Smoothing Parameters Selection for Dimensionality Reduction Method based on Probabilistic Distance
Application to Handwritten Recognition

Faycel El Ayeb and Faouzi Ghorbel
GRIFT Research Group, CRISTAL Laboratory, Ecole Nationale des Sciences de l'Informatique (ENSI),
La Manouba University, 2010 Manouba, Tunisia

Keywords: Dimensionality Reduction, Feature Extraction, Fisher Criterion, Orthogonal Density Estimation, Patrick-Fisher Distance, Smoothing Parameter, Handwritten Digits Classification, Invariant Descriptors.

Abstract: Here, we intend to give a rule for the choice of the smoothing parameter of the orthogonal estimate of Patrick-Fisher distance in the sense of the Mean Integrate Square Error. The orthogonal series density estimate precision depends strongly on the choice of such parameter which corresponds to the number of terms in the series expansion used. By using series of random simulations, we illustrate the better performance of its dimensionality reduction in the mean of the misclassification rate. We show also its better behavior for real data. Different invariant shape descriptors describing handwritten digits are extracted from a large database. It serves to compare the proposed adjusted Patrick-Fisher distance estimator with a conventional feature selection method in the mean of the probability error of classification.

1 INTRODUCTION

It is well known that the feature extraction obtained by dimensional reduction algorithms is very important task for pattern recognition. Different applications in this field as face analysis, handwritten character recognition, 3D-medical image segmentation have investigated such approach. The famous Linear Discriminate Analysis (*LDA*) method which optimizes the Fisher ratio (Fukunaga, 1990) is very used in practice for reducing dimensionality. However, when one of the conditional probability density functions (*PDFs*) relative to labels follows a non Gaussian distribution, the *LDA* gives generally bad and non stable result (Ghorbel et al., 2012). It has been proved that the maximization of an estimate of the Patrick-Fisher distance (Drira and Ghorbel, 2012) (Aladjem, 1997) (Patrick and Fisher, 1969) (Hillion et al., 1988) could be improves the result because it considers the hole of the statistical information about the conditional observation. However the *LDA* method is generally expressed only according to the first and second statistical moments of *PDFs*. In the present work, we investigate a dimensionality reduction method introduced in (Ghorbel, 2011). It consists on an estimator of the Patrick-Fisher distance ($\hat{d}_{PF}$) using the conventional orthogonal series density estimator.

Among the non parametric density estimation method, the *PDFs* could be approximated by an orthogonal series expansion (Cencov and Nauk, 1962). The performance and smoothness of the orthogonal series density estimate depend strongly on the optimal choice of the parameter k_N which corresponds to the number of terms in the series expansion used. Rather than arbitrarily choosing the value of k_N for each class to estimate the Patrick-Fisher distance (d_{PF}), we propose to select the number of terms that minimize the mean integrated square error (*MISE*) of the $\hat{d}_{PF}$. The remaining of the paper is organized as follows. In section 2, we review the orthogonal series density estimator. The *LDA* method and the one based on the $\hat{d}_{PF}$ are recalled in Section 3. After that, we propose a novel rule for selecting the optimal values of k_N for $\hat{d}_{PF}$. Experimental results both on simulated data and on handwritten digits database are given in Section 4. Section 5 gives a conclusion of the paper.

2 ORTHOGONAL SERIES DENSITY ESTIMATION

In the following, we just recall the orthogonal series density estimation method by presenting its essen-

El Ayeb F. and Ghorbel F. (2013).
Smoothing Parameters Selection for Dimensionality Reduction Method based on Probabilistic Distance - Application to Handwritten Recognition.
In *Proceedings of the 2nd International Conference on Pattern Recognition Applications and Methods*, pages 325-330
DOI: 10.5220/0004333503250330

tial convergence studies detailed in (Beauville, 1978). The orthogonal series estimator of the *PDF* of a given sample X_i assumed to follow the same distribution could be obtained by the following limited Fourier series expansion:

$$\hat{f}_{k_N}(x) = \sum_{m=0}^{k_N} \hat{a}_{m,N} e_m(x) \tag{1}$$

Where $\{e_m(x)\}$ is a complete and orthogonal basis of functions. Here k_N is called the truncation value or sometimes the smoothing parameter. It represents an integer depending on the sample size N. The Fourier coefficients estimators $\{\hat{a}_{m,N}\}$ could be written according to the sample set of random variable $X = (x_1, ..., x_N)$ as:

$$\{\hat{a}_{m,N}\} = \frac{1}{N} \sum_{i=1}^{N} e_m(x_i) \tag{2}$$

The convergence of this orthogonal series density estimator depends on the choice of k_N. Kronmal and Tarter investigated a method for the determination of the optimal choice of k_N (Kronmal and Tarter, 1968).

2.1 Smoothing Parameter Selection for Orthogonal Density Estimation

Convergence theorems have been established to find the optimal value of k_N for several error criteria. Among these criteria the *MISE* between the theoretical *PDF* and its estimate $\hat{f}_{k_N}$ can be expressed by:

$$MISE = E(\int |f(x) - \hat{f}_{k_N}(x)|^2 dx) \tag{3}$$

Where $E(.)$ is the expectation operator.

By replacing $\hat{f}_{k_N}(x)$ by its orthogonal density estimate and after some calculations given in (Kronmal and Tarter, 1968), the *MISE* could be written as follow:

$$MISE(\hat{f}_{k_N}(x)) \simeq \frac{1}{N-1} \sum_{i=0}^{k_N} [2\hat{d}_i - (N+1)\hat{a}_{i,N}^2] + \sum_{i=0}^{\infty} a_i^2 \tag{4}$$

Where

$$\hat{d}_i = \frac{1}{N} \sum_{j=1}^{N} e_i^2(x_j) \tag{5}$$

As $\sum_{i=0}^{\infty} a_i^2$ does not depend on k_N then searching for k_N which minimize the $MISE(\hat{f}_{k_N}(x))$ is the same to minimize the following expression:

$$J(k_N) = \frac{1}{N-1} \sum_{i=0}^{k_N} [2\hat{d}_i - (N+1)\hat{a}_{i,N}^2] \tag{6}$$

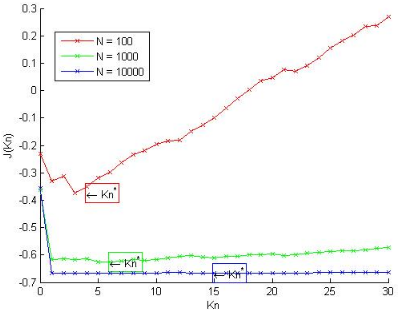

Figure 1: Behavior of $J(k_N)$ against k_N (k_N^* correspond to the minima of $J(k_N)$).

Kronmal and Tarter (Kronmal and Tarter, 1968) indicated the existence of an optimal value of k_N that minimize the $J(k_N)$. They proved that k_N has a value much smaller than the sample size N. Their strategy for determining the optimal value of k_N consists on the following rule. Starting from $k_N = 1$ and increase by 1 this value until $J(k_N)$ increase. The optimal value will be the value of k_N just before $J(k_N)$ increase. This strategy is called stopping rule. Since $J(k_N)$ may have multiple local minima, to avoid being trapped at a local minimum, one cannot simply increase m incrementally until $J(k_{N+1}) > J(k_N)$. For this reason, Kronmal and Tarter give an improvement on this rule by suggesting to stop the rule only if we obtain a certain sequence length of $\Delta J(k_N) = (J(k_N) - J(k_{N+1})) < 0$. This rule is adopted latter in (Beauville, 1978) and (Wong and Wang, 2005).

We give the following experience to illustrate this strategy. We generate a sample set of random variable X from a multimodal distribution composed of the superposition of two Gaussians with the following parameters $\mu_1 = 1, Var_1 = 2$ and $\mu_2 = 3, Var_2 = 1$. Here μ and Var correspond respectively to the mean and the variance of X. In Figure 1, we plot $J(k_N)$ against k_N for different values of sample size. The optimal values obtained are equal to 4, 6 and 15 respectively to sample size equal to 100, 1000 and 10000. This experimental results show that the optimal number of terms for the orthogonal density estimation is much smaller than the sample size. In addition the error $J(k_N)$ increases monotonically from all values greater than the optimal k_N. In the next section, we will review a standard method for dimensionality reduction and the one we investigated based on the $\hat{d}_{PF}$.

3 DIMENSIONALITY REDUCTION

The goal of a dimensionality reduction is to project high dimensional data samples in a low dimensional space in which groups of data are the most separated. In the following subsections, we recall the *LDA* method and the one based on the $\hat{d}_{PF}$. We also present a novel rule for selecting the optimal smoothing parameters values to improve the convergence of the $\hat{d}_{PF}$.

3.1 Linear Discriminant Analysis Method

The *LDA* is a widely used method for dimensionality reduction. It intends to reduce the dimension, so that in the new space, the between class distances are maximized while the within class ones are minimized. To that purpose, *LDA* considers searching for orthogonal linear projection matrix *W* that maximizes the following so-called Fisher optimization criterion (Fukunaga, 1990) :

$$J(W) = \frac{trace(W^T S_b W)}{trace(W^T S_W W)} \tag{7}$$

S_W is the within class scatter matrix and S_b is the between class scatter one. Their two well known expressions are given by:

$$S_W = \sum_{k=1}^{c} \pi_k E((X-\mu_k)(X-\mu_k)^T) \tag{8}$$

$$S_b = \sum_{k=1}^{c} \pi_k (\mu_k - \mu)(\mu_k - \mu)^T \tag{9}$$

Where μ_k is the conditional expectation of the original multidimensional random vector *X* relative to the class *k*. μ corresponds to the mean vector over all classes. c is the total number of classes and π_k denote the prior probability of the k^{th} class. $E(.)$ is the expectation operator.

Because it's not practical to find an analytical solution *W* that maximizes the criteria *J(W)*, one possible suboptimal solution is to choose *W* formed by the *d* eigenvectors of $S_W^{-1} S_b$ those correspond to the *d* largest eigenvalues. In general, the value of *d* is chosen to be equal to the number of classes minus one. After computation of *W*, the *LDA* method proceeds to the projection of the original data onto the reduced space spanned by the vectors of *W*. Note that this method is based only on first and second order moments and thus it assumes that the different underlying distributions of classes are normally distributed. This restrictive assumption constitutes a limitation to using *LDA* and makes it fail when dealing with non-Gaussian classes distributions.

3.2 Dimensionality Reduction using Patrick-Fisher Distance based on Orthogonal Series

Let recall the expression of the Patrick-Fisher distance d_{PF}:

$$d_{PF} = [\int_X |\pi_1 P_1(x/w_1) - \pi_2 P_2(x/w_2)|^2 dx]^{1/2} \tag{10}$$

Here π_i is the prior probability of class w_i and $P_i(x/w_i)$ denotes its conditional probability density.

The $\hat{d}_{PF}$ (Ghorbel, 2011) is obtained by substituting the conditional probability density by its orthogonal estimation into the expression of the d_{PF}. After some computations given in (Ghorbel, 2011), the expression of the $\hat{d}_{PF}$ could be written as follow:

$$\begin{aligned}\hat{d}_{PF}(W) = &\frac{1}{N^2}(\sum_{i=1}^{N1}\sum_{j=1}^{N1} K_{k_{N1}}(<W|x_i^1>,<W|x_j^1>) + \\ &\sum_{i=1}^{N2}\sum_{j=1}^{N2} K_{k_{N2}}(<W|x_i^2>,<W|x_j^2>) - 2Re(\\ &\sum_{i=1}^{N1}\sum_{j=1}^{N2} K_{min(k_{N1},k_{N2})}(<W|x_i^1>,<W|x_j^2>)))\end{aligned} \tag{11}$$

Where $<|>$ denotes the scalar product operator. x_i^k is the i^{th} observation of the k^{th} class. $Re(.)$ correspond to the real part of complex number and $K_{k_{Ni}}(x,y)$ is the kernel function associated to the orthogonal system of functions $\{e_m(x)\}$ used (Ghorbel, 2011). $\{N_i\}_{i\in\{1,2\}}$ is the sample size of the i^{th} class and N is the total size of all classes.

The $\hat{d}_{PF}$ expression depends on $\{k_{Ni}\}_{i\in\{1,2\}}$ which represent the number of terms to be used to estimate the *PDF* of the i^{th} class.

For dimensionality reduction purpose, this $\hat{d}_{PF}$ is considered as the criterion function to be maximized with respect to a linear projection matrix *W* that transform original data space onto a *d*-dimensional subspace so that classes are most separated. A linear projection matrix *W* that maximizes the $\hat{d}_{PF}$ should be found numerically. Since the equation of this estimator is highly nonlinear according to the element of *W* and an analytical solution is often practically not feasible, we will resort to an optimization algorithm to compute a suboptimal projection matrix *W*.

3.3 Smoothing Parameter Selection for Orthogonal Patrick-Fisher Distance Estimator

To determine the optimal numbers of $\{k_{Ni}\}_{i\in\{1,2\}}$ to be used for estimating the $\hat{d}_{PF}$, we propose to consider those minimizing the *MISE* criteria of this estimator. We define this latter as:

$$MISE = J(k_{N1}, k_{N2}) = E(|d_{PF} - \hat{d}_{PF}|^2) \quad (12)$$

Note that the simplicity of the *MISE* expression for orthogonal density estimator does not seem to be the same for the $\hat{d}_{PF}$. Hence, a numerical evaluation of k_{N_1} and k_{N_2} is extremely complex. For solving the optimal choice problem for the $\hat{d}_{PF}$, we propose to use for each class orthogonal density estimator the optimal value determined by the method described in section 2. Rather than used a pre-specified values, this choice seems to be reasonable to minimize the *MISE* of the $\hat{d}_{PF}$. To verify this purpose, we give the following simulation study. We generate two samples data from two different Gaussians distributions with parameters $\mu_1 = 1, Var_1 = 3$ and $\mu_2 = 3, Var_2 = 1$. Each sample have a size equal to 1000. We vary k_{N1} from 1 to $\sqrt{N1}$ and k_{N2} from 1 to $\sqrt{N2}$ and we calculate $\hat{d}_{PF}$ for each pair (k_{N1}, k_{N2}) by considering k_{N1} terms and k_{N2} terms to estimate respectively the *PDF* of the first sample and the second one. The theoretical d_{PF} can be calculated since we have the analytical expression of the Gaussian *PDF* of each sample. We approximate the integral in the expression of d_{PF} by using the Simpson's method (Atkinson, 1989). To estimate the expectation in the expression of $J(k_{N1}, k_{N2})$, we generate samples one hundred times and we calculate the means of the square difference between the d_{PF} and its orthogonal estimation $\hat{d}_{PF}$. Figure 2 shows the values of $J(k_{N1}, k_{N2})$. The pair (k_{N1}, k_{N2}) that minimizes $J(k_{N1}, k_{N2})$ is selected to be used as the optimal values of k_{N1} and k_{N2}.

Based on an extensive simulation, the values of k_{N1} and k_{N2} which minimize respectively the orthogonal density estimate of the first class and the second one give a sub-optimal solution to minimize $J(k_{N1}, k_{N2})$. This choice could be useful when we have no information about the *PDFs* of data which corresponds generally to the case of real world data.

4 EXPERIMENTAL RESULTS

In this section, we intend to compare the performances of the dimensionality reduction method based on the $\hat{d}_{PF}$ described above with the *LDA* both on simulated data and on real world dataset. To do that, we evaluate the classification accuracy of a nonparametric Bayesian classifier that is applied on the projected data onto the reduced space. We evaluate the classification accuracy by counting the number of misclassified samples obtained by the classifier over all classes of the projected data.

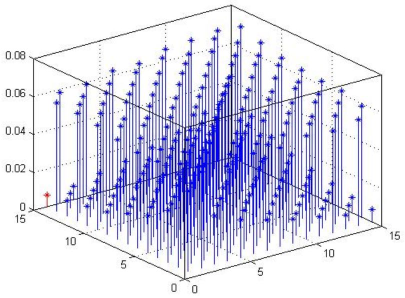

Figure 2: Values of $J(k_{N1}, k_{N2})$ against the different values of pair (k_{N1}, k_{N2}). In red color the selected minima of $J(k_{N1}, k_{N2})$.

4.1 Experiment with Simulated Data

This experiment concerns the two-class case. Vectors data from the first class are drawn from a multivariate Gaussian distribution with mean vector $\mu_1 = (3...3)^T$. For the second class, vectors data are generated from a mixture of two multidimensional Gaussians distributions. The first distribution has a mean vector $\mu_2 = (2...2)^T$ and the second has a mean vector $\mu_3 = (4...4)^T$. We consider for all these distributions the same covariance matrix $\sum = 2I$ where I denotes the identity matrix. The sample size for each class is equal to 1000 and generated vectors have a dimension equal to 14. We search for the projection vector W that map the generated data onto the optimal one-dimensional subspace according to the two methods of reduction studied. Note that the used system of the orthogonal functions is the trigonometric one (Hall, 1982). After finding the projection vector W according to each method, simulated data are projected onto the reduced space. Then we applied a Bayesian classifier on the projected data obtained. Classification results are summarized in Table 1. We remark that the dimensional reduction accuracy of the method based on the $\hat{d}_{PF}$ is better than the *LDA*.

Table 1: Classification results of experiment with simulated data.

	LDA method	Method based on $\hat{d}_{PF}$
Misclassification rate	0.47	0.22

The *LDA* method fails to find an optimal subspace in which satisfactorily class separation is obtained since the original simulated data contain multimodal distribution. However, the method based on the $\hat{d}_{PF}$ succeeds to overcome the restriction of unimodality. The success of this latter method can be explained by the fact that the $\hat{d}_{PF}$ based method accounts for higher order statistics and not just for the second order as in the *LDA*.

4.2 Experiment with Real World Data

In this experiment, we consider a sample set selected from the publicly available MNIST database containing binary images of handwritten digits. From this database we consider a subset formed by two classes of digits. Each class contains 1000 randomly selected digits. Figure 3 shows some examples of selected digits. Each digit is described by a features vector which is invariant under planar rotation, translations and scale factors. We denote the features vector by I_k. Among the large proposed invariant descriptors for planar contour shape in the literature, we consider here three different kinds of contour-descriptors. The first one is well known and is called Fourier descriptors (Ghorbel and Bougrenet, 1990). The second one introduced by Crimmins admits the completeness property (Crimmins, 1982). Third one introduced in (Ghorbel, 2011) gives in the same time the completeness and the stability properties to the descriptors. In the following, we recall the definitions of these three set of invariants descriptors. Let denote by γ a normalized arc length of a closed contour which represents the exterior handwritten boundary and by $C_k(\gamma)$ its corresponding Fourier coefficient with order k.

4.2.1 Fourier Descriptors Set

$$\{I_k\} = \frac{|C_k(\gamma)|}{|C_1(\gamma)|} \qquad \forall\ k \geq 2 \tag{13}$$

4.2.2 Crimmins Descriptors Set

$$\begin{cases} I_{k_0} = |C_{k_0}(\gamma)| \\ I_{k_1} = |C_{k_1}(\gamma)| \\ \begin{cases} \forall\ k \neq k_0 \ and\ k \neq k_1\ : \\ I_k = C_k^{k_0-k_1}(\gamma) C_{k_0}^{k_1-k}(\gamma) C_{k_1}^{k-k_0}(\gamma) \end{cases} \end{cases} \tag{14}$$

4.2.3 Ghorbel Descriptors Set

$$\begin{cases} I_{k_0} = C_{k_0}(\gamma) \\ I_{k_1} = C_{k_1}(\gamma) \\ \begin{cases} \forall\, k \neq k_0\ and\ k \neq k_1 : \\ I_k = 0\ \ if\ |C_{k_0}(\gamma)| = 0\ or\ |C_{k_1}(\gamma)| = 0 \\ I_k = \dfrac{C_k^{k_0-k_1}(\gamma) C_{k_0}^{k_1-k}(\gamma) C_{k_1}^{k-k_0}(\gamma)}{|C_{k_0}(\gamma)|^{k_1-k-p} |C_{k_1}(\gamma)|^{k-k_0-q}}\ \ otherwise \\ (p\ and\ q\ are\ two\ fixed\ positive\ floats) \end{cases} \end{cases} \tag{15}$$

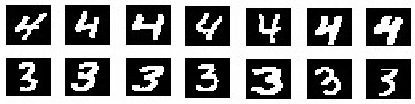

Figure 3: Some examples of digits selected from MNIST database.

We compute Fourier coefficients from digit outline boundary and we construct the three invariants descriptors sets as defined above. We consider for each digit shape from our selected sample the first fourteen Fourier coefficients. After dimensional reduction with the two methods studied, classification results are computed and are illustrated in Table 2.

We notice that the two dimensionality reduction methods perform similar when using the Fourier descriptors dataset and Ghorbel descriptors. These results can be justified by the fact that these two sets of descriptors verify the property of stability introduced in (Ghorbel, 2011). This property expresses the fact that low level distortion of the shape does not induce a noticeable divergence in the set of descriptors. Hence, the distribution of the invariant descriptors associated to each class has one mode since in this case we can assume that only the first and the second statistical moments are needed to estimate their conditional *PDFs*. This is not the case for Crimmins descriptors since it does not verify the stability property so that the induced classes *PDFs* could be multi-modal.

Table 2: Misclassification rate results of experiment with real dataset.

	LDA method	Method based on $\hat{d}_{PF}$
Fourier descriptors set	0.26	0.25
Crimmins descriptors set	0.38	0.15
Ghorbel descriptors set	0.1	0.1

5 CONCLUSIONS

In this paper we propose a rule for the determination of the optimal number of terms in an orthogonal series for best approximation of the orthogonal Patrick-Fischer distance estimator. When data are multi-modal, experimental results both on simulated data and on real dataset have shown that the orthogonal Patrick-Fisher distance estimator gives better performance in the mean of misclassification rate since it increases the probabilistic measure between the projected classes onto the reduced space and decreases the number of the misclassified samples. Otherwise, when the different conditional distributions are with one mode, the *LDA* performance becomes similar. Thus, *LDA* becomes preferable because of its relative algorithmic simplicity. Simulation data and real data basis are tested in order to prove the importance of the adjustment of the orthogonal Patrick-Fischer distance estimator. In our future work, we will consider the multi-class case.

REFERENCES

Aladjem, M. (1997). Linear discriminant analysis for two classes via removal of classification structure. In *In IEEE PAMI*.

Atkinson, K. (1989). *An Introduction to Numerical Analysis*. John Wiley and Sons.

Beauville, J. P. A. (1978). Estimation non paramtrique de la densit et du mode exemple de la distribution gamma. In *In Revue de Statistique Applique*.

Cencov, N. N. and Nauk, D. Z. (1962). Estimation of an unknown density function from observations. In *In SSSR*.

Crimmins, T. (1982). A complete set of fourier descriptors for two-dimensional shapes. In *In IEEE Trans. Syst.*

Drira, W. and Ghorbel, F. (2012). Dimension reduction by an orthogonal series estimate of the probabilistic dependence measure. In *ICPRAM'12, Int. Conf. on Pattern Recognition Applications and Methods*, Portugal.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Classification*. Academic Press, New York.

Ghorbel, F. (2011). *Vers une approche mathmatique unifie des aspects gomtrique et statistiques de la reconnaissance des formes planes*. arts-pi, Tunisie.

Ghorbel, F. and Bougrenet, J. T. (1990). Automatic control of lamellibranch larva growth using contour invariant feature extraction. In *Pattern Recognition*.

Ghorbel, F., Derrode, S., and Alata, O. (2012). *Rcentes avances en reconnaissance de formes statistique*. arts-pi, Tunisie.

Hall, P. (1982). Comparison of two orthogonal series methods of estimating a density and its derivatives on an interval. In *In Journal of Multivariate Analysis*.

Hillion, A., Masson, P., and Roux, C. (1988). Une mthode de classification de textures par extraction linaire non paramtrique de caractristiques. In *In Colloque TIPI*.

Kronmal, R. and Tarter, M. (1968). The estimation of probability densities and cumulatives by fourier series methods. In *In JASA, Journal of the American Statistical Association*.

Patrick, E. A. and Fisher, F. P. (1969). Non parametric feature selection. In *In IEEE Trans. Information Theory*.

Wong, K. W. and Wang, W. (2005). Adaptive density estimation using an orthogonal series for global illumination. In *In Computers and Graphics*.