Sparse Multinomial Kernel Discriminant Analysis (sMKDA)

Robert F Harrison [a,*]

[a] *Department of Automatic Control & Systems Engineering, The University of Sheffield, Mappin Street, Sheffield, England, S1 3JD, UK.*

Kitsuchart Pasupa [b]

[b] *School of Electronics & Computer Science, The University of Southampton, Southampton, England, SO17 1BJ, UK.*

Abstract

Dimensionality reduction via canonical variate analysis (CVA) is important for pattern recognition and has been extended variously to permit more flexibility, e.g. by "kernelizing" the formulation. This can lead to over-fitting, usually ameliorated by regularization. Here, a method for sparse, multinomial kernel discriminant analysis (sMKDA) is proposed, using a sparse basis to control complexity. It is based on the connection between CVA and least-squares, and uses forward selection via orthogonal least-squares to approximate a basis, generalizing a similar approach for binomial problems. Classification can be performed directly via minimum Mahalanobis distance in the canonical variates. sMKDA achieves state-of-the-art performance in terms of accuracy and sparseness on 11 benchmark datasets.

Key words: linear discriminant analysis, kernel discriminant analysis, multi-class, multinomial, least-squares, optimal scaling, sparsity control.

1 Introduction

Dimensionality reduction is an important step in pattern recognition, classification and data visualization where data may exist in high-dimensional feature spaces, and linear discriminant analysis (LDA) has played a central

* Corresponding Author
Email address: r.f.harrison@sheffield.ac.uk (Robert F Harrison).

role in achieving this for over 70 years. LDA seeks a linear projection that maximizes the separation between data belonging to each of $c \geq 2$ classes while minimizing the separation of those belonging to the same class. The resulting co-ordinate frame – the canonical variates – has well-documented properties and, under certain circumstances, the associated discriminant functions prove optimal for classification [1] while performing well in many realistic situations. Nonetheless, in many cases of practical interest, recognition performance can be substantially enhanced by allowing more flexible, non-linear, discriminant functions. While explicit expansion of data in basis functions can resolve this for problems of low dimension, the combinatorial increase in the number of coefficients to be estimated may make this impractical in general. Over the past decade a large body of theory based on reproducing kernel Hilbert spaces has given a new perspective on many non-linear problems in machine learning [e.g. 2, 3] and much work has been carried out in generalizing LDA within this framework. One key advantage of these kernel methods over explicit expansions is that they avoid the need to work explicitly in very high, possibly infinite, dimensional feature spaces, instead leading to problems whose "size" is bounded by the sample size, n. This may, itself bring problems when n is large. In addition, kernels can be defined that deal with much more general data types than those that are simply represented in a vector of numbers, e.g. sequences, trees, graphs and more general data still.

The problem of generalizing LDA to provide more flexible discrimination has received much attention and a plethora of solutions has been presented, some of which address the multinomial problem [e.g. 4–11], while others consider only the binomial ($c = 2$) case [e.g. 12–18]. Interestingly, while a key feature of the most widespread kernel machine – the Support Vector Machine – is its sparse nature, most of these generalizations pay no attention to this point. Sparsity control is of particular importance in kernel-based formulations which, in their most straightforward implementations, depend on the entire sample and can, therefore, strongly violate the principle of parsimony. Such control is also necessary to ameliorate tendency to over-specialization, can improve numerical conditioning and, of course, can reduce computational burden in operation – of particular interest in real-time applications. We focus attention on the question of sparsity in the literature on kernelized discriminant analysis for reasons of space. Of the approaches mentioned above, only [10, 12, 15, 17–19] address, explicitly, the question of sparsity control and, of these, only [10, 17] address the multinomial case. However, it is not clear that the latter technique can be used other than in a "one-vs-all" (OVA) or "all-vs-all" (AVA) strategy. We seek to address the multinomial problem directly while achieving sparsity in the resulting classifier. In doing so we aim to avoid the following problems associated with the use of binomial classifiers for multinomial tasks.

The first is the need to train c or $\frac{1}{2}c(c-1)$ individual classifiers, depending on

choice of strategy. However, in a "train-once, use-many" task, this may not be of particular concern for moderate c and the associated clock-time would be highly dependent on the particular choice of classification engine anyway. Nonetheless, considerable saving can be made when, as is the case here, the dominant computation only has to be performed once. An associated training issue relates to the balance between the classes. Even for an equiprobable class distribution, the OVA approach leads to severe imbalance for large c. The question of how best to handle this therefore arises, e.g. via stratification. A second problem arises in operation, when a number of classifiers must be run and their results analyzed to provide a predicted class value. A number of methods (see [20] for a critical overview) has been championed for the post-processing, each of which is claimed to have merit, although in [20] it is argued that OVA performs equally as well as any of these. In general, while the overhead of running numerous classifiers may not be an issue, for kernel machines involving, potentially, the entire sample, it may be unacceptable so that achieving sparsity can be operationally important. To be clear, the idea of sparseness is related to how few sample instances need to be retained to form an adequate basis for the underlying feature space or, equivalently, to compute the necessary kernels during operation. An additional motivation for solving the sparse multinomial problem directly, without recourse to multi-classifiers lies in the fact that, even if each binomial classifier is itself sparse, the union of the sets of retained samples across all classifiers may not be. A further motivator for the use of LDA, rather than, say, a direct application of multi-response least-squares, arises from a problem highlighted in [21, §5.1]: if class centroids happen to be nearly collinear in feature space a catastrophic collapse of the associated discriminant functions can occur leading to poor performance. This has been called the "masking" problem.

The objective of this paper is to generalize the method devised by Billings and Lee [12] that made use of the fact that, in the binomial situation, the canonical variate directions are available form a simple least-squares formulation. This is attractive because a widely used and effective method of sparsity control in the from of a forward selection procedure based on the modified Gram-Schmidt procedure – orthogonal least-squares (OLS) – can be applied [22]. To extend the method to the multinomial situation ($c > 2$) we were inspired, initially, by a result of Crownover [23] that permits a two-stage procedure for finding the directions of the canonical variates involving a least-squares problem followed by a small, easily managed, eigen-decomposition of order $\mathcal{O}(c)$. This contrasts with some approaches [4, 11] which involve large eigen-decompositions of order $\mathcal{O}(n)$. Our intention was then to apply a regressor selection technique such as OLS to the least-squares stage. Subsequently, a more convenient formulation [21] based on penalized optimal scoring was shown via canonical correlation analysis to be able to provide *exactly* the canonical variates [24], i.e. identical direction and scaling to LDA. This facilitates the computation of Mahalanobis distance in canonical variates leading to a simple method for

computing classification performance and hence speeding the cross-validation and operational stages. Roth and Steinhage [8] developed a kernelized version of the method of [21] but did not attempt to control sparsity thereby resulting in a classifier that requires access to the full sample in operation.

In §2 the method of optimal scoring for discriminant analysis is is outlined and re-expressed in dual co-ordinates preparatory to "kernelization" in §2.2. The method of orthogonal least-squares is described in §2.3 and applied to the formulation of §2.2 leading to the sparse multinomial kernel discriminant analysis formulation. Section four examines the performance of the proposed algorithm on nine benchmark multinomial classification problems and compares this with the best results on these same benchmarks gleaned from 11 recently published articles in the machine learning literature and a smaller comparison with classifiers specifically designed to be sparse. A further three-class problem is examined in more detail including a visualization demonstrating, graphically, the value of sMKDA in this domain. All results are comparable with the current state-of-the-art in terms of accuracy and sparsity (where the comparison is possible) [1] .

2 Discriminant Analysis via Optimal Scoring

The idea that LDA might be affected through linear regression is a natural one since the key component is a linear transformation of the data onto a space of lower dimension that satisfies certain conditions – maximal separation between the classes in this case. In the binomial case the link between the least-squares problem and Fisher's discriminant function is well-known and explicit [e.g. 1] provided that class membership is coded appropriately. The extension to more than two classes is less straightforward. What seemed obvious in the binomial case – a linear regression onto quantitative values that indicate membership of one class or not – does not translate directly to the multinomial situation. The multivariate regression onto, say, a binary-valued indicator matrix representing membership does not, in general, determine the canonical variates and performing discrimination in the resulting space can lead to very poor classification performance under certain circumstances (masking) [21]. The method of optimal scoring, which arises from canonical correlation analysis of categorical variables, provides an answer to the question of how best to code classes by assigning, optimally, real, quantitative values to category labels. This leads to a two-stage optimization process, a linear regression followed by a low-dimensional eigen-decomposition. The details of this are well described in [21] so we simply sketch the steps of the derivation needed in the sequel.

[1] Sparsity, if any, is not always reported in the articles reviewed.

For notational ease, we assume at this stage that the data are "centered", i.e. that they have had their mean value removed and we ignore, for the moment, the question of regularization.

Consider a sample from $\mathbb{R}^d$ of size n, categorized into c groups each containing $n_i, i = 1...c$ vectors. The task is to determine a projection of the data into a space of lower dimension that maximizes, in some sense, the separation of the groups. Denote the $(n \times d)$-dimensional data matrix by: $\mathrm{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]^T$ and the $(n \times c)$-dimensional indicator matrix, Y indicate group membership such that y_{ij} indicates that instance $\boldsymbol{x}_i$ belongs to group j. The elements of the c-dimensional vector, $\boldsymbol{\theta}$, assign numerical scores to each of the groups so that the vector $\mathrm{Y}\boldsymbol{\theta}$ represents the "scored" outcomes. Then in [21] it is shown that the canonical variates can be obtained by finding the set of vectors, $\boldsymbol{\theta}_i$, $\boldsymbol{\beta}_i$ that minimize the average-squared-residual, ASR,

$$ASR\left(\theta, \beta\right) = \frac{1}{n} \left\| Y\boldsymbol{\theta} - X\boldsymbol{\beta} \right\|_2^2 \tag{1}$$

subject to the normalization $\frac{1}{n} \left\| \mathrm{Y}\boldsymbol{\theta} \right\|_2^2 = 1$, and re-scaling.

The minimization can be achieved through a two-stage process, first by considering $\boldsymbol{\theta}$ as fixed and solving the unconstrained minimization *w.r.t.* $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = \left(\mathrm{X}^\mathsf{T}\mathrm{X}\right)^{-1} \mathrm{X}^\mathsf{T}\mathrm{Y}\boldsymbol{\theta} \tag{2}$$

assuming X has full-rank. It then remains to optimize w.r.t. θ.

Inserting the RHS of equation (2) into equation (1) and applying the constraint, $\frac{1}{n} \left\| \mathrm{Y}\boldsymbol{\theta} \right\|_2^2 = 1$, gives the partially optimized ASR:

$$ASR\left(\boldsymbol{\theta}\right) = 1 - \frac{1}{n} \left(\boldsymbol{\theta}^\mathsf{T}\mathrm{Y}^\mathsf{T}\mathrm{H}\mathrm{Y}\boldsymbol{\theta} \right) \tag{3}$$

where $\mathrm{H} = \mathrm{X} \left(\mathrm{X}^\mathsf{T}\mathrm{X}\right)^{-1} \mathrm{X}^\mathsf{T}$ denotes the "hat" matrix arising from the regression. To minimize equation (3) *w.r.t.* $\boldsymbol{\theta}$ the constraint is attached through the Lagrange multiplier, μ, giving

$$1 - \frac{1}{n} \left(\boldsymbol{\theta}^\mathsf{T}\mathrm{Y}^\mathsf{T}\mathrm{H}\mathrm{Y}\boldsymbol{\theta} + \mu\boldsymbol{\theta}^\mathsf{T}\mathrm{Y}^\mathsf{T}\mathrm{Y}\boldsymbol{\theta} \right) \tag{4}$$

and by taking its gradient and equating to zero the following generalized eigenproblem is obtained:

$$\mathrm{Y}^{\mathsf{T}}\mathrm{H}\mathrm{Y}\boldsymbol{\theta} = \mu\mathrm{Y}^{\mathsf{T}}\mathrm{Y}\boldsymbol{\theta} \tag{5}$$

We express the result in this explicit form simply to establish that it is indeed identical to that of Crownover [23, equation (12)]. While the eigen-vectors obtained from solving equation (5) have the same directions as the LDA solution, they are scaled differently and this is where the results of Crownover [23] and of Hastie et al. [21] diverge.

The overall procedure can be expressed slightly differently, avoiding the, albeit trivial, *generalized* eigen-problem, as follows [21]:

(1) Choose a matrix Θ_0 that satisfies the constraint, $\frac{1}{n}\Theta_0^{\mathsf{T}}\mathrm{Y}^{\mathsf{T}}\mathrm{Y}\Theta_0 = I$, e.g. $\Theta_0 = \mathrm{diag}\left\{\frac{1}{\sqrt{n_1}}, \frac{1}{\sqrt{n_2}}, \ldots, \frac{1}{\sqrt{n_c}}\right\}$ and let $\mathrm{Y}_0 = \mathrm{Y}\Theta_0$.
(2) Carry out a multivariable (multi-response) regression on Y_0 and compute $\hat{\mathrm{Y}}_0 = \mathrm{X}\mathrm{B}$, where $\mathrm{B} = \left(\mathrm{X}^{\mathsf{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathsf{T}}\mathrm{Y}_0$.
(3) Compute the eigen-decomposition of $\mathrm{Y}_0^{\mathsf{T}}\hat{\mathrm{Y}}_0$ to obtain W, the matrix of eigen-vectors corresponding to the $k \leq (c-1)$ largest eigen-values [2], λ_i, $i = 1, 2, \ldots, (c-1)$, in descending order.
(4) Adjust the matrix of regression coefficients thus: $\mathrm{B}^{\mathrm{os}} \leftarrow \mathrm{B}\mathrm{W}$.

The matrix, B^{os}, is shown [21, 24] to be identical to the canonical variates up to a diagonal scaling, $D = \mathrm{diag}\left\{\frac{1}{\sqrt{\lambda_i(1-\lambda_i)}}\right\}$, where the λ_i, $i = 1, 2, \ldots (c-1)$ are sorted into descending order, giving

$$\mathrm{B}^{\mathrm{cv}} = \mathrm{B}^{\mathrm{os}}\mathrm{D} = \mathrm{B}\mathrm{W}\mathrm{D} \tag{6}$$

This leads to an expression for the expression of a new datum in canonical variates, thus:

$$\boldsymbol{y}^{\mathrm{cv}} = \boldsymbol{x}^{\mathsf{T}}\mathrm{B}^{\mathrm{cv}} \tag{7}$$

2.1.1 A Dual Formulation

Having established the basic idea we first note that, for a *sparse* kernelized version, it is imperative to avoid the "centering" operation since this is an inherently non-sparse procedure in dual co-ordinates involving the entire sample for every new instance. Since our ultimate aim is classification and/or visualization, the absolute position of the sample in feature space is of little interest and so an explicit offset will be maintained by augmenting the data matrix

[2] A single, zero eigen-value will always exist owing to the centering of X.

with a column of n ones in the usual fashion and extending the row dimension of the coefficient matrix, B, from d to $(d+1)$. Therefore, the trivial eigenvalue found above is changed to a unit eigen-value which, along with its associated eigen-vector, must be excluded before searching for the $c-1$ largest.

In anticipation of our final objective, we re-formulate the results so that they depend only upon inner-products between sample vectors. This is achieved by noting that the coefficient matrix, B can itself be expressed as a linear combination of the data matrix, X, i.e. $\mathrm{B} = \mathrm{X}^{\mathsf{T}}\mathrm{A}$, $\dim \mathrm{A} = (n \times c)$. Substitution for B and augmenting the model to incorporate an explicit offset leads to the multivariable regression problem of minimizing:

$$\left\| \mathrm{Y}_0 - \begin{bmatrix} \mathrm{X} & \mathbf{1}_N \end{bmatrix} \begin{bmatrix} \mathrm{X}^{\mathsf{T}}\mathrm{A} \\ \boldsymbol{\alpha}^{\mathsf{T}} \end{bmatrix} \right\|_F^2 = \left\| \mathrm{Y}_0 - \begin{bmatrix} \mathrm{X}\mathrm{X}^{\mathsf{T}} & \mathbf{1}_N \end{bmatrix} \begin{bmatrix} \mathrm{A} \\ \boldsymbol{\alpha}^{\mathsf{T}} \end{bmatrix} \right\|_F^2 \tag{8}$$

with respect to $\boldsymbol{\alpha}$ and A, where $\|\cdot\|$ indicates the Frobenius norm, and $\boldsymbol{\alpha}^{\mathsf{T}}$ denotes the c-dimensional (row) vector of constant offsets. While such a formulation requires the solution of a degenerate N-dimensional system with, typically, $N \gg d$, we shall not use it in its present form and defer the issue for later.

2.2 Multinomial Kernel Discriminant Analysis (MKDA)

It is well-known that linear discriminants are likely to prove suboptimal in many cases of practical interest. Indeed, it is only when the within class samples are normally distributed with common covariance matrices that the linear development is optimal. Rather than appeal to a parametric form we permit flexibility in the solution through a linear expansion in basis functions so that the discriminant analysis is carried out in the space of basis functions rather than in the original "measurement" space. Nonetheless, Hastie et al. [21] appeal to the fact that canonical variates obtained in this way are themselves sums of random variables which will have the effect of making the regressors appear more normally distributed.

Consider the mapping $\boldsymbol{f} : \mathbb{R}^d \rightarrow \mathcal{F}$ – the feature (inner-product) space – whose evaluations at the data samples are $\boldsymbol{f}_i = \boldsymbol{f}(\boldsymbol{x}_i)$, $i = 1, 2, \ldots, n$ with $\dim(\boldsymbol{f}_i) = \nu$ – a large, possibly infinite, value. This yields a new data matrix, $\mathrm{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n]^T$. Direct determination of the projection vectors using the explicitly mapped data involves solving a ν-dimensional least-squares problem. Commonly, $\nu > n$ so the dual formulation of § 2.1.1 offers a potential solution provided that the inner products, $\langle \boldsymbol{f}_i, \boldsymbol{f}_j \rangle$ can be computed efficiently. This "kernel trick" has been widely examined elsewhere [e.g. 2, 3] so we do not

dwell on the technical details here.

Inner products between a wide class of functions are amenable to straightforward evaluation through the use of a reproducing kernel. Conversely, kernel functions possessing certain properties correspond to inner products in some feature space even if the explicit mapping, $\boldsymbol{f}$, remains obscure. For example, imagine a polynomial expansion of degree p of the original d-dimensional data. This leads to feature space dimension of $\nu = \frac{(d+p)!}{d!p!}$, however, inner products in this space can be carried out via the polynomial kernel, $k_{\text{poly}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(1 + \boldsymbol{x_i}^T \boldsymbol{x_j}\right)^p$. On the other hand, application of the gaussian radial basis function kernel of "width", p, $k_{\text{grbf}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(\frac{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{2p^2}\right)$, computes the inner product in an infinite dimensional ($\nu = \infty$) feature space spanned by the eigen-functions of k_{grbf}. Many additional kernel functions can be found in [e.g. 2].

The kernel (Gram) matrix, K, with elements, $k_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $i, j = 1, 2, \ldots, n$, is implicitly given by $\mathrm{K} = \mathrm{F}\mathrm{F}^T$ thus replacement of $\mathrm{X}\mathrm{X}^T$ by K in equation (8) and application of the algorithm, using a suitable pseudo-inverse, permits the computation of linear discriminants in a feature space of arbitrary complexity and to project the data in a non-linear fashion onto $\mathbb{R}^\delta$, $1 \leq \delta \leq c - 1$ for visualization or decision making. However, it is likely that in the current formulation, severe over-fitting will occur leading to poor generalization.

2.2.1 Ridge Penalty

It is common to introduce a ridge-type penalty to combat ill-conditioning and to induce smoothness (or reduce complexity) in the implied mapping, $\boldsymbol{f}$. This yields the following least-squares problem:

$$\arg\min_{\mathrm{A}, \boldsymbol{\alpha}} \left\| \begin{bmatrix} \mathrm{Y}_0 \\ 0_{n \times c} \end{bmatrix} - \begin{bmatrix} \mathrm{K} & \mathbf{1}_n \\ \mathrm{R}^{\frac{1}{2}} & 0_{n \times 1} \end{bmatrix} \begin{bmatrix} \mathrm{A} \\ \boldsymbol{\alpha}^\mathsf{T} \end{bmatrix} \right\|_F^2 \tag{9}$$

where R is an $n \times n$-dimensional, positive semi-definite matrix. NB we do not penalize the offset to avoid distorting the estimated outcomes [e.g. 25, p. 59].

Denoting the augmented regressor matrix as $\tilde{\mathrm{K}} = \begin{bmatrix} \mathrm{K} & \mathbf{1}_n \end{bmatrix}$, the optimal coefficient matrix, $\begin{bmatrix} \hat{\mathrm{A}}^\mathsf{T} & \hat{\boldsymbol{\alpha}} \end{bmatrix}^\mathsf{T}$, as B and the penalty matrix, $\begin{bmatrix} \mathrm{R} & 0_{n \times 1} \\ 0_{1 \times n} & 0 \end{bmatrix}$, as $\tilde{\mathrm{R}}$, then the resulting modification to step (2) of the algorithm in § 2.1 is:

$$\mathrm{B} = \left(\tilde{\mathrm{K}}^{\mathsf{T}}\tilde{\mathrm{K}} + \tilde{\mathrm{R}}\right)^{-1} \tilde{\mathrm{K}}^{\mathsf{T}}\mathrm{Y}_0 \tag{10}$$

which corresponds to the result of Roth and Steinhage [8]. All other steps in the algorithm and the final scaling follow as before.

While regularization in this way addresses the problem of overspecialization, an, $(n+1) \times (n+1)$, least-squares problem must be solved. In cases where n is large this may be infeasible or undesirable. Besides, sparse models are appealing in the sense that they attempt to fulfil the principle of parsimony (c.f. Occam's Razor) as well as being computationally less demanding in operation.

2.3 Sparse Multinomial Kernel Discriminant Analysis (sMKDA)

A key reason for approaching the MKDA problem from a least-squares perspective is that there exists a well-established method for regressor subset selection known as the orthogonal least-squares (OLS) forward selection algorithm – Billings and Lee [12] have very successfully addressed the binomial kernel Fisher discriminant problem in this way. The forward selection procedure has been developed over many years and has been widely applied in the field of non-linear systems identification. It is based on the observation that the regressor matrix, $\tilde{\mathrm{K}}$, has a decomposition, $\tilde{\mathrm{K}} = \mathrm{PS}$ where P has orthogonal columns and S is upper-triangular with unit diagonal [3]. From this, the proportion of the total sum-of-squared-error (SSE) explained by each of the columns of P can be determined independently of all other columns. This permits a greedy, sequential approach that admits new regressors by choosing the one that maximizes the reduction in SSE, at each stage. The algorithm is terminated when SSE is adequately accounted for or a maximum acceptable number of terms has entered the reduced basis. This is determined by a user-defined threshold, retaining n_{s} terms in the subset. There is a number of ways the orthogonalization can be achieved and we use the one based on the modified Gram-Schmidt procedure [26] for multivariable (multi-response) systems [4]. The resulting $(n_{\mathrm{s}} \times c)$ coefficient matrix, $\mathrm{B}_{\mathrm{ols}}$ now takes the place of B in the algorithm in § 2.1 and a new datum can be expressed in canonical variates, thus:

$$\boldsymbol{y}^{\mathrm{cv}} = \boldsymbol{k}^{\mathsf{T}}\mathrm{B}_{\mathrm{ols}}\mathrm{WD} \tag{11}$$

[3] Here we consider only the matrix, $\tilde{\mathrm{K}}$ but the idea applies equally to the data matrix, $\tilde{\mathrm{X}}$ or to $\tilde{\mathrm{F}}$ if explicitly available.

[4] The algorithm has been presented many times in the past and we see no reason to repeat this here.

where $\boldsymbol{k}^{\mathsf{T}}$ denotes the vector of kernel evaluations, $k(\boldsymbol{x}_i, \boldsymbol{x})$ and i belongs to the index set of the retained samples.

The extension of OLS to a regularized cost function is possible [27, 28] but is only straightforward in the space of orthogonalized regressors where a diagonal penalty can be applied. This means that a direct comparison with, e.g. the result from equation (10) is only possible when $\mathrm{R} \propto \mathrm{S}^{\mathsf{T}}\mathrm{S}$, i.e. if the (diagonal) penalty in orthogonal co-ordinates is Λ then the choice, $\mathrm{R} = \mathrm{S}^{\mathsf{T}}\Lambda\mathrm{S}$, gives an exactly equivalent optimization problem. This is mentioned only for information and is not pursued further. Instead we focus on controlling complexity through the construction of a suitable sparse basis via OLS in conjunction with an appropriate choice of kernel parameter, p.

From the foregoing, it is clear that a means of selecting both the kernel parameter, p, and the sparsity of the basis, n_{s}, is required. In [21] the use of generalized cross-validation is proposed alongside an additional "goal directed" (e.g. classification performance) means of selecting a parameter that influences the rôle of the effective number of degrees-of-freedom. We, instead, adopt a direct, many-fold cross-validation approach ignoring smoothing at the regression stage by using misclassification rate as the figure-of-merit during cross-validation. This is particularly convenient in the current formulation because, as shown in [24], the Mahalanobis distance from class centroids (adjusted for class prior distribution) can be simply and quickly computed in canonical variates by the following expression:

$$\left\| \boldsymbol{y}^{\mathrm{cv}} - \overline{\boldsymbol{y}}_j^{\mathrm{cv}} \right\|^2 - 2\log \hat{p}_j \tag{12}$$

where $\overline{\boldsymbol{y}}_j^{\mathrm{cv}}$ is the mean of the projected training sample for the j^{th} class, i.e. the j^{th} class centroid, and $\hat{p}_j$ is the estimated prior probability of the j^{th} class.

Further convenience arises from during cross-validation because computing the OLS solution for $n_{\mathrm{s}}^{\max}$ and fixed p delivers the solutions for all values of n_{s} up to and including the maximum permitted. The computational overhead for the basic implementation of the OLS algorithm is estimated to be $\mathcal{O}(n_{\mathrm{s}}n)^2$ so, when a high degree of sparsity is either demanded or achievable, a considerable saving in computational effort can be made compared with a full two-dimensional search over p and, e.g. a regularization parameter.

Clearly, the dominant computation arises in the OLS algorithm and for certain tasks, when both n and n_{s} are sizeable, this will be become infeasible as is true of many algorithms. We have computed the *full* OLS solution ($n_{\mathrm{s}} = n$) for a range of sample sizes up to $n = 5000$ and with $n_{\mathrm{s}} = 752$ for a samples of size n=10000, and the CPU times are shown in figure A.1. It is clear from this that after approximately 10% of samples have been processed the time to process an additional sample reduces considerably. This is a consequence of the forward

selection strategy of the algorithm. As an example, the full OLS procedure for a 5000×5000 Gram matrix takes approximately 55 hours. The computations are carried out in Matlab v7.5 running on a 2.6GHz dual-core AMD Opteron processor with 4GB RAM. The major difficulty here is in maintaining the Gram matrix in memory. It is important to note that this is not a limitation of the OLS algorithm, simply one arising from our implementation. While it may take a substantial amount of time to train sMKDA, this is not necessarily a barrier to its adoption in a "train once, use many" scenario providing that the resulting classifier is sufficiently compact and fast in operation.

3 Experimental Results

For illustration of our method we examine its performance first on a widely studied set of multinomial benchmark problems followed by a more challenging task both in size and distribution. Here we exploit the simplicity of the minimum Mahalanobis distance from class centroids for all decision making, i.e. during cross-validation and test phases. We note, though, that this may not provide the best possible performance for any given sample.

3.1 UCI Repository Benchmarks

Here the method described above is applied to nine datasets selected from the UCI repository of machine learning databases [29]. The results are compared with those of the best performing method on the same sets as reported in 11 recently published articles [30–40]. Because there is no standard for experimentation across these articles we have adopted the following procedure. First, instances containing missing data are simply excised from each set and each variable is normalized to zero mean and unit variance. Each set is then randomly divided into a training and a testing sample five times, to ameliorate sample bias. Results are then averaged across these replications. Prior distribution of classes is maintained (see Table A.5 in the Appendix). All experiments use the gaussian radial basis function kernel and five-fold cross-validation is used to estimate the optimal kernel parameter, p, and the number of samples retained by the OLS procedure, n_{s}. To ensure numerical stability, a small diagonal penalty, (10^{-9}), is applied in the orthogonal co-ordinates (excluding the offset). Finally, classification is achieved by assigning a sample to the class corresponding to the nearest projected class centroid as per equation(12). The experimental results are shown in Table A.1. The table compares, for each of the nine datasets, the current published best results (based on overall accuracy) and those of sMKDA. For each dataset, two rows are presented, the upper containing the overall percentage accuracy (proportion of correctly

assigned samples), the lower, the percentage retained samples (the number of samples required to implement the model) which is the complement of sparsity. Sparsity is not reported in [40] or in [34] (Dermatology, Thyroid and Vowel) and is not applicable in the case of LDA (Zoo, [37]). We have reported 100% retained samples for KNN and counted it in the overall average even though KNN is not a "sparse" technique. In summary, sMKDA achieves greater accuracy than the best reported in five out of nine cases, is marginally worse in three more and has an approximately 3.5% shortfall in the Glass benchmark. This latter can probably be explained by the extreme level of sparsity achieved when compared with the published result for SMLR [36]. For every dataset where the comparison could be made (five), sMKDA achieves substantially better sparsity than competing methods. These summaries are reflected in the grand means at the foot of Table A.1 indicating that the overall accuracy of sMKDA is not significantly different from the best published techniques but that, on average, it delivers more than three times the sparsity of the competing techniques.

It is worth noting that, while table A.1 gives a comparison of sMKDA against the best performing classifiers in terms of accuracy, some of these are not specifically designed to be sparse. Even the basic SVM can often be improved in this respect. Of the entries in table A.1 only Glass and Iris are from a system (SMLR) specifically designed this way. In both cases sMKDA proves to be the sparser. We have also examined results for these datasets and an additional one – Yeast – [36] processed by the Relevance Vector Machine [41] in a "one-versus-all mode". The RVM is also specifically designed to induce sparsity and table A.2 gives the results. In summary, for this small comparison, it can be said that while the RVM is substantially more parsimonious than SMLR at the cost of some accuracy, sMKDA has comparable sparsity but better accuracy than the RVM and is substantially sparser than SMLR but with slightly poorer accuracy. Again, we do not wish to draw too firm a conclusion from these comparisons because there are variations in experimental method, but we believe they further support the competitiveness of our approach.

3.2 Thyroid disease database

An additional, moderately large dataset, also taken from the UCI repository of machine learning databases [29] comprising 7200 records describing patients grouped into three classes, hypo (C_1), hyper (C_2) and normal (C_3) thyroid function through 21 real and categorical attributes is used to evaluate per-

formance. We denote this dataset "Thyroid 2" to avoid confusion with the smaller thyroid-related sample of the previous subsection. As well as its size, Thyroid 2 presents a challenge because the classes are grossly imbalanced with over 92% belonging to the normal group. The dataset is pre-partitioned into a training and a testing sample containing 3772 and 3428 samples, respectively. Again, the class priors are approximately retained in the training and testing samples and classes are assigned according to 12. A baseline, conventional multinomial LDA is first computed, giving a test sample accuracy of 93.84%. The contingency table for this is presented in Table A.3.

Then, applying sMKDA using five-fold cross-validation on the training sample to estimate the optimal kernel parameter and level of sparsity gives an accuracy of 98.67% with p=5 and 111 (2.94%) retained samples. In operation the classifier gives an accuracy of 97.72% on the testing sample. This shows an increase in accuracy over the linear result of 3.88%. The contingency table for this is presented in Table A.4.

What can be seen from Tables A.3 and A.4 is largely obscured by simply reporting the raw accuracies owing to the strong preponderance of normals. It now becomes clear that the introduction of the more flexible sMKDA substantially improves the class conditional accuracies for the minority groups. The hit rate for hypo-thyroid function rises from 58.90% to 76.71%, almost 18%, while for hyper-thyroid function, there is an almost 77% increase, from 1.69%. These improvements are obtained at the cost of a 0.5% drop in the accuracy on the normal population.

An important benefit of projecting data onto the most discriminatory canonical variates is the ability to visualize it. In particular, since here $c = 3$ a two-dimensional plot contains all the useful information about the sample. Figures A.2 and A.3 emphasize the improved separation between classes, in particular between C_2 and C_3. Figure A.2 clearly demonstrates how C_2 is almost entirely embedded in C_3 and, by examining the centroid positions, how its members can easily be mistaken for those of C_3. Figure A.3 then illustrates how sMKDA separates them and the class centroids leading to a substantially less ambiguous situation. The case is less striking for C_1 as is reflected in the statistics quoted above. It is also interesting to note that these figures support the suggestion of Hastie et al. [21] that the effect of taking linear combinations of the transformed data makes this sample at least appear to be more normally distributed.

4 Conclusion

A method for performing sparse, multinomial kernel discriminant analysis has been proposed and shown to achieve state-of-the-art performance in terms of both classification performance and degree of sparsity. The method is based on the connection between multinomial linear discrimination and least-squares and makes use of a widely used and highly successful method of achieving sparsity in least-squares problem, forward selection via orthogonal least-squares. This generalizes the method for the binomial case put forward in Billings and Lee [12]. The approach requires a least-squares solution which is computed in order $\mathcal{O}\left(n_{\mathrm{s}}n\right)^2$ followed by an eigen-decomposition of order $\mathcal{O}(c)$. The choice of a sparse basis is the sole means of complexity control considered here although the formulation admits regularization in the form of a quadratic penalty on the coefficients. The solution is genuinely sparse since it avoids the need to retain all training samples for operational use and can perform classification in a simple and direct way via minimum Mahalanobis distance in the canonical variates. The method has been evaluated on nine benchmark multinomial datasets from the UCI repository of machine learning databases and has been found to deliver comparable performance to the best published results across 11 recent machine learning articles. It has further been compared with methods designed to be sparse and again found to be competitive. An additional, larger and highly imbalanced dataset has also been examined in more detail and, again, performance there has been shown to be comparable with the current state-of-the-art.

Issues for future consideration concern the use of regularization to control, directly, the smoothness of the mapping. This might be approached by extending the current framework and using the Bayesian evidence framework to select the regularizer automatically as suggested in [27, 28] or to attempt to avoid explicit cross-validation by direct optimization of the ordinary cross-validation statistic [42]. Such methods might profitably be extended to a recursive formulation of the orthogonal least-squares algorithm [e.g. 43] to address the computational requirements when n is very large. One disadvantage of the forward selection approach is that it may prove sub-optimal in any given situation. Inducing sparsity as a consequence of optimization might, therefore, prove beneficial and the adoption of a multivariable generalization of, say, the LASSO algorithm [e.g. 44] or the method based on surrogate optimization described in [18]. This would have the advantage of performing regularization directly but would add to the cross-validation burden to select the degree of regularization required. This latter problem could be ameliorated by using a method based on "leave-one-out" cross validation – the PRESS statistic – such as the one suggested in [45].

References

[1] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis.* John Wiley & Sons, New York, NY, 1973.

[2] B. Schölkopf and A. Smola. *Learning with Kernels.* MIT Press, Cambridge, MA, 2002.

[3] J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis.* Cambridge University Press, Cambridge, 2004.

[4] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neur. Comput.*, 12:2385–2404, 2000.

[5] Z. Liang and P. Shi. An effiecient and effective method to solve kernel Fisher discriminant analysis. *Neurocomp.*, 61:485–493, 2004.

[6] Z. Liang and P. Shi. Kernel direct discriminant analysis and its theoretical foundation. *Patt. Recogn.*, 38:445–447, 2005.

[7] J. Lu, K.N. Plataniotis, and A.N. Venetsanopoulos. Face recognition using kernel direct discriminant analysis algorithms. *IEEE T. Neur. Netw.*, 14:117–126, 2003.

[8] V. Roth and V. Steinhage. Nonlinear discriminant analysis using kernel functions. In S. Solla, T.K. Leen, and K-R. Müller, editors, *Advances in neural information processing systems*, volume 12, pages 568–574, 1999.

[9] Y. Xu, J-Y. Yang, J. Lu, and D-J. Yu. An efficient renovation on kernel Fisher discriminant analysis and face recognition experiments. *Patt. Recogn.*, 37:2091–2094, 2004.

[10] Y. Xu, D. Zhang, Z. Jin, M. Li, and J-Y. Yang. A fast kernel-based nonlinear discriminant analysis for multi-class problems. *Patt. Recogn.*, 39:1026–1033, 2006.

[11] W. Zheng, L. Zhao, and C. Zou. A modified algorithm for generalized discriminant analysis. *Neur. Comput.*, 16:1283–1297, 2004.

[12] S.A. Billings and K.L. Lee. Nonlinear Fisher discriminant analysis using a minimum squared error cost function and the orthogonal least squares algorithm. *Neur. Netw.*, 15:263–270, 2002.

[13] G. Cawley and N. Talbot. Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. *Patt. Recogn.*, 36:2585–2592, 2003.

[14] T.P. Centeno and N. Lawrence. Optimising kernel parameters and regularisation coefficients for non-linear discriminant analysis. *J. Mach. Learn. Res.*, 7:455–491, 2006.

[15] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, A. Smola, and K-R. Müller. Constructing descriptive and discriminative nonlinear features: Rayleigh Coefficients in kernel feature spaces. *IEEE T. Patt. Anal.*, 25:623–628, 2003.

[16] T. Van Gestel, J.A. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor, and J. Vandewalle. Bayesian framework for least-squares support vector machine classifiers, Gaussian processes, and kernel Fisher discriminant analysis. *Neur. Comput.*, 14:1115–1147, 2002.

[17] Y. Xu, J-Y. Yang, and J. Yang. A reformative kernel Fisher discriminant

analysis. *Patt. Recogn.*, 37:1299–1302, 2004.

[18] R.F. Harrison and K. Pasupa. A simple iterative algorithm for parsimonious binary kernel Fisher discrimination. *Patt. Anal. and Appl.*, (In Press), 2008.

[19] J.A. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific Publishing Co. Ltd. Pte., Singapore, 2002.

[20] R. Rifkin and A. Klautau. In denfense of one-vs-all classification. *J. Mach. Learn. Res.*, 5:101–141, 2004.

[21] T. Hastie, R. Tibshirani, and A. Buja. Flexible discriminant analysis by optimal scoring. *J. Amer. Stat. Assoc.*, 89:1255–1270, 1994.

[22] S. Chen, S.A. Billings, and W. Luo. Orthogonal least squares methods and their application to non-linear system identification. *Int. J. Cont.*, 50:1873–1896, 1989.

[23] R.M. Crownover. A least squares approach to linear discriminant analysis. *SIAM J. Sci. Stat. Comp.*, 12:595–606, 1991.

[24] T. Hastie, A. Buja, and R. Tibshirani. Penalized discriminant analysis. *Ann. Stat.*, 23:73–102, 1995.

[25] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer-Verlag, New York, NY, 2001.

[26] S. Chen, P.M. Grant, and C.F. Cowan. Orthogonal least-squares algorithm for training multioutput radial basis function networks. *IEE Proc. F*, 139:378–384, 1992.

[27] S. Chen. Local regularization assisted orthogonal least-squares regression. *Neurocomp.*, 69:559–585, 2006.

[28] S. Chen, E.S. Chng, and K. Alkadhimi. Regularized orthogonal least squares algorithm for constructing radial basis function networks. *Int. J. Cont.*, 64:829–837, 1996.

[29] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz. UCI Repository of machine learning databases. World Wide Web, `http://www.ics.uci.edu/~mlearn/MLRepository.html`, 1998.

[30] S. Abe. Sparse least squares support vector training in the reduced empirical feature space. *Patt. Anal. and Appl.*, 10(3):203–214, 2007.

[31] L. Bo, L. Wang, and L. Jiao. Feature scaling for kernel Fisher discriminant analysis using leave-one-out cross validation. *Neur. Comput.*, 18:961–978, 2006.

[32] G.C. Cawley, N.L. Talbot, and M. Girolami. Sparse multinomial logistic regression via bayesian L1 regularisation. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19, pages 209–216. MIT Press, Cambridge, MA, 2007.

[33] R. El-Yaniv, D. Pechyony, and E. Yom-Tov. Superior multi-class classification through a margin-optimized single binary problem. Technical Report H-0243, IBM, 2006.

[34] L. Gonzalez-Abril, C. Angulo, F. Velasco, and J. A. Ortega. A note on

the bias in SVMs for multiclassification. *IEEE T. Neur. Netw.*, 19(4): 723–725, 2008.

[35] C.-W. Hsu and C.-J. Lin. A comparison of methods for multi-class support vector machines. *IEEE T. Neur. Netw.*, 13(2):415–425, 2002.

[36] B. Krishnapuram, L. Carin, M.A. Figueiredo, and A.J. Hartemeink. Sparse multinomial logistic regression: fast algorithms and generalization bounds. *IEEE T. Patt. Anal.*, 27:957–968, 2005.

[37] T. Li, S. Zhu, and M. Ogihara. Using discriminant analysis for multi-class classification: an experimental investigation. *Knowl. Inf. Syst.*, 10 (4):453–472, 2006.

[38] E. Mayoraz and E. Alpaydin. Support vector machines for multi-class classification. In José Mira and Juan Vincente Sánchez-Andrés, editors, *Engineering Applications of Bio-Inspired Artificial Neural Networks, International Work-Conference on Artificial and Natural Neural Networks, IWANN '99*, volume 1607 of *Lecture Notes in Computer Science*, pages 833–842. Springer, 1999.

[39] J. Weston and C. Watkins. Support vector machines for multi-class pattern recognition. In *Proceedings of 7th European Symposium on Artificial Neural Networks*, pages 219–224, 1999.

[40] W. Zuo, D. Zhang, and K. Wang. On kernel difference-weighted k-nearest neighbor classification. *Patt. Anal. and Appl.*, Online, 2008.

[41] M. Tipping. Sparse Bayesian learning and the Relevance Vector Machine. *J. Mach. Learn. Res.*, 1:211–244, 2001.

[42] X. Hong, S. Chen, and P.M. Sharkey. Automatic kernel regression modelling using combined leave-one-out test score and regularised orthogonal least squares. *ijns*, 14:27–37, 2004.

[43] D.L. Yu, J.B. Gomm, and D. Williams. A recursive orthogonal least squares algorithm for training RBF networks. *Neur. Proc. Lett.*, 5:167–176, 1997.

[44] T. Similä and J. Tikka. Input selection and shrinkage in multiresponse linear regression. *Comp. Stat. Data Anal.*, 52:406–422, 2007.

[45] S. Chen, X. Hong, C.J. Harris, and P.M. Sharkey. Sparse modelling using orthogonal forward regression with PRESS statistic and regularization. 34:898–911, 2006.

A Appendix

The numbers of classes for the UCI datasets along with their prior probabilities are shown in Table A.5

List of Figures

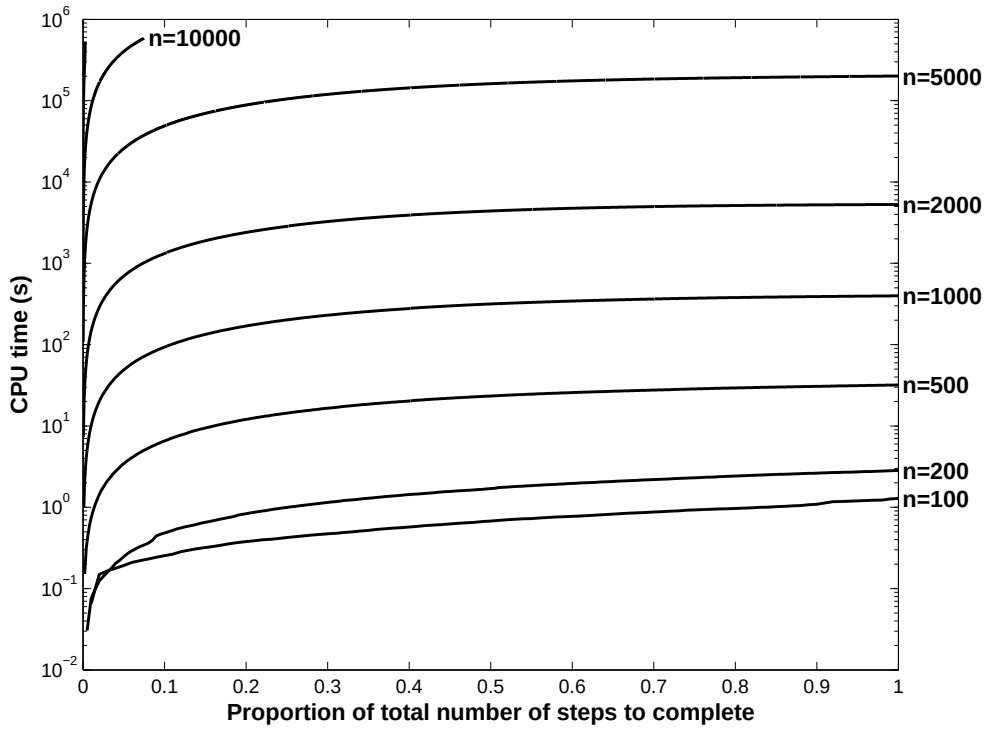

Fig. A.1. CPU times for OLS computations for a range of sample sizes.

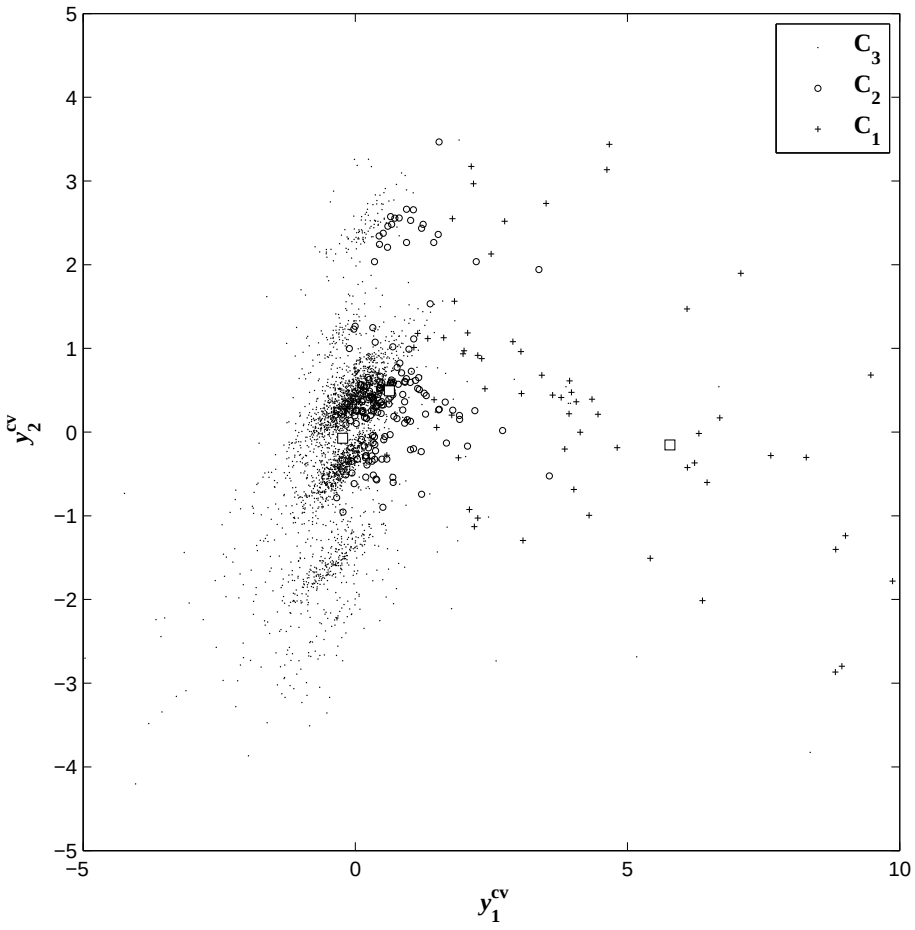

Fig. A.2. Conventional (linear) LDA projection of Thyroid 2 testing sample onto its complete set of $c-1$ canonical variates. The class centroids are denoted by squares.

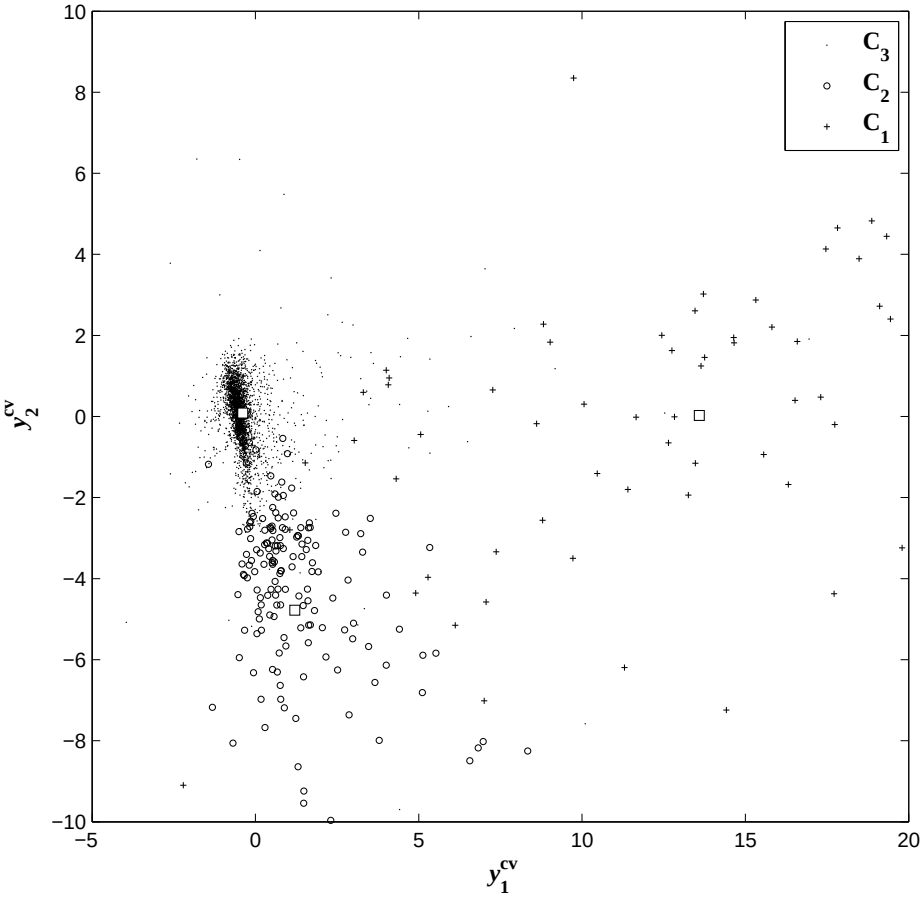

Fig. A.3. sMKDA projection of Thyroid 2 testing sample onto its complete set of $c-1$ canonical variates. The class centroids are denoted by squares.

List of Tables

Table A.1
Results of applying sMKDA to nine datasets from the UCI repository of machine learning databases. Accuracy (%), percentage retained samples (%), and kernel parameter, p, are averaged across the five data partitions and, where available, the standard error is shown. For each dataset, accuracy appears on the first row and percentage retained samples (complement of sparsity) on the second. Best performers are set in bold face type.

Dataset	c	n	d	Published Best		sMKDA	
				Method	Results	Results	p
Dermatology	6	358	34	SVM [40]	97.60 ± 0.42	**98.49 ± 0.15**	14.75±3.5
					Not Reported	**18.30 ± 4.52**	
Glass	6	214	9	SMLR [36]	**76.64**	73.08 ± 1.13	3.85±1.21
					93.37	**9.11 ± 1.68**	
Iris	3	150	4	SMLR [36]	**99.33**	99.20 ± 0.30	5.75±2.70
					50.37	**6.33 ± 2.54**	
Tae	3	151	5	KNN [40]	**64.83 ± 2.72**	63.84 ± 1.52	1.00±0.35
					100.00	**65.07 ± 4.92**	
Thyroid	3	215	5	SVM [34]	96.60 ± 0.40	**97.21 ± 0.33**	2.20±0.62
					Not Reported	**8.95 ± 1.87**	
Vehicle	4	846	18	SVM [35]	**87.47**	86.67 ± 0.56	11.05±1.07
					45.00	**15.16 ± 2.21**	
Vowel	11	990	11	SVM [34]	95.20 ± 0.13	**97.41 ± 0.43**	10.45±0.27
					Not Reported	**23.91 ± 0.99**	
Wine	3	178	13	SVM [35]	99.44	**99.78 ± 0.31**	7.10±1.17
					35.10	**9.55 ± 3.05**	
Zoo	7	101	16	LDA [37]	97.00	**99.41 ± 0.54**	6.40±1.58
					Not Applicable	**14.11 ± 2.85**	
Mean Accuracy (%)					90.46	**90.57 ± 0.59**	
Mean Retained Samples (%)					64.75	**18.94 ± 2.74**	

Table A.2
Comparison of sMKDA with RVM on three datasets: Glass, Iris and Yeast. For each dataset, percentage accuracy appears on the first row and percentage retained samples (complement of sparsity) on the second.

Dataset	c	n	d	RVM	sMKDA
Glass	6	214	9	71.50	73.08 ± 1.13
				7.67	9.11 ± 1.68
Iris	3	150	4	93.33	99.20 ± 0.30
				13.70	6.33 ± 2.54
Yeast	5	208	79	94.23	96.44 ± 0.43
				6.73	4.62 ± 1.54

Table A.3
Contingency Table for Thyroid 2 testing data – LDA

		Estimated			
		C_1	C_2	C_3	Total
Observed	C_1	43	7	23	73
	C_2	1	3	173	177
	C_3	6	1	3171	3178
Total		50	11	3367	3428

Table A.4
Contingency Table for Thyroid 2 testing data – sMKDA, $p = 5$, number of retained samples = 111.

		Estimated			
		C_1	C_2	C_3	Total
Observed	C_1	56	10	7	73
	C_2	0	139	38	177
	C_3	7	16	3155	3178
Total		63	165	3200	3428

Table A.5
Class Prior Probability

Dataset	c	Class Prior Probability
Dermatology	6	{0.310, 0.168, 0.198, 0.134, 0.134, 0.056}
Glass	6	{0.327, 0.355, 0.079, 0.061, 0.042, 0.136}
Iris	3	{0.333, 0.333, 0.333}
Tae	3	{0.325, 0.331, 0.344}
Thyroid	3	{0.698, 0.163, 0.140}
Vehicle	4	{0.235, 0.257, 0.258, 0.251}
Vowel	11	{0.091, 0.091, 0.091, 0.091, 0.091, 0.091, 0.091, 0.091, 0.091, 0.091, 0.091}
Wine	3	{0.331, 0.399, 0.270}
Zoo	7	{0.406, 0.198, 0.050, 0.129, 0.040, 0.079, 0.099}