A HETEROSCEDASTIC EXTENSION OF LDA BASED ON MULTI-CLASS MATUSITA AFFINITY

Mohammad Shahin Mahanta, Konstantinos N. Plataniotis

The Edward S. Rogers Sr. Department of Electrical and Computer Engineering
University of Toronto
Email: {mahanta, kostas} @comm.utoronto.ca

ABSTRACT

Linear discriminant analysis (LDA), a conventional feature extraction technique, is a homoscedastic solution and ignores the second order information of the data. A heteroscedastic extension of LDA has been previously proposed which relies on the average pairwise Chernoff distances of the classes. However, in a multi-class scenario with number of classes $C > 2$, the average of pairwise distances is not directly related to the classification error rate. Furthermore, the corresponding method imposes a high computational complexity of order $O(C(C-1))$. This paper proposes an inherently multi-class heteroscedastic extension of LDA based on Matusita's separability measure, a multi-class generalization of the Chernoff distance which is related to multi-class error bounds. The proposed feature extractor can be trained non-iteratively with computational complexity of $O(C)$. Experimental comparisons with the Chernoff method demonstrate both a performance improvement when estimated parameters are used, and a reduction of factor $C-1$ in the computational load as predicted.

Index Terms— Heteroscedastic feature extraction, Chernoff distance, Matusita affinity, Gaussian quadratic classifier, multi-class separability measure.

1. INTRODUCTION

Linear feature extraction techniques such as linear discriminant analysis (LDA) are widely used to simplify the classification of high-dimensional data and simultaneously reduce the computational complexity [1]. These techniques are essential in applications such as automated medical diagnosis, data mining, bioinformatics, personal identification from biometrics, and speech recognition [2] where the high dimensionality of the data poses a major challenge to the classification problem. A popular feature extraction method, LDA, assumes that the data are homoscedastic, i.e. the covariances of the different classes are the same [3]. Thus, in a heteroscedastic scenario with different class covariances, LDA ignores and hence eliminates the significant information in the covariances of the data. Removal of covariance information also restricts the number of LDA features to the dimensionality of the subspace spanned by the class means.

Several feature extraction methods have been proposed for heteroscedastic data. Iterative procedures have been proposed to maximize criteria such as the data likelihood [4] or average pairwise Chernoff distances of the classes [5, 6]. But these approaches depart from the efficient training algorithm of LDA. Thus, simple non-iterative procedures have also been proposed based on criteria such as the mutual information between the features and the classes [7], the geometrical mean of pairwise Mahalanobis distances between the classes [8], or approximation of a linear sufficient statistic [9, 10].

One of the commonly used non-iterative heteroscedastic techniques, the Chernoff method, is based on a modified version of LDA [11]: The LDA criterion is expressed as an average of pairwise Euclidean distances. These pairwise distances are replaced with pairwise Chernoff distances in the Chernoff criterion. However, unlike the Chernoff distance, the average of pairwise Chernoff distances is not directly related to a bound on the classification error which could provide a theoretical ground for the criterion. In fact, maximizing the average pairwise distance may result in separation of the already separated classes, and possibly overlap of the originally adjacent classes [12].

These drawbacks of the Chernoff method can be obviated by minimizing a multi-class bound on the probability of error. Such a bound can be derived using the union bound or the equivocation [12]. However, this approach leads to much higher computational complexity due to the required iterative optimization.

In our proposed non-iterative feature extractor, we use Matusita's multi-class separability measure instead of the average Chernoff distance. Matusita's separability measure is a multi-class generalization of the Chernoff distance [13], and can be used to find an upper bound on the multi-class probability of error [14]. Based on this measure, we will propose a non-iterative procedure for linear feature extraction which is related to the corresponding error bounds. Moreover, the pairwise formulation of the Chernoff method imposes computations of order $O(C(C-1))$ for C classes, which is reduced to $O(C)$ for our proposed method. This reduction of computational complexity by a factor of $C-1$ is accurately verified using experiments on synthetic and real-world data.

2. REVIEW OF CHERNOFF FEATURE EXTRACTOR

In this section, first the typical linear feature extraction problem is formulated. Then, LDA and the need for heteroscedastic solutions such as the Chernoff method are reviewed.

Consider classification of the data[1] $\mathbf{x} \in \mathbb{R}^n$ into one of the classes ω_i, $1 \leq i \leq C$. The non-zero prior probability for ω_i is shown as p_i, and can be estimated as the fraction of the training samples belonging to ω_i, i.e. $\frac{N_i}{N}$. Also, the sample mean and covariance of class ω_i are denoted respectively as $\mathbf{m}_i$ and $\mathbf{S}_i$. The average mean is shown as $\mathbf{m}$, and the average sample covariance, also called within-class scatter, can be written as $\mathbf{S}_W$. The following formulation assumes $\mathbf{S}_W = \mathbf{I}_{n\times n}$, which can be guaranteed in the general case through initial normalization of the data by the factor $\mathbf{S}_W^{-1/2}$ [11].

[1] In this paper, scalars, vectors, and matrices are respectively shown in regular lowercase/uppercase (e.g. a or A), boldface lowercase (e.g. $\mathbf{a}$), and boldface uppercase (e.g. $\mathbf{A}$). Also, the matrix logarithm for positive definite (p.d.) matrix $\mathbf{S}_{n\times n}$, as defined in [15], is denoted as $\log(\mathbf{S})$.

Linear features of the data are generally calculated as

$$\mathbf{y}_{d\times 1} = \mathbf{T}_{d\times n}\mathbf{x}_{n\times 1}. \quad (1)$$

In the LDA method, the operator $\mathbf{T}$ is selected so that the criterion $tr\{(\mathbf{T}\mathbf{T}^T)^{-1}\mathbf{T}\mathbf{S}_B\mathbf{T}^T\}$ is maximized, with the between class scatter $\mathbf{S}_B = \sum_{i=1}^{C} p_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T$ [1, 11]. However, the formulation of $\mathbf{S}_B$ ignores possible differences in the class covariances, and also restricts the number of LDA features to $C-1$.

To incorporate the covariance information into LDA, in [11], the LDA criterion is expressed based on an average of pairwise Euclidean distances between the class means. Then, each pairwise distance is replaced by the Chernoff distance between the class pair. The resulting average of pairwise Chernoff distances incorporates the differences between class means as well as covariances. The Chernoff feature extractor based on this criterion retains the desirable non-iterative nature of LDA. In this method, $\mathbf{T}$ is obtained by selecting its rows as the d eigenvectors corresponding to the largest eigenvalues of

$$\mathbf{S}_C = \sum_{i=1}^{C-1}\sum_{j=i+1}^{C} p_i p_j \Big\{ \mathbf{S}_{ij}^{-\frac{1}{2}}(\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{S}_{ij}^{-\frac{1}{2}} + \frac{1}{\pi_i\pi_j}(\log(\mathbf{S}_{ij}) - \pi_i\log(\mathbf{S}_i) - \pi_j\log(\mathbf{S}_j)) \Big\}, \quad (2)$$

where $\pi_i = \frac{p_i}{p_i+p_j}$, $\pi_j = \frac{p_j}{p_i+p_j}$, and $\mathbf{S}_{ij} = \pi_i\mathbf{S}_i + \pi_j\mathbf{S}_j$ [11].

The Chernoff distance between two classes, which is the basis for the Chernoff criterion, is related to upper bounds on the probability of binary classification error [16]. However, the multi-class measure obtained by the average of pairwise Chernoff distances is not similarly related to the probability of multi-class classification error. In fact, the average of pairwise distances can be dominated by the distance of class pairs which are already far apart in the original space, while the contribution of adjacent classes is non-significant [12]. Furthermore, the pairwise formulation of the Chernoff method leads to a computational overhead in the training phase. To alleviate these problems, we propose a heteroscedastic feature extractor based on an inherently multi-class distance measure. Both theoretical and experimental results indicate that this approach leads to a significant improvement in the computational complexity of the Chernoff method.

3. PROPOSED MATUSITA FEATURE EXTRACTOR

This section introduces a heteroscedastic extension of LDA which is a non-iterative procedure as the Chernoff method. Yet, it is related to the maximization of the multi-class Matusita separability measure, and hence minimization of the corresponding error bounds.

As a common assumption in many applications [10] which also underlies the Chernoff criterion, we assume that the data in class ω_i are distributed as $\mathcal{N}(\mathbf{m}_i, \mathbf{S}_i)$, for $1 \le i \le C$. We also assume that $\mathbf{S}_W = \mathbf{I}$. The weighted Matusita *affinity* is a measure of similarity or overlap of distributions, and can be calculated for our assumed Gaussian distributions as [13]

$$\rho_w = \frac{\prod_{i=1}^{C}|\mathbf{S}_i|^{-\frac{p_i}{2}}}{|\mathbf{Q}|^{\frac{1}{2}}} \times \exp\Big\{\frac{1}{2}\mathbf{q}^T\mathbf{Q}^{-1}\mathbf{q} - \sum_{i=1}^{C}\frac{p_i}{2}(\mathbf{m}_i - \mathbf{m})^T\mathbf{S}_i^{-1}(\mathbf{m}_i - \mathbf{m})\Big\}, \quad (3)$$

where $\mathbf{Q} = \sum_{i=1}^{C} p_i\mathbf{S}_i^{-1}$, and $\mathbf{q} = \sum_{i=1}^{C} p_i\mathbf{S}_i^{-1}(\mathbf{m}_i - \mathbf{m})$.

Minimization of ρ_w is equivalent to maximization of the separability measure $-2\log\rho_w = tr(\mathbf{S}_M)$, where

$$\mathbf{S}_M = \sum_{i=1}^{C} p_i\log\mathbf{S}_i + \log\mathbf{Q} - \mathbf{Q}^{-\frac{1}{2}}\mathbf{q}\mathbf{q}^T\mathbf{Q}^{-\frac{1}{2}} + \sum_{i=1}^{C} p_i\mathbf{S}_i^{-\frac{1}{2}}(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i^T - \mathbf{m})\mathbf{S}_i^{-\frac{1}{2}}. \quad (4)$$

Therefore, following a non-iterative procedure similar to LDA, we will define the Matusita criterion as maximization of the projected separability measure $tr\{(\mathbf{T}\mathbf{T}^T)^{-1}\mathbf{T}\mathbf{S}_M\mathbf{T}^T\}$. Linear feature extractor $\mathbf{T}_{d\times n}$ maximizes this criterion if and only if its rows are the d eigenvectors of $\mathbf{S}_M$ with the largest corresponding eigenvalues. Such an operator is the proposed Matusita feature extractor.

It is noteworthy that for data with arbitrary $\mathbf{S}_W$, we first need to normalize the data and parameters by $\mathbf{S}_W^{-1/2}$ to achieve $\mathbf{S}_W = \mathbf{I}$, and then find the linear operator from eigendecomposition of $\mathbf{S}_M$. The Matusita feature extractor will be $\mathbf{T}\mathbf{S}_W^{-1/2}$ in this general case.

Comparison of (2) and (4) reveals that the Chernoff method requires computations of the order $O(C(C-1))$, whereas the complexity of the Matusita method is $O(C)$. This efficiency is achieved through the inherently multi-class formulation of Matusita's measure which avoids the more demanding pairwise formulation.

In the following, we will compare the performance of the proposed Matusita method with that of LDA and its Chernoff-based heteroscedastic extension. All these methods rely on non-iterative procedures on the first two moments of the data.

4. EXPERIMENT ON SYNTHETIC DATA

In this experiment, the data are generated exactly according to the assumed Gaussian distributions $\mathcal{N}(\mathbf{m}_i, \mathbf{S}_i)$. Furthermore, the exact parameters $\mathbf{m}_i$ and $\mathbf{S}_i$ are used for the design of the feature extractors and the classifier. As a result, this experiment excludes the effect of both the deviation of the data from the implied model, and any parameter estimation error.

In each of 100 iterations, $C = 6$ Gaussian-distributed classes are selected. Each class mean $\mathbf{m}_i$ is selected as a uniformly distributed random vector in the unit cube centered at the origin in $\mathbb{R}^{50}$. Each covariance $\mathbf{S}_i$ is selected as $\sigma(\mathbf{A}\mathbf{A}^T + \mathbf{\Gamma}_i\mathbf{\Gamma}_i^T)$. The scaling factor $\sigma = 100$ affects the total variance. The element $\mathbf{A}$, common to covariances of all classes, is a 50×50 random matrix with all its entries selected uniformly from the interval $[-0.5, 0.5]$. The class-specific constituent $\mathbf{\Gamma}_i$ is a 50×50 random matrix with the entries of the 20×20 block on the top-left corner selected like the entries of $\mathbf{A}$, and other entries as zero. The resulting covariance matrices differ only in 20 dimensions. We will use this fact to examine minimal sufficiency of the different methods as described in [10]. In each of 100 iterations, a specific realization of these random class parameters are selected and used to generate 100 testing samples per class.

For each feature extractor and in each iteration, the linear operator is calculated by plugging in the exact $\mathbf{m}_i$ and $\mathbf{S}_i$ values. The resulting operator is applied on each of the corresponding testing samples, and the most prominent extracted features are used by a quadratic classifier based on $\mathbf{m}_i$ and $\mathbf{S}_i$ to determine the corresponding class. The classification error rate for each feature extractor over all the testing samples is calculated, and is averaged over all the 100 iterations.

The average error rate for each algorithm is plotted in Fig. 1 versus the number of extracted features used by the classifier. Also, the

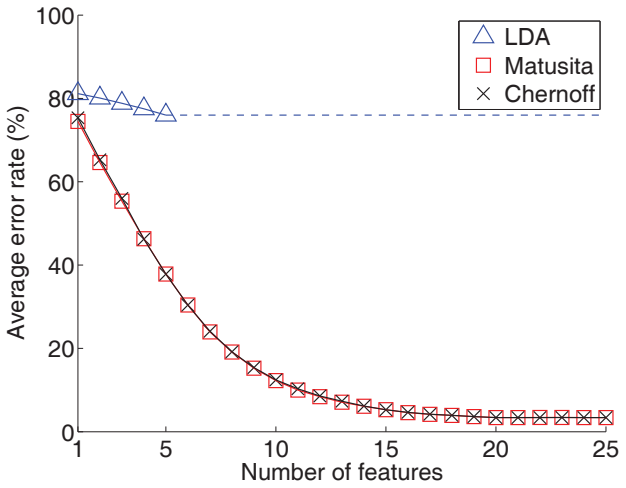

Fig. 1: Average error rate of the quadratic classifier with different feature extractors using actual $\mathbf{m}_i$ and $\mathbf{S}_i$ parameters for six heteroscedastic Gaussian distributions. This experiment does not include the effect of parameter estimation errors.

Method	LDA	Matusita	Chernoff
t (ms)	5.1	343.4	1665.1

Table 1: Average CPU time in milliseconds (t) required to calculate the feature extractors based on synthetic parameters.

required CPU processing time, calculated for the deployed workstation with 3GHz quad core Intel processor and 4GB RAM, is reported in Table 1.

From Fig. 1, it is evident that the heteroscedastic methods have been able to exploit the second order discriminant information of the data to improve LDA's performance. Although the performance of the Chernoff and Matusita methods are quite close to each other in this case, Matusita outperforms the Chernoff method in the presence of estimation errors as shown in the next section.

Furthermore, the complexity of the Matusita method is lower than that of the Chernoff method by an order of $C-1$. The simulation times in Table 1 verify this relationship. Specifically, the ratio of the run time for Chernoff versus Matusita is 4.85 which approximately equals $C-1=5$. Thus, the proposed Matusita method is asymptotically $C-1$ times faster than the Chernoff method. Such a difference will be critical especially for applications with a large number of classes.

Fig. 1 also demonstrates that both the Chernoff and Matusita provide at most 25 features in this case, and hence avoid any redundant features beyond 25 which is the minimum number of sufficient features [10]. Furthermore, both methods have provided the minimal average error rate at this dimension. Therefore, both methods seem to provide a minimum-dimensional linear sufficient statistic. This property of the Chernoff and Matusita methods in practice is the target of future investigation.

5. EXPERIMENTS ON UCI DATA SETS

In this section, real world data are classified into a predetermined set of classes. Therefore, neither the accuracy of the parameters, nor the Gaussianity of the data is guaranteed. Thus, the results incorporate the tolerance of different feature extractors to parameter estimation error and deviation of the data from Gaussianity.

Data set name	C	N	n	d_{PCA}	J_H
(a) Wisconsin breast cancer	2	683	9	9	318.10
(b) BUPA liver disorder	2	345	6	6	7.63
(c) Ionosphere	2	351	34	33	37.71
(d) Iris plants	3	150	4	4	14.06
(e) Thyroid gland	3	215	5	5	93.69
(f) Glass identification	6	214	9	8	43.17
(g) Image segmentation	7	2310	19	14	1,266.04

Table 2: Specifications of the UCI data sets: total number of classes (C), total number of samples (N), original data dimensionality (n), data dimensionality after PCA (d_{PCA}), and multivariate heteroscedasticity measure (J_H).

Fig. 2: Experimental setup for UCI experiments.

The data sets are selected from a subset of University of California, Irvine (UCI) machine learning repository [17] designed for the classification purposes. The specifications of the data sets are outlined in Table 2. This table also includes our calculated heteroscedasticity score based on [18] for each data set. This multivariate test is designed to provide a reliable score even in a small sample size scenario. With a 0.05 significant level, any J_H score higher than 1.64 indicates heteroscedasticity. It can be seen that all the data sets are detected as heteroscedastic.

There are two differences in the experimental setup compared to Section 4. First, the data mean and covariance for each class need to be estimated using a set of training samples. We have used a random 90% subset of the data set for training and the remaining 10% for testing. This random partition is repeated 100 times to alleviate any possible bias in the results. The second difference in the setup arises from the small sample size in some of the data sets, which may lead to singularity in the estimated covariances. Both the Chernoff and Matusita methods, as well as the quadratic classifier, assume a non-singular covariance estimate. Thus, to ensure covariance non-singularity, we precede each feature extractor with a principal component analysis (PCA) step and a regularization step as depicted in Fig. 2. Using PCA, the data are projected into d_{PCA}-dimensional space (ref. Table 2), where d_{PCA} is the highest dimension to ensure that the average class covariance is non-singular. Then, each of the class covariances is slightly regularized with a small regularization parameter of 0.001 toward the average class covariance [19].

Each feature extractor, along with the quadratic classifier, is trained and tested on the transformed data. The resulting error rate for every possible number of extracted features d is calculated, and is averaged over all 100 random partitions of the data set. The minimum average error rate over different d values is reported in Table 3 for each method. The corresponding optimal d denoted by d_o, and the average required processing time for the training phase of each feature extractor is also reported in this table. Furthermore, for each data set, corresponding to a row of the table, the significantly superior error rate performance is boldfaced if it exists. The significance is decided according to the signed ranked test [20] with a significance level of 0.01.

From Table 3, the proposed Matusita method provides an overall performance improvement compared to the Chernoff method. Ma-

DS	LDA			Matusita			Chernoff		
	%AER	d_o	t (ms)	%AER	d_o	t (ms)	%AER	d_o	t (ms)
(a)	2.71	1	0.64	2.49	1	6.85	2.62	1	5.14
(b)	**37.20**	1	0.50	37.64	1	4.30	37.52	1	3.48
(c)	13.26	1	3.93	**8.14**	3	51.12	13.50	3	29.94
(d)	2.20	1	0.44	1.66	1	3.78	1.93	1	6.00
(e)	3.28	1	0.47	3.04	4	4.94	3.00	4	8.62
(f)	**39.94**	2	0.64	41.66	2	15.29	42.55	7	83.10
(g)	8.01	6	0.96	**5.91**	8	31.60	6.27	8	215.85

Table 3: The best percentage average error rate (%AER), the corresponding optimal dimension (d_o), and the CPU time required for training (t) for different methods on UCI data sets.

tusita provides the lowest error rate with a significant margin for data sets (c) and (g), and provides minimum or close to minimum error rates in other rows. In data sets (f) and (g) with relatively larger number of classes, Matusita has outperformed Chernoff. Considering these results in conjunction with the results of Section 4, they indicate that compared to the Chernoff method, the Matusita method provides an improved tolerance against inaccuracies in the Gaussianity and the estimated parameters of the data.

In data sets (b) and (f), the LDA's superior performance can be attributed to respectively low heteroscedasticity ($J_H = 7.63$) and low number of samples per class ($214/6 \approx 35.7$) from Table 2. A small sample size leads to inaccuracy in the estimated covariances used by heteroscedastic methods.

Furthermore, from Tables 2 and 3, the Matusita method is computationally faster than the Chernoff method by a factor of almost $C - 1$ for each data set, although it is slightly slower than the Chernoff method when $C = 2$. This efficiency was achieved by deploying an inherently multi-class separability measure rather than a pairwise extension of a two-class measure.

6. CONCLUSIONS

This paper provided a heteroscedastic extension of LDA based on the Matusita separability measure. This measure is preferred over the previously used pairwise Chernoff distance due to its original multi-class formulation, which is related to multi-class error bounds [14] and is computationally more efficient than the Chernoff criterion by a factor of $C - 1$. A non-iterative approach similar to LDA and the Chernoff method was proposed for minimization of the Matusita criterion. Finally, the simulation results verified improvements in both computational efficiency and tolerance to the parameter estimation errors.

7. REFERENCES

[1] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, Wiley-Interscience, second edition, 2000.

[2] A.K. Jain, R.P.W. Duin, and J. Mao, "Statistical pattern recognition: a review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[3] O. Hamsici and A. Martinez, "Bayes optimality in linear discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, 2008.

[4] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Commun.*, vol. 26, no. 4, pp. 283 – 297, 1998.

[5] L. Rueda and M. Herrera, "Linear dimensionality reduction by maximizing the Chernoff distance in the transformed space," *Pattern Recognition*, vol. 41, no. 10, pp. 3138 – 3152, 2008.

[6] L. Rueda, B. J. Oommen, and C. Henríquez, "Multi-class pairwise linear dimensionality reduction using heteroscedastic schemes," *Pattern Recognition*, vol. 43, no. 7, pp. 2456 – 2465, 2010.

[7] K. Das and Z. Nenadic, "Approximate information discriminant analysis: A computationally simple heteroscedastic feature extraction technique," *Pattern Recognition*, vol. 41, no. 5, pp. 1565–1574, 2008.

[8] H. Brunzell and J. Eriksson, "Feature reduction for classification of multidimensional data," *Pattern Recognition*, vol. 33, no. 10, pp. 1741 – 1748, 2000.

[9] M. Mahanta and K. Plataniotis, "Linear feature extraction using sufficient statistic," in *Acoustics Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2010, pp. 2218 –2221.

[10] M. S. Mahanta, A. S. Aghaei, K. N. Plataniotis, and S. Pasupathy, "Heteroscedastic linear feature extraction based on sufficiency conditions," *Pattern Recognition*, vol. 45, no. 2, pp. 821 – 830, 2012.

[11] R.P.W. Duin and M. Loog, "Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp. 732–739, June 2004.

[12] M. Thangavelu and R. Raich, "Multiclass linear dimension reduction via a generalized Chernoff bound," in *Machine Learning for Signal Processing (MLSP), IEEE Workshop on*, oct. 2008, pp. 350 –355.

[13] K. Matusita, "On the notion of affinity of several distributions and some of its applications," *Annals of the Institute of Statistical Mathematics*, vol. 19, pp. 181–192, 1967.

[14] B. Bhattacharya and G. Toussaint, "An upper bound on the probability of misclassification in terms of Matusita's measure of affinity," *Annals of the Institute of Statistical Mathematics*, vol. 34, pp. 161–165, 1982.

[15] J. Brinkhuis, Z. quan Luo, and S. Zhang, "Matrix convex functions with applications to weighted centers for semidefinite programming," Tech. Rep., 2005.

[16] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press Professional Inc., San Diego, CA, USA, second edition, 1990.

[17] A. Asuncion and D. Newman, "UCI machine learning repository," at http://www.ics.uci.edu/∼mlearn/MLRepository.html, 2007.

[18] J. R. Schott, "A test for the equality of covariance matrices when the dimension is large relative to the sample sizes," *Computational Statistics and Data Analysis*, vol. 51, no. 12, pp. 6535–6542, 2007.

[19] F. van der Heijden, R. Duin, D. de Ridder, and D.M.J. Tax, *Classification, Parameter Estimation and State Estimation: An Engineering Approach Using MATLAB*, John Wiley & Sons, Nov. 2004.

[20] J. Rice, *Mathematical Statistics and Data Analysis (Statistics)*, Duxbury Press, June 1994.