

Density Function Interpretation of Subspace Classification Methods

J. Laaksonen, E. Oja
 Laboratory of Computer and Information Science
 Helsinki University of Technology
 P.O. BOX 2200, FIN-02015 HUT, Finland
Jorma.Laaksonen@hut.fi, Erkki.Oja@hut.fi

Abstract

Subspace methods are a powerful class of statistical pattern classification algorithms. They are traditionally considered nonparametric or semiparametric models, which means that the classifier does not produce the *a posteriori* probability of the input vector. This paper, however, presents a novel density estimation interpretation for the subspace methods. As a byproduct, the problem of unequal priors, which occurs if standard subspace methods are used when the *a priori* probabilities of the classes are unequal, is eliminated. As well, the *a posteriori* probability produced by the suggested method can be used to implement a plausible rejection method for difficultly classifiable patterns. At the end of this presentation, the applicability of the proposed probability density function interpretation is demonstrated with experiments.

1 Introduction

Subspace methods [6] are a powerful class of statistical pattern classification algorithms. They are nonparametric methods in the sense that the subspace formalism does not contain implicit probability density function model, the parameters of which would be estimated from the training sample. Instead, the vectors of the training set are used in a collective manner to span a lower-dimensional subspace for each pattern class. The input vectors are then classified according to the largest projection length on the class-conditional subspaces.

The classification decision is thus based entirely on measurements in the feature space and the concept of probability is not present in any form. As the likelihood of the occurrence of a particular input vector in its decided class is not expressed, there is no way to ground a rejection mechanism on the *a posteriori* probabilities. Another well-known shortcoming of the subspace classifiers is their limited usage in cases in which the *a priori* probabilities of the classes are unequal. This problem is not present in other classification algorithms, such as the density estimation, regression, or prototype based classifiers.

We propose here a probabilistic interpretation of the subspace classification rule. The interpretation is based on modeling the distributions of the squared vector distances from the subspaces. It offers as a byproduct solutions both to the selection of a plausible rejection decision and to the problem of unequal priors. The traditional subspace formalism is shortly reviewed in Section 2 before the novel density estimation interpretation is presented in Section 3. Empirical evidence on the usability of the proposed method is presented in Section 4.

2 Subspace Classifiers

In subspace classifiers, a pattern class j is represented as an ℓ_j -dimensional subspace $\mathcal{L}_j$ of the feature space $\mathbb{R}^d$ by using a set of orthonormal column vectors, $\{\mathbf{u}_{1j}, \dots, \mathbf{u}_{\ell_j j}\}$. These vectors can further be combined into a matrix $\mathbf{U}_j$, which leads to notation

$$\mathcal{L}_j = \mathcal{L}_{\mathbf{U}_j} = \{ \mathbf{x} \mid \mathbf{x} = \sum_{i=1}^{\ell_j} \zeta_{ij} \mathbf{u}_{ij}, \zeta_{ij} \in \mathbb{R} \} = \{ \mathbf{x} \mid \mathbf{x} = \mathbf{U}_j \mathbf{z}_j, \mathbf{z}_j \in \mathbb{R}^{\ell_j} \}. \quad (1)$$

By using orthogonal projection $\mathbf{P}_j = \mathbf{U}_j \mathbf{U}_j^T$, any vector $\mathbf{x} \in \mathbb{R}^d$ can be represented as a sum of two mutually orthogonal vectors, $\hat{\mathbf{x}}_j$ which belongs to the subspace $\mathcal{L}_j$ and $\tilde{\mathbf{x}}_j = \mathbf{x} - \hat{\mathbf{x}}_j$ which is perpendicular to it. The

orthogonal projection of $\mathbf{x}$ on the subspace of the class j is calculated as

$$\hat{\mathbf{x}}_j = \mathbf{P}_j \mathbf{x} = \mathbf{U}_j \mathbf{U}_j^T \mathbf{x} = \sum_{i=1}^{\ell_j} (\mathbf{x}^T \mathbf{u}_{ij}) \mathbf{u}_{ij} = \sum_{i=1}^{\ell_j} \zeta_i \mathbf{u}_{ij} . \quad (2)$$

In CLAFIC [8], the correlation matrix $\mathbf{R}_j = E\{\mathbf{x}\mathbf{x}^T \mid \mathbf{x} \in j\}$ is estimated separately for each class j and its first ℓ_j eigenvectors, $\mathbf{u}_{1j}, \dots, \mathbf{u}_{\ell_j j}$ with the largest eigenvalues, are used as the columns of the basis matrix $\mathbf{U}_j$. The classification function $g(\mathbf{x}) \in \{1, \dots, c\}$, where c is the number of classes, makes the classification decision according to the maximal similarity in terms of projection length:

$$g_{\text{CLAFIC}}(\mathbf{x}) = \underset{j=1, \dots, c}{\operatorname{argmax}} \|\hat{\mathbf{x}}_j\| . \quad (3)$$

The coefficients $\zeta_{ij} = \mathbf{x}^T \mathbf{u}_{ij}$ are the principal components of the input pattern vector on the subspace of the class j . In the classical procedure, the sample mean $\hat{\boldsymbol{\mu}}$ of the pooled training set is subtracted from the pattern vectors before they are classified or used in initializing the classifier. In a more developed version [3], class-conditional means $\hat{\boldsymbol{\mu}}_j$ are used instead of the common mean, and the class-conditional correlations $\mathbf{R}_j$ are thus replaced with the covariances $\boldsymbol{\Sigma}_j$. The classification function is then based on the minimization of the residual length,

$$g_{\text{CLAFIC-}\boldsymbol{\mu}}(\mathbf{x}) = \underset{j}{\operatorname{argmin}} \|\tilde{\mathbf{x}}_j\| = \underset{j}{\operatorname{argmin}} \|(\mathbf{I} - \mathbf{U}_j \mathbf{U}_j^T)(\mathbf{x} - \hat{\boldsymbol{\mu}}_j)\| . \quad (4)$$

This latter formalism is used in the following illustrations. Otherwise, the two versions are mutually exchangeable in all respects of the forthcoming discussion. Basically, there exist two possibilities to select the subspace dimensions $\ell_1, \dots, \ell_c$. Either, each subspace is given an equal dimension ℓ , or, the selection of the dimensions is based on observing the decay of the eigenvalues individually in each class. An iterative search algorithms may also be formulated [4].

3 Density Function Interpretation

Pattern classification methods based on density estimates often assume Gaussian distributed pattern vectors within the classes. This assumption is not as such transferable to subspace formalism, and – possibly therefore – no interpretation of subspace classification in terms of density function estimation has been presented this far. In this section, however, such an interpretation based on the gamma probability density function is introduced. The course of the presentation is as follows: First, we demonstrate that the coefficients ζ_{ij} in Eq. 2 tend to be zero-mean normally distributed. Second, we show that they may be regarded as mutually independent. Third, we argue that the sum of their squares can be approximated with gamma distribution. Fourth, we show that the resulting probability density function estimates prove effective in the sense of classification accuracy.

The particular pattern recognition task studied in this section is the classification of handwritten digits. The binary input images were first segmented and normalized to the size of 32×32 pixels. The Discrete Karhunen-Loève Transform was then used to extract 64-dimensional feature vectors. The total number of vectors in the ten equally represented digit classes was 8940 in both training and testing samples. The same sets of data were earlier used in a benchmarking study [2].

3.1 Distribution of the Coefficients

Fig. 1 shows histograms of the marginal probability densities of the coefficients ζ_{ij} obtained from handwritten digits in class ‘7’ on their own subspace $\mathcal{L}_7$. It can be seen that the distribution of $\zeta_{1,7}$ is somewhat flat, $\zeta_{2,7}$ is positively and $\zeta_{3,7}$ negatively slanted. Otherwise, the larger the i the more the distribution of ζ_{ij} resembles zero-mean normal distribution with decreasing variance. Whether the ζ_{ij} coefficients in a particular application really are normally distributed or not, can be tested statistically. Applicable tests include the parametric χ^2 test and the nonparametric Lilliefors test [1]. In this example case of handwritten ‘7’s, both of the tests show, in general terms, that the first half-dozen coefficients are non-Gaussian, the succeeding coefficients up to index $i \approx 40$ are normally distributed, and the remaining ones are again non-Gaussian.

An appealing explanation for the Gaussianity of the distributions can be found by observing how the feature vectors $\mathbf{x}$ were formed in this particular case: the Karhunen-Loève Transform performs 64 inner products between the 1024-dimensional input image and the estimated principal masks of the training sample. Thus, because the

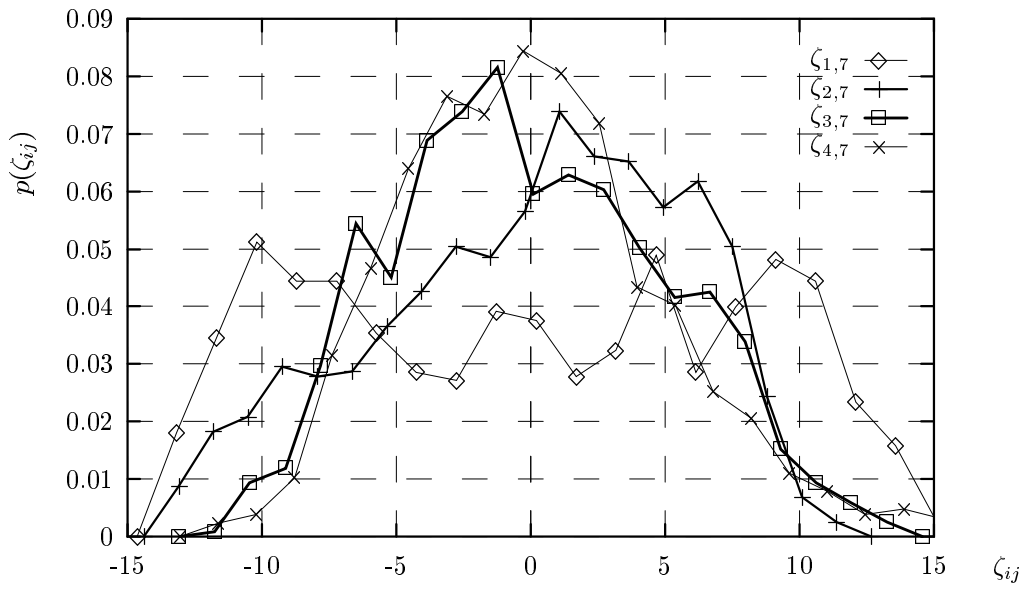


Figure 1: Measured distributions of single-variable marginal probability densities of $\zeta_{1,7}, \dots, \zeta_{4,7}$ of a sample of handwritten digits ‘7’.

coefficients ζ_{ij} are further inner products of the $\mathbf{u}_{ij}$ basis vectors and the feature vector $\mathbf{x}$, they are actually sums of 65536 products of to some extent independent but not identically distributed values. The Central Limit Theorem [5] may be interpreted to state that these kinds of series are asymptotically normally distributed.

3.2 Independence of the Coefficients

Next, the potential mutual independence of the ζ_{ij} coefficients is examined. Again, the projections of the handwritten digits ‘7’ serve as an illustration. Some two-dimensional marginal distributions of the ζ_{ij} s are displayed in Fig. 2. Each subfigure shows the dependence of two coefficients ζ_{i7} and ζ_{k7} for some small indices. The distributions in Fig. 1 were in fact one-dimensional marginal distributions of the distributions of Fig. 2. It is observable that when, say, $i \geq 4$ and $k > i$ the distributions in the $\zeta_{i7}\zeta_{k7}$ plane are quite unimodal and symmetric around the origin, thus resembling the bivariate Gaussian distribution.

In this exemplary case, the coefficient pairs can truly be shown to be uncorrelated by calculating the correlation coefficient for each pair. The resulting t -statistics justifies that the ζ_{ij} variables are uncorrelated. From the uncorrelatedness and the Gaussianity of the coefficients, it follows that they are independent as wished. From the Gaussianity it also follows that the coefficients are not only pairwise independent but mutually independent in all combinations.

3.3 Distribution of the Squared Residual Length

Now, we turn to analyze the lengths of the residual vectors $\tilde{\mathbf{x}}_j$. The squared residual length can be expressed as a sum of a truncated series as in Eq. 2,

$$\|\tilde{\mathbf{x}}_j\|^2 = \sum_{i=\ell_j+1}^d (\mathbf{x}^T \mathbf{u}_{ij})^2 = \sum_{i=\ell_j+1}^d \zeta_{ij}^2 \quad (5)$$

The first ℓ_j coefficients $\zeta_{1j}, \dots, \zeta_{\ell_j j}$, which are normally used in calculating the projection length, are of no interest when the length of the residual is concerned. Consequently, only those terms which above were observed to resemble the Gaussian distribution are involved in forming the squared length $\|\tilde{\mathbf{x}}_j\|^2$. If the ζ_{ij} s really were independent and Gaussian and, additionally, had equal variance, i.e., $\sigma_{\zeta_{ij}}^2 = \sigma_j^2$, then $\sigma_j^{-2} \|\tilde{\mathbf{x}}_j\|^2$ would be distributed as $\chi_{d-\ell_j}^2$. In reality, of course, the $\sigma_{\zeta_{ij}}^2$ s decrease continuously as i increases and cannot be replaced with a common σ_j^2 .

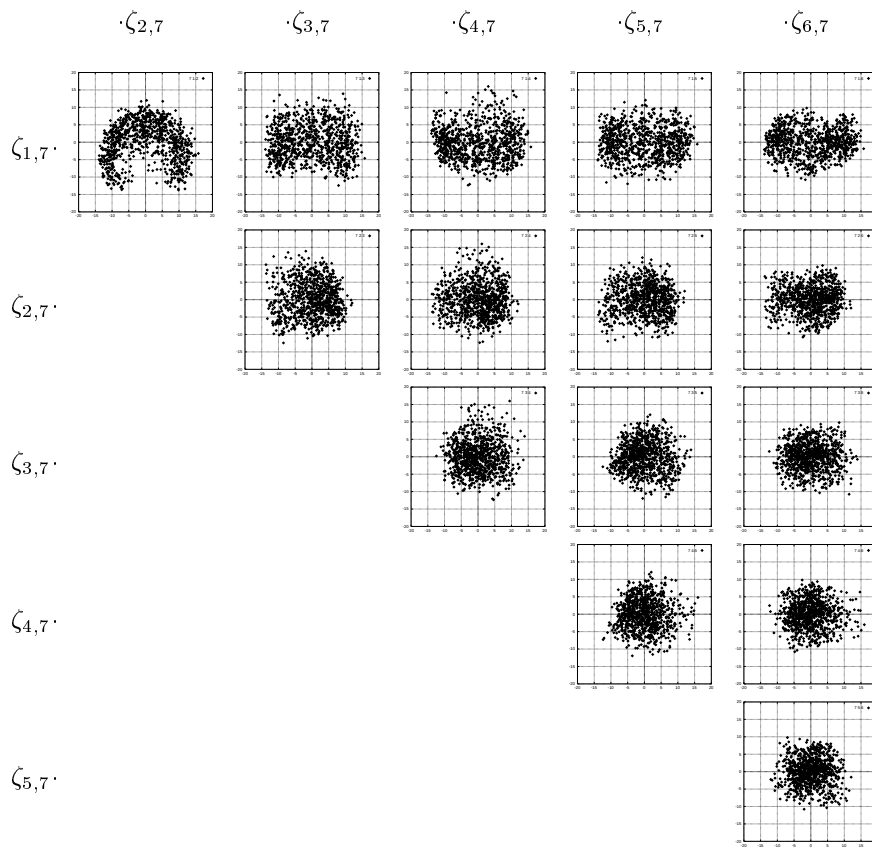


Figure 2: Projections of some ‘7’s on subspace $\mathcal{L}_7$. Each subfigure displays the marginal distribution in the plane of the coefficients ζ_{i7} and ζ_{k7} , shown on the left and on the top, respectively. The ranges of the variables are $[-20, 20]$ in all images.

Therefore, it is not reasonable to try to model the distribution of $\|\tilde{\mathbf{x}}_j\|^2$ with a χ^2 probability density function. Nevertheless, we make here the assumption that $\|\tilde{\mathbf{x}}\|^2$ approximately follows the gamma probability density function $f_\gamma(x; \alpha, \nu)$ which is the underlying form of the χ^2 distribution but has two free parameters instead of the one free parameter of the χ^2 distribution. The gamma density is expressed as

$$f_\gamma(x; \alpha, \nu) = \frac{\alpha^\nu}{\Gamma(\nu)} x^{\nu-1} e^{-\alpha x}, \quad x, \alpha, \nu > 0, \quad (6)$$

$$\Gamma(\nu) = \int_0^\infty t^{\nu-1} e^{-t} dt, \quad \nu \geq 0. \quad (7)$$

The fitting of the gamma distribution to the observed values may be executed either by maximum likelihood estimation or – like here with the continuing example of handwritten digits ‘7’ – using the method of moments [7]. Fig. 3 shows a distribution histogram of $\|\tilde{\mathbf{x}}\|^2$ (with $\ell_j = 24$ and $d = 64$ in Eq. 5) in the case of handwritten digits ‘7’. The gamma probability density function which results from the estimated parameters is plotted aside.

Thus, a model which is able to estimate the value of the probability density function of a pattern class at the location of the input vector $\mathbf{x}$ has been formulated. At the same time, it has been demonstrated that at least the real-world data used for illustration follows that model with some degree of fidelity. These observations can be extrapolated to say that also other natural data may be modeled similarly and their probability function be approximated with the gamma probability density function.

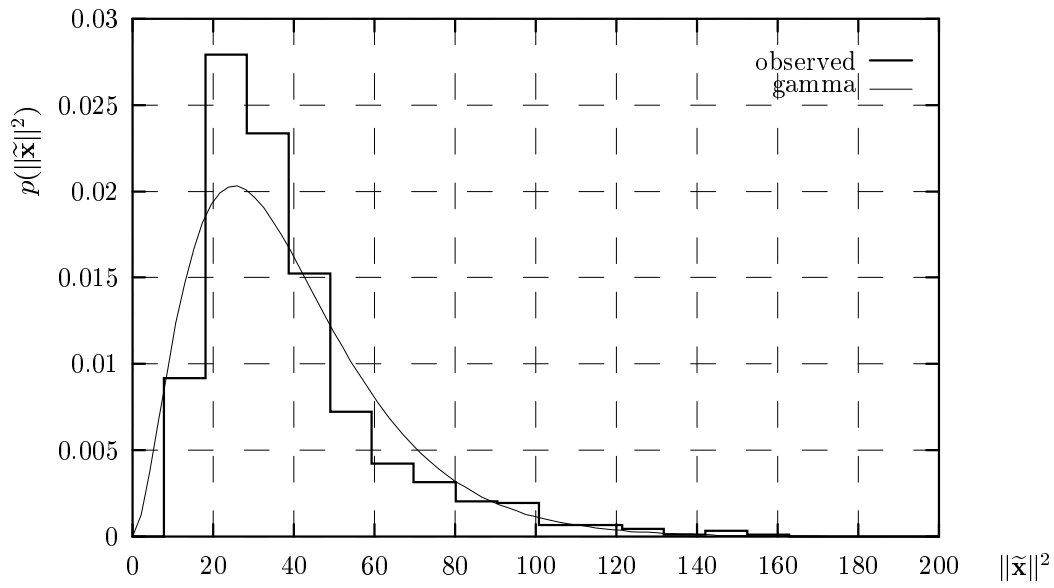


Figure 3: Empirical $\|\tilde{\mathbf{x}}\|^2$ distribution of handwritten ‘7’s plotted as a histogram. The corresponding estimated gamma distribution model is drawn as a continuous function.

3.4 Reformulation of the Classification Rule

In classical Bayesian theory of optimal decision estimation, the classification decision is made under the circumstances that the *a priori* probabilities P_j and the probability density functions $f_j(\mathbf{x})$ are given and the *a posteriori* probabilities $q_j(\mathbf{x})$ can be calculated thereof for all the classes:

$$g_{\text{BAYES}}(\mathbf{x}) = \underset{j=1,\dots,c}{\operatorname{argmax}} q_j(\mathbf{x}) = \underset{j=1,\dots,c}{\operatorname{argmax}} P_j f_j(\mathbf{x}) . \quad (8)$$

We now define a new classification function by inserting the gamma models and the *a priori* probabilities to Eq. 8:

$$g_{\text{CLAFIC-PDF}}(\mathbf{x}) = \underset{j=1,\dots,c}{\operatorname{argmax}} \hat{q}_j(\mathbf{x}) = \underset{j=1,\dots,c}{\operatorname{argmax}} \hat{P}_j f_\gamma(\|\tilde{\mathbf{x}}_j\|^2; \hat{\alpha}_j, \hat{\nu}_j) . \quad (9)$$

The $\|\tilde{\mathbf{x}}_j\|$ values and the gamma distribution model can either be calculated from the projection of Eq. 2, or from the Eq. 4 version, in which the class-conditional means are first subtracted. In either case, the classification is no longer based on maximizing the projection length nor on minimizing the residual length, but on maximizing the *a posteriori* probability of the observed squared residual length. The intrinsic invariance of the subspace classifier concerning the norm of the pattern vector $\mathbf{x}$ has now disappeared. This may be regarded as either a pro or con, depending on the nature of the input data.

The involvement of the estimated *a priori* probability $\hat{P}_j$ of the class j in Eq. 9 allows for the presented classifier to be used in cases in which the classes are not equally represented in the mixture. Additionally, a rejection mechanism can be expressed by setting a rejection threshold $0 \leq \theta \leq 1 - 1/c$ and reject $\mathbf{x}$ if the *a posteriori* probability of the decided class is suspiciously low:

$$1 - \max_{j=1,\dots,c} \hat{q}_j(\mathbf{x}) = 1 - \max_{j=1,\dots,c} \frac{\hat{P}_j \hat{f}_j(\mathbf{x})}{\sum_{i=1}^c \hat{P}_i \hat{f}_i(\mathbf{x})} > \theta. \quad (10)$$

4 Experiments

This section presents the results of experiments performed with handwritten digit data used earlier in a large comparison of neural and statistical classification methods [2]. The ten subspaces $\mathcal{L}_0, \dots, \mathcal{L}_9$ were all generated separately both with the subtraction of the overall mean of the data and with class-conditional means. The former

results are shown in the left column of Table 1 and the latter in the right. First, the classification error percentages were calculated with the standard subspace classification rule. Then the gamma probability density function interpretation was applied and the corresponding classification accuracies evaluated. The results are shown as the two lines of Table 1. All the subspaces were given equal dimensionality ℓ which was varied in order to find the best selection individually for all the four results presented. The value of ℓ is shown in parenthesis after the classification result.

Table 1: Classification error percentages with the traditional and the proposed subspace classifiers. The optimal mutual subspace dimension ℓ is shown in parenthesis.

	CLAFIC	CLAFIC- μ
CLAFIC	4.1% (27)	3.9% (21)
CLAFIC-PDF	4.1% (24)	3.8% (23)

5 Discussion and Future Work

In this paper, we have introduced a new probability density function interpretation of subspace classification. The performed classification experiments show that the presented method has at least the same classification accuracy as the conventional subspace method. Also, it seems that the proposed technique can be combined with both the two subspace classification principles, Eqs. 3 and 4. In an extensive benchmarking study, reported in [2], subspace methods performed well in handwritten digit classification, clearly surpassing such methods as linear discriminant analysis, tree classifier, and multi-layer perceptron networks. More importantly, the theoretical advantages of the proposed method over the conventional subspace classifier include: 1) The new method facilitates plausible rejection of difficultly classifiable input vectors, and 2) The new method is able to compensate for uneven *a priori* distribution of the input vectors within the classes.

According to the Bayesian theory, the misclassifications take place in areas where the class distributions are sparse. In the case of a subspace classifier, these areas are those where the distances to the model subspaces are at the largest. The proposed algorithm tries to model the distributions of the $\|\tilde{\mathbf{x}}\|$ s globally. If only the extreme tail of the distribution would be fitted to a functional model, enhanced classification accuracy could be expected to result.

References

- [1] W. J. Conover. *Practical Nonparametric Statistics 2ed.* John Wiley & Sons Inc., 1980.
- [2] L. Holmström, P. Koistinen, J. Laaksonen, and E. Oja. Neural and statistical classifiers – taxonomy and two case studies. *IEEE Transactions on Neural Networks*, 8(1):5–17, January 1997.
- [3] J. Laaksonen. *Subspace Methods in Recognition of Handwritten Digits.* PhD thesis, Helsinki University of Technology, 1997.
- [4] J. Laaksonen and E. Oja. Subspace dimension selection and averaged learning subspace method in handwritten digit classification. In *Proceedings of the International Conference on Artificial Neural Networks*, pages 227–232, Bochum, Germany, 1996.
- [5] J. M. Mendel. *Lessons in Estimation Theory for Signal Processing, Communications, and Control.* Prentice-Hall, 1995.
- [6] E. Oja. *Subspace Methods of Pattern Recognition.* Research Studies Press Ltd., Letchworth, England, 1983.
- [7] H. W. Sorenson. *Parameter estimation - principles and problems.* Marcel Dekker, New York, 1980.
- [8] S. Watanabe, P. F. Lambert, C. A. Kulikowski, J. L. Buxton, and R. Walker. Evaluation and selection of variables in pattern recognition. In J.T. Tou, editor, *Computer and Information Sciences II.* Academic Press, New York, 1967.