correction preserves the dimensionality and dispersion of raw data.**

Keywords: **similarity matrix, orthogonal decomposition, eigenvector, eigenvalue, Karhunen-Loeve expansion**

I. Introduction

Let raw data be presented by a similarity matrix of a set of a limited size. It is known, the positive definite square matrix has the positive determinant and their eigenvalues are positive too [1]. In this case the set members can be represented in some multidimensional coordinate space by vectors with distances and scalar products between them calculated based on the cosine theorem. The end points of normalized vectors appear to be arranged on the hypersphere of the unit radius. For all positive scalar products all corresponding vectors are located in the positive quadrant of the coordinate space.

In the mathematical sense, paired comparisons must be immersed in some metric (Euclidean) space. This is the well-known theoretical problem [2]. Under modern conditions, this problem becomes practical in machine learning, data mining, image processing, etc. [3].

Nevertheless, empirical functions for paired comparisons are usually not correct mathematical functions of distances or similarities. Using of them usually results in so-called metric violations in the set configuration in some space. Hence, it needs to recover metric by correction of paired comparisons. Violations appear in negative eigenvalues of the similarity matrix of the set elements.

We have developed before and today we are improving the novel end-to-end correction technology for optimal recovering of a violated metric. As a result, the positive definiteness of the corrected matrix is achieved [4–8].

The originality of such approach consists in the following. Indeed, each metric violation is connected with some member of the set which is supposed to be responsible for the violation. This approach differs from the well-known multidimensional scaling problem, since it doesn't need to recover explicitly the feature space itself.

In this paper, another approach is proposed based on the known Karhunen-Loeve expansion in terms of a system of orthogonal functions. It is known, we face the problem of the spectral decomposition of a square matrix based on their eigenvectors [1].

II. Decomposition of a Matrix of Scalar Products Based on Orthogonal Vectors

Let the set of n objects be represented by the normalized matrix $S(n,n)$ of paired comparisons with elements $s_{ii}=1$ for the main diagonal and values $0 \le s_{ij} < 1$ or $-1 < s_{ij} < 1$ for others.

The spectral decomposition of the nondegenerated matrix $S(n,n)$ has the form $S = A\Lambda A^T$, where

This work is supported by RFBR Grants 20-07-00055, 19-07-01178.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

$A(n,n) = (\mathbf{a}_1, \dots \mathbf{a}_n)$ is the orthogonal matrix $A^T A = A^{-1} A = E$ of eigenvectors-columns $\mathbf{a}_i = (a_{1i}, \dots a_{ni})^T$, $|\mathbf{a}_i| = 1$, $E(n,n)$ is the unity matrix, $\Lambda(n,n)$ is the diagonal matrix of eigenvalues sorted in the decreasing order.

It is known, the set represented by the matrix $S(n,n)$ is correctly immersed in some metric (Euclidean) space of the dimensionality not more than *n*, and the matrix itself consists of normalized scalar products of the set elements.

It needs to note immediately, the raw data matrix $X(n,m)$ and the corresponding feature space $\mathbb{R}^m$ with the dimensionality $m \le n$ are not presented here. We suppose they have been lost, otherwise scalar products of objects $S'(n,n) = (1/m) X X^T$ would be properly calculated.

It is easy to see that $\Lambda = A^T S A$ with $tr\,\Lambda = tr\,S = n$ based on the decomposition of the nondegenerated normalized matrix S of scalar products of objects.

III. Correction of Metric Violations

In data analysis problem, the decomposition in terms of orthogonal vectors usually used for a correlation matrix of features $R(m,m) = (1/m) X^T X$ targeted to reduce the dimensionality of the space of eigenvectors of the matrix $R(m,m)$ to the new value $m' < m$ in different tasks.

In particular, the noncorrect matrix $R(m,m)$ has negative eigenvalues. Hence, the projection of the initial data matrix $X(n,m)$ in the new orthogonal subspace $\mathbb{R}^{m'}$ is usually used. The new dimensionality is defined only by positive eigenvalues $\lambda'_1 > \dots > \lambda'_{m'} > 0$ of the initial decomposition (it is the so-called discrete Karhunen-Loeve expansion [9]). The new correlation matrix $\Lambda'(m',m')$ has the diagonal form, where $tr\,\Lambda' = \sum_{i=1}^{m'} \lambda' = m' < m$.

Note, such the projection requires the data matrix $X(n,m)$ to get the new one $X'(n,m')$. In this case, the dispersion of raw data (the total dispersion of normalized features) is reduced to $m' < m$.

Unlike the classical approach, the spectral decomposition in our investigation is used for the set members represented only by scalar products or non-negative similarities (scalar products in the positive quadrant of a metric space) $S(n,n)$. It doesn't matter, what they are: features or objects themselves.

In a case of metric violations in the set configuration in hypothetical space (we have not it), the spectral decomposition $S = A \Lambda A^T$ of the matrix $S(n,n)$ has negative eigenvalues.

In this paper, it is proposed not to reduce initial data, but replace negative eigenvalues in the decomposition by the appropriate positive ones to get as a result the new matrix $\tilde{\Lambda}(n,n)$ of the same dimensionality. After that, the matrix $\tilde{S}(n,n)$ is recovered in the form of $\tilde{S} = A \tilde{\Lambda} A^T$.

Note, the new matrix $\tilde{S}$ appears to be nonnormalized, since their diagonal elements appear to be more, than 1. Therefore, $tr\,\tilde{\Lambda} = tr\,\tilde{S} > n$. After the transformation $\hat{s}_{ij} = \tilde{s}_{ij} / \sqrt{\tilde{s}_{ii} \tilde{s}_{jj}}$, the decomposition of the corrected matrix $\hat{S}(n,n)$ is specified as $\hat{S} = \hat{A} \hat{\Lambda} \hat{A}^T$, $tr\,\hat{\Lambda} = tr\,\hat{S} = n$, where eigenvalues and eigenvectors take the final form.

In fact, based on this approach any set of eigenvalues can be modified by any other values to eliminate not only negative eigenvalues.

Note, based on such the approach, it is possible to formulate any suitable problems to find the corresponding set of eigenvalues for similarity matrices, for example, to provide the right level of the conditionality for the corrected matrix (we demonstrate it below), etc.

Additionally, the raw similarity matrix can have other type of violations, where non-diagonal elements exceed diagonal ones by the module. In this case, the correction recovers the matrix too. Naturally, the raw matrix must be of the full rank. In other case, the matrix dimensionality needs to be reduced before the correction.

IV. Using the Proposed Approach on Real Data

Let the correlation matrix $S(11{,}11)$ be given. It represents statistical interconnections between power of biorhythms of 11 frequencies in a brain like alpha-, beta-, and theta-rhythms in electroencephalograms.

Psychologist V.D. Nebylytsin obtained such data [10] during his investigations of light-sound sensations imposed by rhythms

$$\left|\begin{array}{ccccccccccc}
1 & 0.562 & 0.568 & 0.152 & 0.347 & 0.250 & 0.264 & -0.020 & -0.212 & -0.086 & -0.076 \\
0.562 & 1 & 0.784 & 0.057 & 0.196 & 0.218 & 0.009 & -0.017 & -0.002 & 0.163 & 0.284 \\
0.568 & 0.784 & 1 & 0.288 & 0.475 & 0.264 & 0.066 & 0.144 & 0.114 & 0.228 & 0.151 \\
0.152 & 0.057 & 0.288 & 1 & 0.686 & 0.293 & 0.034 & 0.048 & -0.069 & -0.064 & 0.175 \\
0.347 & 0.196 & 0.475 & 0.686 & 1 & 0.429 & 0.070 & 0.152 & 0.036 & 0.028 & 0.216 \\
0.250 & 0.218 & 0.264 & 0.293 & 0.429 & 1 & 0.788 & 0.197 & 0.154 & 0.109 & 0.035 \\
0.264 & 0.009 & 0.066 & 0.034 & 0.070 & 0.788 & 1 & 0.109 & 0.054 & -0.002 & -0.018 \\
-0.020 & -0.017 & 0.144 & 0.048 & 0.152 & 0.197 & 0.109 & 1 & 0.807 & 0.830 & 0.699 \\
-0.212 & -0.002 & 0.114 & -0.069 & 0.036 & 0.154 & 0.054 & 0.807 & 1 & 0.904 & 0.728 \\
-0.086 & 0.163 & 0.228 & -0.064 & 0.028 & 0.109 & -0.002 & 0.830 & 0.904 & 1 & 0.768 \\
-0.076 & 0.284 & 0.151 & 0.175 & 0.216 & 0.035 & -0.018 & 0.699 & 0.728 & 0.768 & 1
\end{array}\right| .$$

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

This matrix contains ten positive eigenvalues 3.636340, 2.827085, 1.611613, 1.358204, 0.515165, 0.412792, 0.278171, 0.164165, 0.151054, 0.069977, and the last negative eigenvalue -0.024566. We have already studied this matrix in order to correct it and determine the optimal conditionality based on the end-to-end metric correction technology [6-8].

Replacing a negative eigenvalue with a practically zero value 10^{-5} gives, after normalization, eigenvalues 3.629217, 2.821157, 1.605834, 1.355937, 0.514139, 0.411444, 0.277702, 0.163727, 0.150881, 0.069932, 0.00003, and the corrected matrix

$$\begin{vmatrix}
1 & 0.558 & 0.568 & 0.152 & 0.346 & 0.251 & 0.263 & -0.021 & -0.212 & -0.086 & -0.074 \\
0.558 & 1 & 0.776 & 0.058 & 0.198 & 0.212 & 0.013 & -0.015 & -0.0001 & 0.164 & 0.277 \\
0.568 & 0.776 & 1 & 0.286 & 0.472 & 0.267 & 0.063 & 0.142 & 0.112 & 0.227 & 0.154 \\
0.152 & 0.058 & 0.286 & 1 & 0.686 & 0.291 & 0.035 & 0.048 & -0.069 & -0.064 & 0.173 \\
0.346 & 0.198 & 0.472 & 0.686 & 1 & 0.425 & 0.072 & 0.153 & 0.037 & 0.029 & 0.212 \\
0.251 & 0.212 & 0.267 & 0.291 & 0.425 & 1 & 0.782 & 0.195 & 0.152 & 0.108 & 0.039 \\
0.263 & 0.013 & 0.063 & 0.035 & 0.072 & 0.782 & 1 & 0.110 & 0.055 & -0.001 & -0.021 \\
-0.021 & -0.015 & 0.142 & 0.048 & 0.153 & 0.195 & 0.110 & 1 & 0.807 & 0.830 & 0.696 \\
-0.212 & -0.0001 & 0.112 & -0.069 & 0.037 & 0.152 & 0.055 & 0.807 & 1 & 0.904 & 0.724 \\
-0.086 & 0.164 & 0.227 & -0.064 & 0.029 & 0.108 & -0.001 & 0.830 & 0.904 & 1 & 0.765 \\
-0.074 & 0.277 & 0.154 & 0.173 & 0.212 & 0.039 & -0.021 & 0.696 & 0.724 & 0.765 & 1
\end{vmatrix}$$

with the large conditionality value 122035.7, which is calculated as the ratio of the maximal and minimal eigenvalues under the assumption that they are nonnegative ones.

Earlier in [7, 8], a statistically inspired conditionality value of 59.409 was found, which can be taken as a basis here, although it was found for another correction method.

Replacing a negative eigenvalue with a positive value 0.0612 gives after normalization eigenvalues 3.611811, 2.806524, 1.591663, 1.350369, 0.511535, 0.408003, 0.276734, 0.162535, 0.150530, 0.069870, 0.060426, and the corrected matrix

$$\begin{vmatrix}
1 & 0.550 & 0.569 & 0.150 & 0.343 & 0.253 & 0.259 & -0.022 & -0.214 & -0.087 & -0.070 \\
0.550 & 1 & 0.757 & 0.061 & 0.204 & 0.197 & 0.022 & -0.011 & 0.005 & 0.166 & 0.260 \\
0.569 & 0.757 & 1 & 0.283 & 0.463 & 0.273 & 0.056 & 0.139 & 0.108 & 0.224 & 0.162 \\
0.150 & 0.061 & 0.283 & 1 & 0.686 & 0.287 & 0.037 & 0.049 & -0.067 & -0.063 & 0.169 \\
0.343 & 0.204 & 0.463 & 0.686 & 1 & 0.416 & 0.076 & 0.155 & 0.039 & 0.030 & 0.204 \\
0.253 & 0.197 & 0.273 & 0.287 & 0.416 & 1 & 0.768 & 0.190 & 0.147 & 0.105 & 0.050 \\
0.259 & 0.022 & 0.056 & 0.037 & 0.076 & 0.768 & 1 & 0.113 & 0.058 & 0.001 & -0.030 \\
-0.022 & -0.011 & 0.139 & 0.049 & 0.155 & 0.190 & 0.113 & 1 & 0.807 & 0.830 & 0.687 \\
-0.214 & 0.005 & 0.108 & -0.067 & 0.039 & 0.147 & 0.058 & 0.807 & 1 & 0.904 & 0.715 \\
-0.087 & 0.166 & 0.224 & -0.063 & 0.030 & 0.105 & 0.001 & 0.830 & 0.904 & 1 & 0.757 \\
-0.070 & 0.260 & 0.162 & 0.169 & 0.204 & 0.050 & -0.030 & 0.687 & 0.715 & 0.757 & 1
\end{vmatrix}$$

with the practically optimal conditionality 59.77235484.

It is easy to see, all matrices are practically the same. The optimality of the last correction case is supported by the Karhunen-Loeve expansion properties.

V. Conclusion

In this paper, we propose a new approach to the metric correction of matrices of paired comparisons ased on the spectral decomposition of the square matrix of scalar products in terms of its eigenvectors.

Optimality of the correction is supported by the properties of the Karhunen-Loeve expansion and the correct selecting of the new value of the corresponding eigenvalue.

From the other side, here we face the well-known perturbation theory of eigenvalue problems [11, 12]. In the framework of this general theory, the novelty of our approach consists in the attempt to make a perturbation in eigenvalues first to recover then similarity matrices used in machine learning.

In further research, all developed techniques is planned to implement for some actual problems, for example, for using different quality metrics in machine learning, for multimodal analysis of heterogeneous data in formal concept analysis [13], etc.

References

[1] J. N. Franklin, Matrix Theory. Mineola, N.Y., Dover Publications, 2000. 292 p.

[2] R. L. Bishop, R. J. Crittenden, Geometry of manifolds. N.Y., Acade-mic Press, 1964. 273 p.

[3] E. Pekalska, R. P. W. Duin, The dissimilarity representation for pattern recognition. Foundations and applications. Singapore: World Scientific, 2005. 607 p.

[4] S. D. Dvoenko, D. O. Pshenichny, "A recovering of violated metric in machine learning", Proc. of SoICT'16. ACM, N.Y., 2016, pp. 15-21. DOI: 10.1145/3011077.3011084

[5] S. D. Dvoenko, D. O. Pshenichny, "On metric correction and conditi-onality of raw featureless data in machine learning", Pattern Recognition and Image Analysis, vol. 28(4), 2018, pp. 595–604. DOI: 10.1134/S1054661818040089.

[6] S. D. Dvoenko, D. O. Pshenichny, "Optimal correction of metrical violations in matrices of pairwise comparisons", Machine Learning and Data Analysis, 1 (7), 2014, pp. 885-890 (in Russian).

[7] S. D. Dvoenko, D. O. Pshenichny, "The conditionality of matrices of pairwise comparisons after metric corrections", Machine Learning and Data Analysis, 3(1), 2017, pp. 50-60. DOI: 10.21469/ 22233792.3.1.04 (in Russian).

[8] S. D. Dvoenko, D, O. Pshenichny, "A corrrection of metric violations based on statistical hypotheses testing", Bulletin of the Tula State University. Technical science, 10, 2018, pp. 100-107 (in Russian).

[9] J. T. Tou, R. C. Gonzalez, Pattern Recognition Principles, Addison-Wesley, 2nd ed., 1974. 377 p.

[10] V. D. Nebylitsyn, Selected psychological proceedings. Moscow, Pedagogika, 1990, 408 p. (in Russian).

[11] R. Bellman, Introduction to Matrix Analysis, N.Y., McGrow-Hill, 1970, 440 p.

[12] F. Rellich, Perturbation Theory of Eigenvalue Problems, Routledge, 1969, 138 p.

[13] M. Bogatyrev, D. Orlov, "Application of formal contexts in the analysis of heterogeneous biomedical data", Proc. of CEUR Workshop, RCAI, 2020, Oct. 10-16, Moscow, vol. 2648, pp. 315-329.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.