Comments

On Distributional Assumptions and Whitened Cosine Similarities

Marco Loog

Abstract—Recently, an interpretation of the whitened cosine similarity measure as a Bayes decision rule was proposed [1]. This communication makes the observation that some of the distributional assumptions made to derive this measure are very restrictive and, considered simultaneously even inconsistent.

Index Terms—Decision rule, class distributions, distributional assumptions, consistency, whitened cosine similarity.

1 Introduction

A paper [1] published recently in this journal postulates four specific assumptions in order to relate the so-called whitened cosine similarity measure to a Bayes decision rule for classification. In this, the whitened cosine similarity measure between two vectors x and y is defined as

$$d(x,y) = \frac{(W^{\mathrm{t}}x)^{\mathrm{t}}(W^{\mathrm{t}}y)}{\|W^{\mathrm{t}}x\|\|W^{\mathrm{t}}y\|} = \frac{x^{\mathrm{t}}T^{-1}y}{\|W^{\mathrm{t}}x\|\|W^{\mathrm{t}}y\|}, \tag{1}$$

which is the arccosine of the angle between the vectors $W^{\mathrm{t}}x$ and $W^{\mathrm{t}}y$. Here, T is the overall, or total, covariance matrix taken over all data and W is a whitening transformation which diagonalizes T, i.e., $W^{\mathrm{t}}TW = I$, where I is the identity matrix.

The specific assumptions are (as quoted from [1]):

1. The conditional probability density functions of all of the classes are multivariate normal.
2. The prior probabilities of all of the classes are equal.
3. The covariance matrices of all of the classes are identical to the covariance matrix of all samples, regardless of their class membership.
4. The whitened pattern vectors in the Bayes decision rule are normalized to unit norm.

Based on the first three, the following decision rule can be derived (see, for instance, [2]):

$$c(x) = \underset{i\in 1,\ldots,K}{\operatorname{argmin}} (x-m_i)^{\mathrm{t}}T^{-1}(x-m_i). \tag{2}$$

In [1], the fourth assumption is added to the list in order to carry out the following further manipulations to the rule:

$$\begin{aligned} c(x) &= \underset{i\in 1,\ldots,K}{\operatorname{argmin}} (x-m_i)^{\mathrm{t}}T^{-1}(x-m_i) \\ &= \underset{i\in 1,\ldots,K}{\operatorname{argmin}} (W^{\mathrm{t}}(x-m_i))^{\mathrm{t}}(W^{\mathrm{t}}(x-m_i)) \\ &= \underset{i\in 1,\ldots,K}{\operatorname{argmin}} \|W^{\mathrm{t}}x\|^2 + \|W^{\mathrm{t}}m_i\|^2 \\ &\quad - 2\|W^{\mathrm{t}}x\|\|W^{\mathrm{t}}m_i\|d(x,m_i) \\ &= \underset{i\in 1,\ldots,K}{\operatorname{argmin}} d(x,m_i), \end{aligned} \tag{3}$$

where the last equality follows because Assumption 4 is meant to be interpreted as $\|W^{\mathrm{t}}y\| = 1$, both when y is a normal pattern vector x and a mean m_i (see [1]).

In what follows, we question the usefulness of the previous set of assumptions, offer some insight into some of the theoretical intricacies involved, and, ultimately, make the observation that the assumptions are inconsistent. That is, there are no data distributions that can possibly fulfill all four requirements.

2 Analysis of the Assumptions

2.1 Assumption 3

To start with, we note that Assumption 3 in itself is already very restrictive. It states—besides the often-made assumption that all class covariance matrices C_i are equal—that these should also be equal to the "covariance matrix of all samples." That is, they should all equal the total covariance matrix T.

This matrix T is the sum of the pooled within-class covariance and the between-class covariance [3], i.e., $T = \sum_{i=1}^{K} p_i C_i + \sum_{i=1}^{K} p_i (m_i - m)(m_i - m)^{\mathrm{t}}$, in which p_i are the class priors and m is the overall mean and, because $T = C_i$, for all i. This leads to $T = \sum_{i=1}^{K} p_i T + \sum_{i=1}^{K} p_i (m_i - m)(m_i - m)^{\mathrm{t}} = T + B$ with B the between class covariance $\sum_{i=1}^{K} p_i (m_i - m)(m_i - m)^{\mathrm{t}}$. Consequently, it holds that $B = 0$ and there is no variation in the class means. In other words, Assumption 3 implies that all class means are the same.

Obviously, this seems too restrictive an assumption. Together with Assumptions 1 and 2, this would, for instance, imply that all classes are completely overlapping and no discrimination is possible.

Incidentally, note that not much is gained from the relaxation of Assumption 3 to allow all C_i to be proportional to the total covariance, i.e., $C_i = \alpha T$ for all i. This would imply that B is proportional to T as well. In the general setting, where the number of classes is smaller than or equal to the number of feature dimensions, the rank of B is strictly smaller than T's and, therefore, B can only be proportional to T if it equals the null matrix, i.e., it reduces to the case already considered. On the other hand, in the less common situation where the number of classes is larger than the dimensionality, B can be in proportion to T in a nondegenerate way. It is, however, nontrivial to construct such data configurations in which the within, between, and total covariances are all equal apart from certain multiplicative constants.

2.2 Assumption 4

Assumption 4 is even more troublesome and what follows are two main difficulties that stem from imposing this requirement.

First of all, initially it is not made precise what is meant by "pattern vector." From the paper's context, it at least seems the case that all samples x are included. As indicated at the end of Section 1, the derivation in [1] in addition assumes that the whitened means are normalized as well. The paper states (using current notation): "Now, if we add another assumption, 4) the whitened pattern vectors, $W^{\mathrm{t}}x$ and $W^{\mathrm{t}}m_i$, are normalized to unit norm, then [...]." That is, both samples and class mean vectors are assumed to have unit length after applying the whitening transform. But, this is problematic as if all pattern vectors are located on a unit hypersphere, all samples from the same class have to coincide in order for

• *The author is with the Pattern Recognition Group, Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology, Delft, The Netherlands. E-mail: m.loog@tudelft.nl.*

Manuscript received 2 Aug. 2007; revised 18 Oct. 2007; accepted 3 Dec. 2007; published online 12 Dec. 2007.
Recommended for acceptance by B. Triggs.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2007-08-0472.
Digital Object Identifier no. 10.1109/TPAMI.2007.70838.

the mean to lie on the hypersphere as well. As soon as there is some spread in the class samples, the mean will lie inside the sphere and not on it and the mean will not have unit length. Again, this seems an assumption that is too restrictive.

Furthermore, even if we would only consider the samples to lie on the unit hypersphere, Assumption 4 would be difficult to fulfill. In this case, however, the problem is somewhat more subtle.

Consider the interval $[-1, 1]$ and the distribution $P = \frac{1}{2}(\delta_{-1} + \delta_1)$ on this interval. As all mass is distributed in the endpoints -1 and 1 with equal probabilities $\frac{1}{2}$, the variance of this distribution is one. Moreover, it is not hard to see that, as soon as some of the mass shifts, e.g., to the inside of the interval, the variance (the average spread) will decrease. That is, for all distributions other than P, the variance is strictly smaller than one.

Now, the assumption states that the data on the hypersphere is whitened, i.e., that the total covariance equals the identity matrix I. This, in turn, implies that, for every unit vector v, the variance of the projected data $v^{\mathrm{t}}x$ equals $v^{\mathrm{t}}Iv = v^{\mathrm{t}}v = 1$. The projected data, coming from a hypersphere, lies of course in the interval $[-1, 1]$, but is generally not concentrated in the endpoints of this interval and will therefore have variance less than 1, which contradicts the whitening assumption.

We can conclude that Assumption 4 can hold only for one-dimensional data and then merely in the case where it is distributed like P, a single degenerate distribution.

Finally, we may remark that another possibility is to merely require the whitened class means to be normalized to unit norm. Obviously, this is an assumption less limitative than the original one and there exist distributions that fulfill this requirement. Nonetheless, it still is overly restrictive. As given in [4], B can be expressed in terms of pairwise class mean differences: $B = \sum_{i=1}^{K-1}\sum_{j=i+1}^{K} p_i p_j (m_i - m_j)(m_i - m_j)^{\mathrm{t}}$. Using this, we can give an upper bound for the trace of the between-class covariance in the whitened feature space:

$$\begin{aligned}
&\mathrm{tr}\left(\sum_{i=1}^{K-1}\sum_{j=i+1}^{K} p_i p_j (W^{\mathrm{t}}m_i - W^{\mathrm{t}}m_j)(W^{\mathrm{t}}m_i - W^{\mathrm{t}}m_j)^{\mathrm{t}}\right)\\
&\quad = \sum_{i=1}^{K-1}\sum_{j=i+1}^{K} p_i p_j (W^{\mathrm{t}}m_i - W^{\mathrm{t}}m_j)^{\mathrm{t}}(W^{\mathrm{t}}m_i - W^{\mathrm{t}}m_j) \qquad (4)\\
&\quad \leq \sum_{i=1}^{K-1}\sum_{j=i+1}^{K} p_i p_j 4 \leq \frac{1}{2}K(K-1),
\end{aligned}$$

where the first inequality follows because the squared distance between every two class means, i.e., $(W^{\mathrm{t}}m_i - W^{\mathrm{t}}m_j)^{\mathrm{t}}(W^{\mathrm{t}}m_i - W^{\mathrm{t}}m_j)$, is always smaller than or equal to 4. The last one holds because $p_i p_j$ is always smaller than or equal to $\frac{1}{4}$. Also, as the total covariance in the whitened space equals its dimensionality d, the trace of the within-class must be larger than $d - \frac{1}{2}K(K-1)$.

Apart from the unpleasant restrictions on the between and within-class covariances, it also follows, for example, that, for large dimensionality and relatively few classes, the assumption implies that the total covariance has to be dominated by the within-class covariance. There is, however, no need to impose this unit norm restriction with such undesirable consequences. A preferable alternative is to simply assume the norms of the whitened means to be the same, but not necessarily equal to one.

2.3 Inconsistency

As illustrated, Assumptions 3 and 4 are very limiting. Individually, however, they do not render the set of assumptions inconsistent. Yet, together they do.

As we saw, Assumption 4 implies that all classes have to be degenerate with all class samples lying in the same point, i.e., their class mean. Moreover, Assumption 3 implies that there is no spread in the class means and they as well have to be located in one and the same point. In a word, all data from all classes is situated within a single point. This, however, contradicts Assumption 4, which states that the data has been whitened, implying the spread to be one. We therefore conclude that the set of assumptions postulated in [1] is inconsistent and no configuration of data can possibly fulfill them.

ACKNOWLEDGMENTS

Dr. Chengjun Liu is kindly acknowledged for his appraisal of an earlier draft by the author on which the current paper is based. His comments and suggestions helped to strengthen the overall exposition. Two suggestions of his are considered and commented on at the end of Sections 2.1 and 2.2, respectively. The author would also like to thank the three anonymous reviewers for their thoughtful comments and associate editor Dr. Bill Triggs, in addition, for making some pointed remarks. Largest parts of this work were carried when the author was with the Datalogical Institute, University of Copenhagen, Copenhagen, Denmark, and with Nordic Bioscience, Herlev, Denmark.

REFERENCES

[1] C. Liu, "The Bayes Decision Rule Induced Similarity Measures," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1086-1090, June 2007.
[2] G.J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, 1992.
[3] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.
[4] M. Loog, R.P.W. Duin, and R. Haeb-Umbach, "Multiclass Linear Dimension Reduction by Weighted Pairwise Fisher Criteria," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 7, pp. 762-766, July 2001.