

# A Statistical Analysis of Fukunaga Koontz Transform

Xiaoming Huo

Dr. Xiaoming Huo is an assistant professor at the School of Industrial and System Engineering of the Georgia Institute of Technology, Atlanta, Georgia, USA. (email: [xiaoming@isye.gatech.edu](mailto:xiaoming@isye.gatech.edu), phone: 404 385 0354.)

## Abstract

The Fukunaga Koontz Transform (FKT) has been proposed as a feature selection methodology for nearly 32 years. There are a huge number of citations. We have not seen a direct analysis between FKT, and a well established statistical method: Fisher Quadratic Discriminant Analysis (QDA). In this paper, under certain assumptions, we establish such a connection. We speculate that a link in a more general situation is hard to find.

## I. INTRODUCTION

Since Fukunaga and Koontz published their paper [5] on feature selection in 1970, many researchers have used their method in various applications, such as target recognition, face detection [4], etc. A citation search of this paper can easily generate hundreds of references. On the other hand, in statistics, it is well understood that by using the likelihood ratio, when two distributions are multivariate normal, *equal* variance/covariance matrices will generate a Linear Discriminant Analysis (LDA) (or the Fisher analysis) and *unequal* variance/covariance matrices will generate a Quadratic Discriminant Analysis (QDA) [6]. Apparently the method of Fukunaga and Koontz belongs to the second case. Note that Fukunaga and Koontz transform (FKT) is not a QDA—it accomplish feature selection. In this paper, we show that the FKT, in an appropriate sense, is the ‘best’ low-rank approximation to the QDA. In applications of target recognition, it is important to realize a low rank approximate. Because each feature is obtained through a convolution of an imagery with a filter; the number of features determines the complexity of the algorithm.

In Section II, the method of Fukunaga and Koontz, which is also widely known as Tuned Basis Functions (TBF), is reviewed, together with its implementation in target recognition. In Section III, we argue that given the rank of the classifier, the method of TBF (or FKT) is the best low rank approximation. We give some concluding marks in Section IV.

## II. FUKUNAGA AND KOONTZ TRANSFORM & TUNED BASIS FUNCTIONS

In this section, we review the basic principle of the Fukunaga and Koontz transform, and an associated architecture.

Consider a library of target image chips,  $\{x_i, i = 1, \dots, T\}$  where each  $x_i$  is an image chip of size  $m \times n$  and contains a target. Consider a set of clutter chips,  $\{y_i, i = 1, \dots, C\}$ , where each  $y_i$  is an image chip of size  $m \times n$  and contains a clutter. We can re-order the pixels in each image chip into an  $mn \times 1$  vector and without ambiguity, let  $x_i, y_j$  denote these vectors. The structure of a detector can be viewed as projecting an image onto a set of basis vectors and accumulating the energy in the coefficient sequence

of the projection. The TBF is a systematic methodology to find such a basis set. It will separate targets from clutters. A description of the TBF are as follows.

Recall vectors  $\{x_i\}$  and  $\{y_i\}$  are target and clutter vectors. For simplicity, assume that the mean chip has been removed from each class. Note that due to this preprocessing step, we can make a zero-mean assumption for the rest of this paper. Let

$$\Sigma_1 = E[xx'] \text{ and } \Sigma_2 = E[yy'].$$

The sum of these matrices  $\Sigma_1 + \Sigma_2$  is positive semi-definite and can be factorized as follows:

$$\Sigma_1 + \Sigma_2 = \Phi D \Phi', \quad (1)$$

where  $\Phi$  is the matrix of eigenvectors of  $\Sigma_1 + \Sigma_2$  and  $D$  is a diagonal matrix with diagonal elements being equal to the eigenvalues. We can define a transformation operator  $P$  as

$$P = \Phi D^{-1/2}, \quad (2)$$

and new data vectors

$$\tilde{x} = P'x \text{ and } \tilde{y} = P'y.$$

The sum of the variance and covariance matrices for  $\tilde{x}$  and  $\tilde{y}$  becomes

$$P'(\Sigma_1 + \Sigma_2)P = I. \quad (3)$$

The covariance matrices for the transformed data  $\tilde{x}$  and  $\tilde{y}$  are  $T = P'\Sigma_1P$  and  $C = P'\Sigma_2P$  respectively. From equation (3) it is easy to see that  $T + C = I$ .

If  $\vec{\theta}$  is an eigenvector of  $T$  with corresponding eigenvalue  $\lambda$ , then it is also an eigenvector of  $C$  with eigenvalue  $(1 - \lambda)$ . This relationship guarantees that the covariance matrices of the transformed data will have the same eigenvectors; the sum of the two eigenvalues (of  $T$  and  $C$ ), which are associated with the same eigenvector, is equal to 1. Consequently, the dominant eigenvector of  $T$  is the weakest eigenvector of  $C$ , and vice versa. In the language of target detection, the dominant eigenvector of  $T$  contains maximal information about the target space, while containing the least information about the clutter space. Therefore, the first several dominant eigenvectors of  $T$  (as *target basis functions*) should be used to correlate an input image; and a high correlation coefficient suggests the presence of a target. Similarly, the weakest eigenvectors of  $T$  (as *anti-target basis functions*) should be correlated with an input image, and a high correlation reflects the presence of a clutter, or equivalently the absence of a target. The TBF detector utilizes both facts to create a detection algorithm.

In target recognition, one typically wants to use a small number of filters. Because adding one more filter means doing one more convolution with the image. In the matrix language, it is equivalent to say that one only wants to consider target recognition in a subspace. Note that the set of eigenvectors of the matrix  $T$ , which are also the eigenvectors of the matrix  $C$ , form an orthogonal basis in  $\mathcal{R}^{mn}$ . From previous analysis, the eigenvalue associated with a particular eigenvector yields a measure of the amount of target and/or clutter information that is described by that eigenvector. In TBF, only a small subset of dominant target and clutter basis functions are chosen. Specifically, one chooses the  $N_1$  basis functions that best represent targets and the  $N_2$  basis functions which best represent clutters. A matrix  $\Theta$  is defined as

$$\Theta = [\vec{\theta}_1, \dots, \vec{\theta}_{N_1}, \vec{\theta}_{mn-N_2+1}, \dots, \vec{\theta}_{mn}]. \quad (4)$$

Note that matrix  $\Theta$  is a  $mn$  by  $N_1 + N_2$  matrix, and it determines a  $(N_1 + N_2)$ -dimensional subspace in  $\mathcal{R}^{mn}$ .

A test image vector  $z$  ( $z \in \mathcal{R}^{mn}$ ) is projected onto this set, to obtain a feature vector  $v$ , which is of length  $N_1 + N_2$ , i.e.,  $v = \Theta'z = (v_1, v_2, \dots, v_{N_1+N_2})$ . The detection metric is defined as

$$\phi = \sum_{i=1}^{N_1} v_i^2 - \sum_{i=N_1+1}^{N_1+N_2} v_i^2. \quad (5)$$

The first summation on the right hand side of the above metric is the energy concentrated on *target*-like basis functions. The second summation is the energy projected on *clutter*-like basis functions. The metric is the difference between the two projected energies, and is expected to be large for targets and small for clutters.

### III. OPTIMAL QUADRATIC CLASSIFIERS

In the FKT framework, an important property is the dimension reduction. In this section, we show that FKT, under some assumption, is nearly the best classifier that one can do under the principle of likelihood ratio method and Gaussian distributional assumption.

In Section III-A, we start with a review of Bayes classifier and its application in target and clutter classification. We then prove the ‘best’ low rank approximation in Section III-B and III-C.

#### A. Bayes classifier on TBF transformed data

The Bayes classifier is the classifier which minimizes the probability of classification error, given the distribution of each population. Bayes classifier can be derived from likelihood ratio statistics.

Assuming that the populations are Gaussian with equal means, and that they have been normalized by  $(\Sigma_1 + \Sigma_2)^{-1/2}$ , so that their covariance matrices  $T$  and  $C$  satisfy  $T + C = I$ , the Bayes classifier takes the form  $S = \sum_i w_i v_i^2 \leq \alpha$ , where

- 1)  $S$  is the detector statistic, which is a weighted sum of squares of projection: for an image chip  $z$ ,

$$v = (\Sigma_1 + \Sigma_2)^{-1/2} z = (v_1, v_2, \dots, v_N)'$$

Here  $N$  is the dimension of the image chip (or an image).

- 2) The constant  $\alpha$  depends on the clutter/target prior probability, cost of missclassification, and on other factors.
- 3) The weights  $w_i$  depends on  $\lambda_i$ , which is the  $i$ -th largest eigenvalue of  $T$ :

$$w_i = \frac{(2\lambda_i - 1)}{\lambda_i(1 - \lambda_i)} = \frac{1}{1 - \lambda_i} - \frac{1}{\lambda_i}.$$

- 4) *The statistic  $S$  can be derived by considering the difference of the negative log-likelihoods of the two classes.* Due to space limitation, we have to leave out the details of this derivation. Basically, the difference of two negative log-likelihood functions of multivariate normal with mean zero and variance  $\Sigma_1$  and  $\Sigma_2$  is a quadratic function. The exact form of this quadratic function is computable.

Note that  $0 \leq \lambda_i \leq 1$ , so

- 1) for small  $\lambda_i \sim \epsilon$ , a weight  $w_i$  behaves like  $-1/\lambda_i$  ( $\sim -1/\epsilon$ );
- 2) for large  $\lambda_i \sim 1 - \epsilon$ , a weight  $w_i$  behaves like  $1/(1 - \lambda_i) \sim 1/\epsilon$ ;
- 3) for  $\lambda_i \sim 1/2$ , we have  $w_i \approx 0$ .

Let  $I_T$  denote the group of indices  $i$  for which  $\lambda_i \sim (1 - \epsilon)$  and  $I_C$  denote the group of indices for which  $\lambda_i \sim \epsilon$ . If for all the other indices, we have  $\lambda_i \sim 1/2$ , then

$$S \sim \frac{1}{\epsilon} * \left( \sum_{I_T} v_i^2 - \sum_{I_C} v_i^2 \right).$$

In short, the optimal detector statistic is approximately the TBF classifier.

Now, what happens if the  $\lambda_i$  are not exactly distributed the way that was just supposed? Will it still be true that the statistic

$$\Delta = \left( \sum_{I_T} v_i^2 - \sum_{I_C} v_i^2 \right) \tag{6}$$

is a low-rank approximation to the optimal detector? To prove a theoretical result, we introduce the following assumption.

*Condition 3.1 (Plateau condition):* Let  $n$  be the total number of eigenvalues  $\lambda_i$ 's. For an integer  $k$  ( $k < n$ ), the *Plateau condition* is satisfied if there exist at least  $k$   $\lambda_i$  that yield the maximal value among the set  $\{|\lambda_i - \frac{1}{2}|, i = 1, 2, \dots, n\}$ :

$$\lambda_{i_t} = \max_{1 \leq i \leq n} |\lambda_i - \frac{1}{2}|, \quad t = 1, 2, \dots, k.$$

We can show that when the above condition is satisfied, the TBF gives the low rank classifier that minimizes the classification error. This indicates that TBF is the optimal low rank approximate. Note that the above condition does not assume anything on the  $\lambda_i$ 's that are associated with relative small values of  $|\lambda_i - \frac{1}{2}|$ .

This condition is very restrictive. It is made so that we can compare an TBF approach, as in (5), with an optimal classifier.

### B. Low rank approximation to the optimal classifier

We start formulating our problem. Let us suppose that we have rotated our data by  $(\Sigma_1 + \Sigma_2)^{-1/2}$ , such that the two covariance matrices – target and clutter – are diagonal:  $\lambda_i$ 's are diagonal entries for targets, and  $(1 - \lambda_i)$ 's are entries for clutters. The discrimination problem can be formulated as a hypothesis testing problem as follows.

For random variables  $z_i \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$ ,

$$\begin{aligned} H_0 &: y_i = \sqrt{\lambda_i} z_i, \quad i = 1, 2, \dots, n, \quad \text{and} \\ H_1 &: y_i = \sqrt{1 - \lambda_i} z_i, \quad i = 1, 2, \dots, n, \end{aligned} \quad (7)$$

where  $0 \leq \lambda_i \leq 1, \forall i$ .

Suppose the optimal decision rule is in the form

$$D(Wy),$$

where  $W \in \mathcal{R}^{k \times n}$  is a rank  $k$  ( $k < n$ ) matrix,  $WW^T = I_k$  and

$$y = (y_1, y_2, \dots, y_n)^T.$$

We need to show that the function  $D(\cdot)$  is a quadratic function and that  $W$  is ‘nearly’ a diagonal matrix, which ‘picks’  $y_i$ 's that are associated with the largest- $k$  values of  $|\frac{1}{2} - \lambda_i|$ .

More specifically, let  $\lambda_{(j)}$  denote the  $\lambda_i$  that has the  $j$ -th largest value of  $|\frac{1}{2} - \lambda_i|$ . Let  $y_{(j)}$  denote the  $y_i$  that is associated with  $\lambda_{(j)}$  (as in (7)). We show that the optimal decision rule is

$$D(Wy) = \sum_{i=1}^k \left( \frac{1}{\lambda_{(i)}} - \frac{1}{1 - \lambda_{(i)}} \right) y_{(i)}^2. \quad (8)$$

Note that the above is similar to the decision rule when matrix  $W$  is allowed to have full rank; the difference is that the number of terms is reduced. Also note that by taking into account the fact that we have either  $\lambda_{(i)} = \epsilon \sim 0$  or  $\lambda_{(i)} = 1 - \epsilon \sim 1$ , following a similar argument, the (8) is numerically close to the  $\Delta$  in (6).

### C. Proof of the low rank approximation

Under the Plateau condition, we have the following.

*Theorem 3.2:* When the plateau condition is satisfied, the rank  $k$  decision takes has the form in (8), and  $\lambda_i$ 's take maximal values in the set  $\{|\lambda_i - \frac{1}{2}|, i = 1, 2, \dots, n\}$ .

We prove the above result step by step. There are four steps. We first specify the optimality condition. Then the problem is rewritten in multivariate analysis, and the new eigenvalues are analyzed. We prove a greedy incremental result for the objective function. The final proof is based on utilizing all the above.

1) *Optimality condition:* The optimal decision rule is the one that maximizes the value of the following objective function:

$$\begin{aligned} & \max_{D(\cdot), W, t} && P_{H_1}\{D(Wy) > t\} \\ \text{subject to} &&& \text{rank}(W) = k, \quad WW^T = I_k, \\ &&& P_{H_0}\{D(Wy) > t\} \leq \alpha. \end{aligned}$$

Using the idea of Lagrangian multiplier, the above is equivalent to the following,

$$\begin{aligned} & \min_{D(\cdot), W, t} && P_{H_1}\{D(Wy) < t\} + c_1 \cdot P_{H_0}\{D(Wy) > t\} \\ \text{subject to} &&& \text{rank}(W) = k, \quad WW^T = I_k, \end{aligned} \quad (9)$$

where appropriate value of  $c_1$  will render the exact solution to the previous optimization problem. In the rest of this note, we consider the second optimality condition.

2) *Rewritten in vectors:* The original problem in (7) is equivalent to the following:

$$\begin{aligned} H_0 & : y \sim mN(\vec{0}, D), \quad \text{and} \\ H_1 & : y \sim mN(\vec{0}, I_n - D), \end{aligned}$$

where

$$D = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix},$$

$y$  is a  $n$ -dimensional random vector,  $\vec{0}$  is a zero vector of the same dimension, and  $mN$  stands for multivariate normal.

For  $Wy$ , we have

$$\begin{aligned} H_0 & : Wy \sim mN(\vec{0}, WDW^T), \quad \text{and} \\ H_1 & : Wy \sim mN(\vec{0}, W(I_n - D)W^T). \end{aligned}$$

The matrix  $WDW^T$  has to be semi-definite. Consider its Jordan decomposition

$$WDW^T = O\tilde{D}O^T,$$

where  $O \in \mathcal{R}^{k \times k}$ ,  $OO^T = I_k$ . Apparently, we have

$$WDW^T + W(I_n - D)W^T = I_k.$$

Hence

$$W(I_n - D)W^T = O(I_k - \tilde{D})O^T.$$

Note that an orthogonal matrix will not change the density function of a multivariate normal distribution. From all the above, for  $\tilde{y} = Wy$ , it is equivalent to consider the following hypothesis testing problem:

$$\begin{aligned} H_0 & : \tilde{y} \sim mN(\vec{0}, \tilde{D}), \quad \text{and} \\ H_1 & : \tilde{y} \sim mN(\vec{0}, I_k - \tilde{D}), \end{aligned} \tag{10}$$

where

$$\tilde{D} = \begin{pmatrix} \tilde{\lambda}_1 & & & \\ & \tilde{\lambda}_2 & & \\ & & \ddots & \\ & & & \tilde{\lambda}_k \end{pmatrix}.$$

Since

$$\tilde{D} = O^T WDW^T O.$$

We have

$$\tilde{\lambda}_i = \sum_{j=1}^n w_{ij}^2 \lambda_j, \quad i = 1, 2, \dots, k,$$

where  $(w_{i1}, w_{i2}, \dots, w_{in})$  form the  $i$ th row of the matrix  $O^T W$ . Recall that

$$\sum_{j=1}^n w_{ij}^2 = 1, \quad i = 1, 2, \dots, k,$$

and

$$\sum_{i=1}^k w_{ij}^2 \leq 1, j = 1, 2, \dots, n.$$

Suppose  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ , we can easily prove the following,

$$\tilde{\lambda}_1 \leq \lambda_1, \quad \tilde{\lambda}_1 + \tilde{\lambda}_2 \leq \lambda_1 + \lambda_2, \quad \dots,$$

and

$$\tilde{\lambda}_k \geq \lambda_n, \quad \tilde{\lambda}_k + \tilde{\lambda}_{k-1} \geq \lambda_n + \lambda_{n-1}, \quad \dots.$$

3) *Optimal decision rule:* To minimize the objective function in (9), following a Neymann-Pearson type of argument (considering the log of the likelihood ratio), it is easy to see that the decision rule must be

$$D(Wy) = \sum_{j=1}^k \left( \frac{1}{\tilde{\lambda}_j} - \frac{1}{1 - \tilde{\lambda}_j} \right) \tilde{y}_j^2 \leq \text{a threshold.}$$

Now we decompose the objective function

$$\begin{aligned} & P_{H_1}\{D(Wy) < t\} + c_1 \cdot P_{H_0}\{D(Wy) > t\} \\ &= \int f_1 I\{D(Wy) < t\} + c_1 f_2 I\{D(Wy) > t\} dy \\ &= \int \left[ I\{D(Wy) < t\} \prod_{j=1}^k \frac{1}{\sqrt{1 - \tilde{\lambda}_j}} \phi\left(\frac{y_j}{\sqrt{1 - \tilde{\lambda}_j}}\right) \right. \\ &\quad \left. + c_1 I\{D(Wy) > t\} \prod_{j=1}^k \frac{1}{\sqrt{\tilde{\lambda}_j}} \phi\left(\frac{y_j}{\sqrt{\tilde{\lambda}_j}}\right) \right] d\tilde{y}_1 \cdots d\tilde{y}_k, \end{aligned}$$

where  $f_1$  and  $f_2$  are the density functions under  $H_1$  and  $H_0$  respectively,  $I\{\cdot\}$  is an indicator function, and  $\phi(\cdot)$  is the probability density of the standard normal. When  $k = 1$ , one can see that the above becomes comparing two normal distributions with common mean zero and different variances. It is easy to verify that the larger the value of  $|\frac{1}{2} - \lambda|$  is, the smaller the value of the following function

$$(*) = \int \left[ I\left\{\left(\frac{1}{\lambda} - \frac{1}{1 - \lambda}\right)y^2 > \tilde{t}\right\} \frac{1}{\sqrt{\lambda}} \phi\left(\frac{y}{\sqrt{\lambda}}\right) + I\left\{\left(\frac{1}{\lambda} - \frac{1}{1 - \lambda}\right)y^2 < \tilde{t}\right\} \frac{1}{\sqrt{1 - \lambda}} \phi\left(\frac{y}{\sqrt{1 - \lambda}}\right) \right] dy$$

will be. This can be shown by the following, when  $\lambda < 1/2$ ,

$$\begin{aligned} (*) &= \int I\left(\frac{1 - 2\lambda}{\lambda(1 - \lambda)}y^2 > \tilde{t}\right) \frac{1}{\sqrt{\lambda}} \phi\left(\frac{y}{\sqrt{\lambda}}\right) + I\left(\frac{1 - 2\lambda}{\lambda(1 - \lambda)}y^2 < \tilde{t}\right) \frac{1}{\sqrt{1 - \lambda}} \phi\left(\frac{y}{\sqrt{1 - \lambda}}\right) \\ &= \int I\left(x^2 > \frac{(1 - \lambda)\tilde{t}}{1 - 2\lambda}\right) \phi(x) + I\left(x^2 < \frac{\lambda\tilde{t}}{1 - 2\lambda}\right) \phi(x) \\ &= 1 - \int I\left(\frac{\lambda\tilde{t}}{1 - 2\lambda} < x^2 < \frac{(1 - \lambda)\tilde{t}}{1 - 2\lambda}\right) \phi(x). \end{aligned}$$

When  $\lambda \rightarrow 0$ ,  $\lambda/(1 - 2\lambda)$  decreases to 0. In the mean time, the difference  $\frac{(1-\lambda)\tilde{t}}{1-2\lambda} - \frac{\lambda\tilde{t}}{1-2\lambda}$  is  $\tilde{t}$ . Hence the value of (\*) decreases as  $\lambda$  decreases. For  $\lambda > 1/2$ , similar result can be drawn.

This analysis demonstrates that by substituting  $\tilde{\lambda}_j$  with a  $\lambda_i$  that has larger deviation from  $\frac{1}{2}$  (i.e. larger  $|\frac{1}{2} - \lambda|$ ), the value of the objective function in (9) is reduced. This shows that the minimum can only be achieved when  $\tilde{\lambda}_j, j = 1, 2, \dots, k$  are associated with the largest  $k$  values of  $|\frac{1}{2} - \lambda_i|, i = 1, 2, \dots, n$ .

4) *Final argument:* From all the above, we proved the Theorem 3.2.

The Plateau condition seems to be a very strong assumption. It is hard to obtain a more generic result. The difficulty in obtaining a generic result is to study the interplay between values of different  $\lambda_i$ 's to the value of the objective function in (9).

#### IV. CONCLUSION

We prove that under certain conditions, the Fukunaga Koontz transform is the optimal low rank approximation to an optimal classifier. This gives an interesting statistical interpretation on a widely-used pattern recognition method.

The current result can be generalized to multi-class classifier. We will leave details for future publication.

#### ACKNOWLEDGMENT

Professor David Donoho first formulates the problem as a low rank approximation problem. The author would also like to thank Abhijit Mahalanobis, Bob Muise, and Robert Stanfill for introducing us the Fukunaga Koontz transform. The comments from the associate editor and two anonymous referees helped improving the presentation of the paper.

#### REFERENCES

- [1] T. Kailath and H.L. Weinert (1975). An RKHS (reproducing kernel Hilbert space) approach to detection and estimation problems. II. Gaussian signal detection. *IEEE Transactions on Information Theory*, IT-21 (1), 15-23, January.
- [2] L. A. Shepp (1966). Radon-Nikodým derivatives of Gaussian measures. *Ann. Math. Statist*, 37, 321-354.
- [3] G. Kallianpur and H. Oodaira (1963). The equivalence and singularity of Gaussian measures. *Proc. Sympos. Time Series Analysis (Brown Univ., 1962)*, 279-291, Wiley, New York.
- [4] M.-H. Yang, D.J. Kriegman, and N. Ahuja (2002). Detecting faces in Images: A Survey. *IEEE Trans on Pattern Anal. and Machine Intell.*, 24 (1), January.
- [5] F. Fukunaga and W. Koontz (1970). Applications of the Karhunen-Loève Expansion to Feature Selection and Ordering. *IEEE Trans. Computers*, 19 (5), 311-318.
- [6] E.L. Lehmann (1986). *Testing statistical hypotheses*. Wiley, New York.