Optimal Reduced-Rank Quadratic Classifiers Using the Fukunaga-Koontz Transform, with Applications to Automated Target Recognition

Xiaoming Huo[2], Michael Elad[1], Ana Georgina Flesia[1],
Bob Muise[3], Robert Stanfill[3], Jerome Friedman[1], Bogdan Popescu[1],
Jihong Chen[2], Abhijit Mahalanobis[3], David L. Donoho[1],
[1] Stanford University, Stanford, CA 94305,
[2] Georgia Institute of Technology, Atlanta, GA 30332,
[3] Lockheed Martin Co., Orlando, FL.

March 24, 2003

Abstract

In target recognition applications of discriminant or classification analysis, each 'feature' is a result of a convolution of an imagery with a filter, which may be derived from a feature vector. It is important to use relatively few features.

We analyze an optimal reduced-rank classifier under the two-class situation. Assuming each population is Gaussian and has zero mean, and the classes differ through the covariance matrices: Σ_1 and Σ_2. The following matrix is considered:

$$\Lambda = (\Sigma_1 + \Sigma_2)^{-1/2}\Sigma_1(\Sigma_1 + \Sigma_2)^{-1/2}.$$

We show that the k eigenvectors of this matrix whose eigenvalues are most different from $1/2$ offer the best rank k approximation to the maximum likelihood classifier. The matrix Λ and its eigenvectors have been introduced by Fukunaga and Koontz; hence this analysis gives a new interpretation of the well known Fukunaga-Koontz transform.

The optimality that is promised in this method hold if the two populations are exactly Gaussian with the same means. To check the applicability of this approach to real data, an experiment is performed, in which several 'modern' classifiers were used on an Infrared ATR data. In these experiments, a reduced-rank classifier–Tuned Basis Functions–outperforms others.

The competitive performance of the optimal reduced-rank quadratic classifier suggests that, at least for classification purposes, the imagery data behaves in a nearly-Gaussian fashion.

Acknowledgement

This work was sponsored by the US Army Research Office. Financial support was received by DARPA under contract DAAD19-02-C-0025. The contents of this paper do not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

1 Introduction

Since Fukunaga and Koontz published their paper [4] on feature selection in 1970, many researchers have used their method in various applications, such as target recognition, face detection [10], 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 the two distributions are multivariate normal, *equal* variance and covariance matrices lead to Linear Discriminant Analysis (LDA) (or Fisher analysis) and *unequal* variance and covariance matrices lead to Quadratic Discriminant Analysis (QDA) [7]. Apparently the method of Fukunaga and Koontz belongs to the second case. Note that Fukunaga and Koontz transform (FKT) is not a QDA — it chooses a subset of features. In this paper, we show that the FKT, in an appropriate sense, is the 'best' low-rank approximate to QDA. Moreover, in the application of target recognition, it is important to realize a low rank approximate. Because each feature is obtained through a convolution of the imagery with a filter; the number of the features determines the complexity of the algorithm.

In Section 2, the method of Fukunaga and Koontz, which is also called Tuned Basis Functions (TBF), is reviewed, together with its implementation in a target recognition scenario. In Section 3, we argue that given the rank of the classifier, the method of TBF (or FKT) is the most optimal low rank approximate. In Section 4, an experiment on infra-red image data is reported, in which several other machine learning methods (e.g. maximal rejection, LDA, QDA, support vector machine, multiple additive regression trees) are used to compare with the FKT approach. It is found that FKT gives the best overall performance, which indicates that the transformed infrared ATR imagery data behaves in a nearly-Gaussian fashion. We give some concluding marks in Section 5.

2 Fukunaga and Koontz Transform & Tuned Basis Functions

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

Consider a library of target image chips, $\{x_i, i = 1, \ldots, T\}$ where each x_i is an image chip of size $m \times n$, containing a target in the library. Similarly, we assume the existence of a set of clutter chips, $\{y_i, i = 1, \ldots, C\}$, where each y_i is an image chip of size $m \times n$, containing an example of clutter. We can re-order the pixels in each image chip into an $mn \times 1$ vector and without ambiguity, represent this vector by the same variable as the image.

The structure of the detector can be viewed as projecting the image onto a set of basis vectors and accumulating the energy in the coefficient sequence of the projection. There are many different orthogonal basis sets that may be utilized to generate the filters in Figure 1. The TBF is a systematic methodology to find such a basis set, which separates targets from clutter. The details of the TBF are as follows.

Let the vectors $\{x_i\}$ and $\{y_i\}$ represent target and clutter vectors. For simplicity, assume that the mean chip has been removed from each class. Let

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

The sum of these matrices $\Sigma_1 + \Sigma_2$ is positive definite and can be factorized in the form

$$\Sigma_1 + \Sigma_2 = \Phi D \Phi'. \tag{1}$$

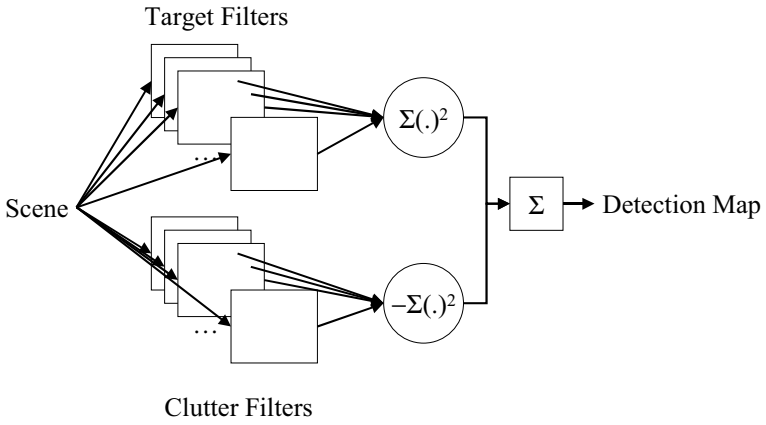

Figure 1: Architecture of a TBF detector.

where Φ 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

$$P = \Phi D^{-1/2}, \tag{2}$$

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

$$P'(\Sigma_1 + \Sigma_2)P = I. \tag{3}$$

The covariance matrices for the transformed data $\tilde{x}$ and $\tilde{y}$ are $T_{=}P'\Sigma_1 P$ and $C_{=}P'\Sigma_2 P$ respectively. From (2) it is easy to show that

$$T + C = I.$$

It is easy to verify that if $\vec{\theta}$ is an eigenvector of T with corresponding eigenvalue λ, then it is also an eigenvector of C but with eigenvalue $(1 - \lambda)$. This relationship guarantees that the covariance matrices of the transformed data will have the same eigenvectors. It should be noted that the the sum of the corresponding eigenvalues of T and C 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 (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 (Anti-target basis functions) should be correlated with an input image, and a high correlation reflects the presence of clutter, or equivalently the absence of a target. The TBF detector utilizes both facts to create a detection algorithm, which is tuned to the available training samples. To give readers a sense of eigenvalues in real applications, in Figure 2, eigenvalues for some images chips with targets and clutters are plotted.

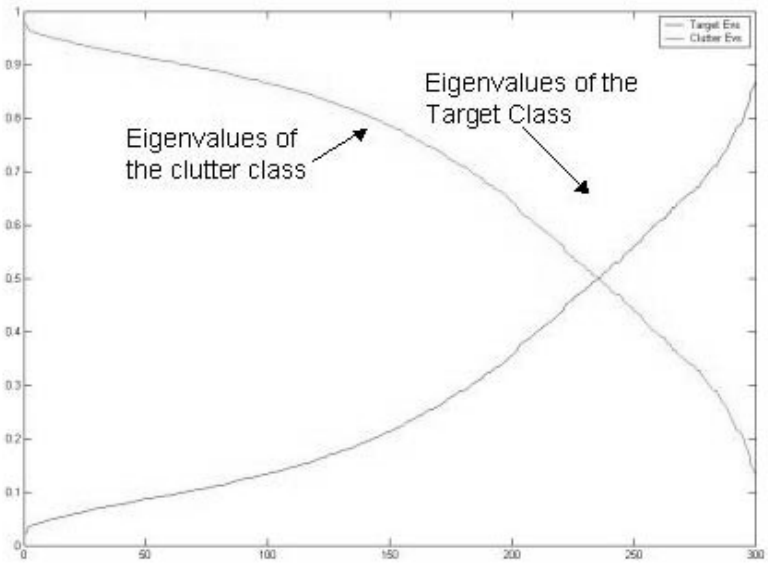

Figure 2: Eigenvalues for targets and clutters.

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 we only want to consider the 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 for $\mathcal{R}^{mn}$. From a 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 the TBF formulation, 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 Θ is defined as

$$\Theta = [\vec{\theta}_1, \ldots, \vec{\theta}_{N_1}, \vec{\theta}_{mn-N_2+1}, \ldots, \vec{\theta}_{mn}]. \tag{4}$$

Note that matrix Θ is an 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, $v = \Theta' z = (v_1, v_2, \ldots, v_{N_1+N_2})$. The detection metric is defined as

$$\phi = \sum_{i=N_1+1}^{N1+N_2} v_i^2 - \sum_{i=1}^{N_1} v_i^2, \tag{5}$$

which also is the output of the diagram that is depicted in Figure 3.

The first summation on the right hand side of the above metric is the net energy in the projections of the test image on the *target*-like basis function. The second summation is the net energy projected on the *clutter*-like basis functions. The metric is thus the difference in the two projected energies, and is expected to be large for targets and small for clutter.

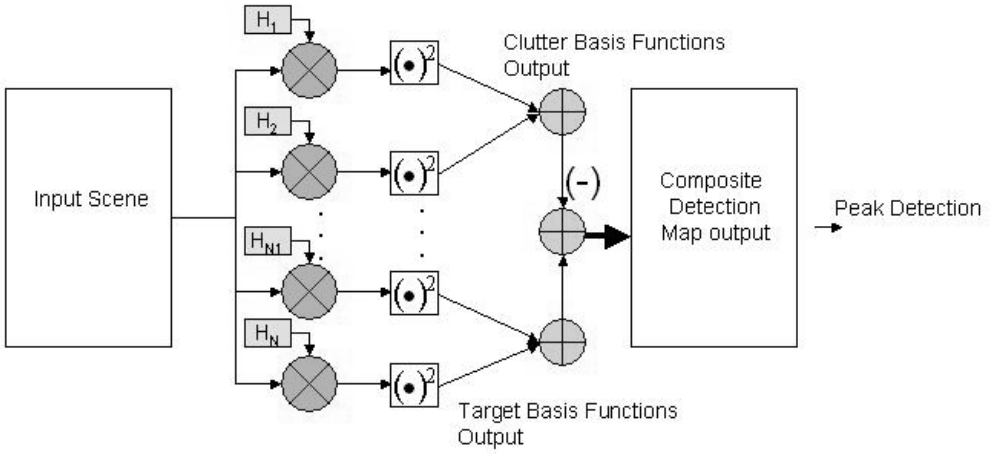

Figure 3: Another illustration of a TBF detector.

3 Optimal quadratic classifiers

In the FKT framework, an important property is that it realizes 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 a Gaussian distributional assumption.

In Section 3.1, 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 Sections 3.2 and 3.3.

3.1 Bayes classifier on TBF transformed data

A classical result states that the Bayes classifier is the classifier which minimizes the probability of classification error, assuming that the distribution under each population is given. Normally, the Bayes classifier can be derived from likelihood ratio statistics.

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

$$S = \sum_i w_i v_i^2 \lessgtr \alpha,$$

where

1. S is the detector statistic, which is a weighted sum of squares of projection: for an image chip z,
$$v = (\Sigma_1 + \Sigma_2)^{-1/2} z = (v_1, v_2, \ldots, v_N)'.$$
Here N is the dimension of an image chip (or an image).

2. The constant α depends on the clutter/target prior probability, cost of misclassification, as well as other factors.

3. The weights w_i depends on λ_i =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.

For detailed derivation of the above results, we refer to Appendix A.

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

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

Hence, if it happens that there is a group I_T of indices where $\lambda_i \sim (1-\epsilon)$ and a group I_C of where $\lambda_i \sim \epsilon$ and if all the other indices have $\lambda_i \sim 1/2$, then

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

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

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

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

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

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

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 λ_i's that are associated with relative small values of $|\lambda_i - \frac{1}{2}|$.

3.2 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 covariance matrix of our new data is diagonal, with λ_i diagonal entries on the target class and $(1-\lambda_i)$ entries on the clutter class. Then we may see our discrimination problem as a hypothesis testing problem as follows.

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

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

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

Suppose the optimal decision rule is of the form

$$D(Wy),$$

where W is a rank $k (k < n)$ matrix, $WW^T = I_k$ and

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

We need to show that function $D(\cdot)$ is a quadratic function and 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 λ_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

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

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 Δ in (6).

3.3 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 λ_i's takes maximal values in the set $\{|\lambda_i - \frac{1}{2}|, i = 1, 2, \ldots, n\}$.*

We prove the result step by step. There are four steps in the proof. 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 all the above, and is given in the last step.

3.3.1 Optimality condition

The optimal decision rule is the one that will maximize the value of the following objective function:

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

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

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

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 latter optimality condition.

3.3.2 Rewritten in vectors

The original problem in (7) is equivalent to the following:

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

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

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

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:

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

where

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

Since

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

We have

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

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

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

and

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

Suppose

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n,$$

we can easily prove the following,

$$\begin{aligned} \tilde{\lambda}_1 &\leq \lambda_1, \\ \tilde{\lambda}_1 + \tilde{\lambda}_2 &\leq \lambda_1 + \lambda_2, \\ &\vdots \end{aligned}$$

and

$$\begin{aligned} \tilde{\lambda}_k &\geq \lambda_n, \\ \tilde{\lambda}_k + \tilde{\lambda}_{k-1} &\geq \lambda_n + \lambda_{n-1}, \\ &\vdots \end{aligned}$$

3.3.3 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

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

Now we decompose the objective function

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

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 reduces to comparing two normal distributions with zero means 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

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

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 $\lambda \to 0$, $\lambda/(1-2\lambda)$ 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 λ decreases. For $\lambda > 1/2$, a similar result can be drawn.

This analysis demonstrates that by substituting $\tilde{\lambda}_j$ with a λ_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 $\tilde{\lambda}_j, j = 1, 2, \ldots, k$ are associated with the largest k values of $|\frac{1}{2} - \lambda_i|, i = 1, 2, \ldots, n$.

3.3.4 Final argument

From all the above, we proved 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 λ_i's to the value of the objective function in (9).

4 Simulation

For an infra-red (IR) image dataset, we compare the TBF with six other classifiers. They are

1. **(SVM)** *Support Vector Machine* [1]. SVM based methods recently have gained significant popularity. It is a maximal margin based classification method. Its derivation is significantly different from other methods that are derived in statistics, e.g. Fisher quadratic classifier. SVM can be formulated as a Quadratic Programming problem. Hence it can be solved efficiently.

2. **(MRC)** *Maximal Rejection* classifier [2]. MRC builds a hierarchy of classifiers (rejectors). At each stage, a proportion of input data is classified into one class, by utilizing a simple classification rule. The remaining inputs are transferred to the next classifier. The advantage of this approach is the speed and performance in many application domain, such as face detection.

3. **(FQC)** *Fisher Quadratic Classifier*. FQC is a classical method, which is rooted in statistics. This has become standard material in many statistical textbook. The basic idea is to utilize a Gaussian model and to consider the likelihood ratio. The result is a classifier, which depends on a quadratic function of the input.

4. **(MART)** *Multiple Additive Regression Trees* [3]. We consider two implementational strategies:

   (a) a MART with 50 features **(MART-**50**)**, and

   (b) a MART with 25 features **(MART-**25**)**.

   MART is an additive tree model. It takes advantages of both the tree models and the generality of additive model. In spirit, it is close to a recently emerged powerful technique: *boosting*. Its supreme performance has been observed in many applications.

5. **(k-NN)** *k-Nearest Neighbor*. k-NN is a classical non-parametric statistical method. For each input x_0, it considers the inputs that fall in a neighborhood of x_0. It is a standard benchmark.

The pairs of error rates—the false alarm rate and the mis-detection rate—for various classifiers are provided in Figure 4. It can be seen that the TBF classifier in general, gives the best overall performance. This shows that in the applications of IR based target recognition, the transformed image chips are nearly Gaussian distributed, and the TBF method catches it well.

5 Conclusion

The problem of low-rank approximation to a statistically optimal classifier is studied. We present a condition, under which the well-known Fukunaga-Koontz transform (also known as Tuned Basis Function) can be proven to be the *best* approximation. This condition is still restrictive, but enough to provide some conceptual insight. In an IR database, we compare our TBF with several existing dominant techniques. It is found that in the IR database, the TBF outperforms nearly all other methods. This shows that in ATR practice, a relatively cheap TBF is sufficient. Of course, this conclusion is limited by the data that we are working on.

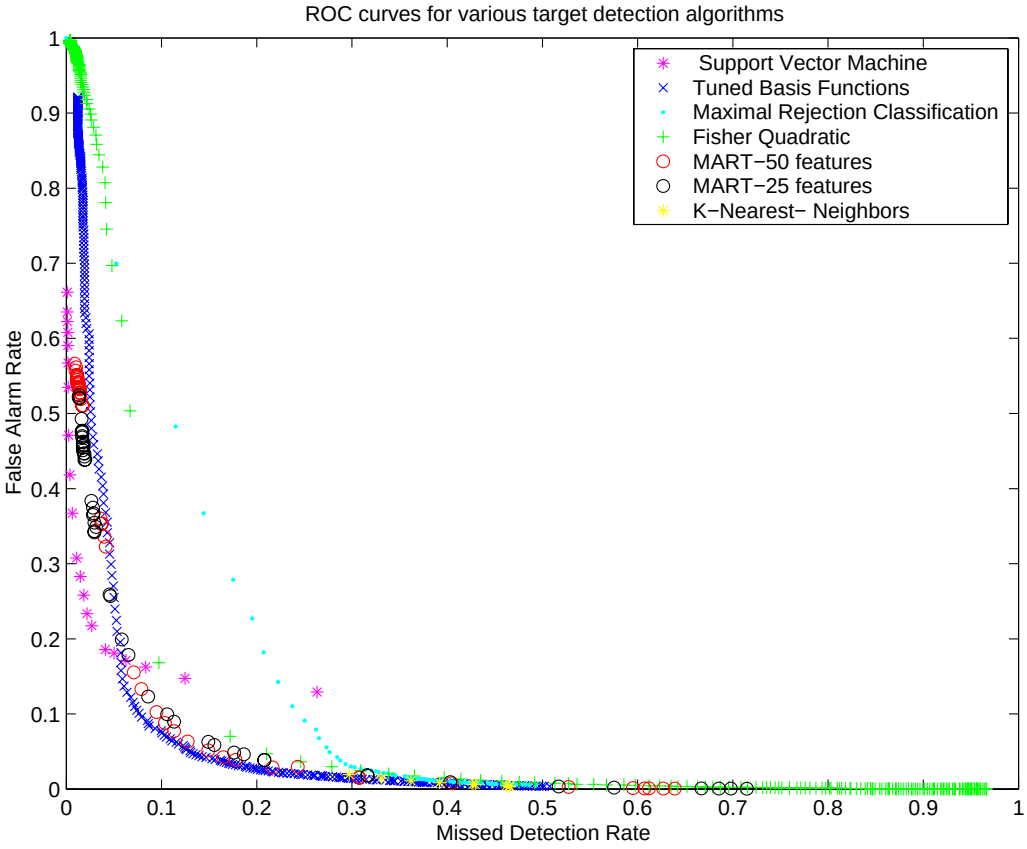

Figure 4: Comparison of various classifiers for IR target/clutter imageries.

A Optimal Quadratic Classifier – Gaussian Case

We now develop the Maximum-Likelihood (ML) classifier which minimizes the classification error. We assume hereafter that both the *target* and *clutter* sets are drawn from the Gaussian distributions:

$$\begin{aligned} P_T(\underline{X}) &= \left[\frac{1}{(2\pi)\cdot det\{\Sigma_T\}}\right]^{n/2} \exp\left\{-(\underline{X}-\underline{M}_T)^T\Sigma_T^{-1}(\underline{X}-\underline{M}_T)\right\}, \quad \text{and} \\ P_C(\underline{X}) &= \left[\frac{1}{(2\pi)\cdot det\{\Sigma_C\}}\right]^{n/2} \exp\left\{-(\underline{X}-\underline{M}_C)^T\Sigma_C^{-1}(\underline{X}-\underline{M}_C)\right\}. \end{aligned}$$

The optimal ML classifier is given by

$$\begin{aligned} f(\underline{X}) &= \frac{P_T(\underline{X})}{P_C(\underline{X})} \\ &= C\cdot\exp\left\{-(\underline{X}-\underline{M}_T)^T\Sigma_T^{-1}(\underline{X}-\underline{M}_T)+(\underline{X}-\underline{M}_C)^T\Sigma_C^{-1}(\underline{X}-\underline{M}_C)\right\} \\ &\lessgtr \frac{p_c}{p_t}, \end{aligned} \tag{11}$$

where p_t and p_c are the classes a-priori probabilities. Using the decomposition proposed earlier, we can pre-multiply the vectors $(\underline{X}-\underline{M}_T)$ and $(\underline{X}-\underline{M}_C)$by $\mathbf{P}^T$. Define $\underline{Z}_T = \mathbf{P}^T(\underline{X}-\underline{M}_T)$ and $\underline{Z}_C = \mathbf{P}^T(\underline{X}-\underline{M}_C)$. Thus we get

$$\begin{aligned} f(\underline{X}) &= C\cdot\exp\left\{-\underline{Z}_T^T\mathbf{P}^{-1}\Sigma_T^{-1}\mathbf{P}^{-T}\underline{Z}_T+\underline{Z}_C^T\mathbf{P}^{-1}\Sigma_C^{-1}\mathbf{P}^{-T}\underline{Z}_C\right\} \\ &= C\cdot\exp\left\{-\underline{Z}_T^T\mathbf{T}^{-1}\underline{Z}_T+\underline{Z}_C^T\mathbf{C}^{-1}\underline{Z}_C\right\}. \end{aligned}$$

If we now further assume that the means of the two classes is the same (i.e., $\underline{M}_T = \underline{M}_C$), we have that $\underline{Z}_T = \underline{Z}_C = \underline{Z}$ and thus the ML decision rule becomes

$$\exp\left\{-\underline{Z}^T(\mathbf{T}^{-1}-\mathbf{C}^{-1})\underline{Z}\right\} \lessgtr \text{threshold}.$$

Now, recall that we know that $\mathbf{T}+\mathbf{C}=\mathbf{I}$, and we know that these pair of matrices share the same eigenvectors, implying that they are jointly diagonalizable. Assume that

$$\mathbf{T} = \Theta^T\mathbf{D}\Theta, \quad \text{and} \quad \mathbf{C} = \Theta^T(\mathbf{I}-\mathbf{D})\Theta,$$

where $\mathbf{D}$ is a diagonal matrix with real values λ_k on the main diagonal in the range $[0,1]$. Putting these relationships into the ML decision rule we have

$$\exp\left\{-(\Theta^{-T}\underline{Z})^T\left[\mathbf{D}^{-1}-(\mathbf{I}-\mathbf{D})^{-1}\right](\Theta^{-T}\underline{Z})\right\} \lessgtr \text{threshold}.$$

Defining $\underline{W} = \Theta^{-T}\underline{Z}$ we have

$$\exp\left\{-\underline{W}^T\left[\mathbf{D}^{-1}-(\mathbf{I}-\mathbf{D})^{-1}\right]\underline{W}\right\} \lessgtr \text{threshold}.$$

Or better yet

$$-\underline{W}^T\left[\mathbf{D}^{-1}-(\mathbf{I}-\mathbf{D})^{-1}\right]\underline{W} = \sum_{k=1}^{n}\left(\frac{1}{1-\lambda_k}-\frac{1}{\lambda_k}\right)w_k^2 \lessgtr \text{threshold}.$$

So we see that by taking the incoming vector $\underline{X}$, removing the mean $\underline{M}_T$ (or $\underline{M}_C$, which is assumed to be the same), and linearly transforming the obtained vector by the matrix $\Theta^{-T}\mathbf{P}^T$, we obtain a simple elementwise quadratic decision rule.

References

[1] N. Cristianini and J. Shawe-Taylor (2000). *Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, New York.

[2] M. Elad, Y. Hel-Or, and R. Keshet (2002). Pattern Detection Using a Maximal Rejection Classifier. *Pattern Recognition Letters,* 23(12), 1459-1471, October.

[3] J. H. Friedman (2001). Greedy function approximation: a gradient boosting machine. *Ann. Statist.* 29 (5), 1189–1232.

[4] F. Fukunaga and W. Koontz (1970). Applications of the Karhunen-Loève Expansion to Feature Selection and Ordering. *IEEE Trans. Computers*, 19 (5), 311-318.

[5] 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.

[6] 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.

[7] E.L. Lehmann (1986). *Testing statistical hypotheses*. Wiley, New York.

[8] A. Mahalanobis, R.R. Muise, R.S. Stanfill, and A.V. Nevel (2002). Design and application of quadratic correlation filters for target detection. Submitted.

[9] L. A. Shepp (1966). Radon-Nikodým derivatives of Gaussian measures. *Ann. Math. Statist*, 37, 321-354.

[10] 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.