WCCI 2010 IEEE World Congress on Computational Intelligence
July, 18-23, 2010 - CCIB, Barcelona, Spain

Cohort-based Kernel Visualisation with Scatter Matrices

Enrique Romero, Ana Sofia Fernandes, Tingting Mu and Paulo J.G. Lisboa

***Abstract*— A key question in medical decision support is how best to visualise a patient database, with especial reference to cohort labelling, whether this is an indicator function for classification or a cluster index. We propose the use of the kernel trick to visualise complete patient databases, in low-dimensional projections, with class labelling, given a non-linear classifier of choice. The results show that this method is useful both to see how individual patient cases relate to each other with reference to the classification boundary, and also to obtain a visual indication of the separation that can be obtained with difference choices of kernel functions.**

I. Introduction

A central task in medical decision support is to provide statistical inferences of class membership, if at all possible supplemented by a direct visualisation of the patient data base with specific reference to the chosen classifier. For linear classifiers, this can be achieved with linear projective methods [16]. However, with non-linear classifiers e.g. kernel-based Support Vector Machines (SVM), it is non-trivial to obtain a linearly-separable visualisation of the data. Ideally, this will use the kernel trick to effect the linear separation. This has the advantage, first, of presenting the totality of the individual cases recorded in the data base in a single low-dimensional projection that relates directly to a classification map. Secondly, the degree of separation that is apparent in the map will provide useful guidance in the choice of kernels.

Low-dimensional visualisation methods generally fall into three categories. Purely linear methods frequently utilise singular values spanning the largest variance in the data or preserving pairwise inner products, such as the widely used Principal Component Analysis-based bi-plots [11] or classical Multi-Dimensional Scaling [22]. A second approach is to relax the linearity restriction and to define a non-linear projection to optimise the correspondence between distances in the original input space and distances in the projected space, such as in generalised (metric and non-metric) Multi-Dimensional Scaling [6] (including for example Sammon's mapping [18] and Kruskal's approach [14]) or Isomap [21]. A third approach generates topographic maps by projecting data onto a curved surface weaving through the data and cutting through noise, such as in Self-Organizing Maps [13] or in Generative Topographic Mapping [4].

Most visualisation methods construct the mapping to low-dimensional spaces in an unsupervised manner (i.e., without using the class labels) and then represent differently objects from different classes for comparison purposes [15]. Although this property makes the method generically applicable, it does not always use all of the information that is available, for instance clustering labels or prior knowledge about class membership. This is particularly important if the objective is to visualise the separation between cohorts, be they different partition clusters of data from multiple classes. The most common supervised visualisation methods are based on Linear Discriminant Analysis (LDA) (see, for example, [7]) and its non-linear (kernel) version, Kernel Linear Discriminant Analysis (KLDA) [19]. While KLDA searches for optimal directions in feature space for which separation between classes is maximal, this method does not directly attempt to perform dimensionality reduction, which is important in data visualisation. It is possible to combine KLDA with feature space selection by optimising the kernel parameters directly [24] or, alternatively, to map the data onto a high-dimensional kernel space and then optimise a scatter-matrix separation index similar to that used in this paper, based on the Kernel PCA transformation [23]. Some relationships between KLDA, Kernel PCA (KPCA) [20] and LDA can be found in [25]. These methods have been further extended to kernel quadratic discriminant analysis [17].

This paper takes advantage of a recent result derived for linear visualisation of labelled data cohorts, typically defined by cluster membership or class labels, in the context of scatter-matrix separation measures. In the Cluster-based Visualisation with Scatter Matrices (CVSM) method, described in [16], it was shown in that the space spanned by the cohort means forms a useful basis for dimensionality reduction while retaining much of the cohort separation measured by a quadratic index. In particular, when the covariance matrix of the original data is non-singular, it was proven that this data compression exactly preserves the value of the separation index, thus preserving the cohort separation at the level of second-order statistics. This is a surprising amount of separation for what can be a drastic reduction in dimensionality, raising the possibility that a similarly efficient compression may be possible for non-linearly separable data cohorts, through a suitable application of kernel methods. The application of kernels to CVSM may lead to insights about the data, by generating linearly separable views of non-linearly separable data.

In this paper, a kernel extension of CVSM [16] is presented. Although the kernel trick cannot be directly applied, this drawback can be avoided by representing the data in dual form. With this representation, the projections of the data onto the sub-space spanned by the (orthonormalised) class means in the feature space can be easily computed. The

Enrique Romero is with the Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, Spain.

Ana Sofia Fernandes is with the Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal

Tingting Mu is with the National Centre for Text Mining, School of Computer Science, University of Manchester, UK

Paulo J.G. Lisboa is with the School of Computing and Mathematical Sciences, Liverpool John Moores University, UK

978-1-4244-8126-2/10/$26.00 ©2010 IEEE

only parameters of the whole method are the parameters of the kernel function, and usually requires only the inversion of an $N_c \times N_c$ matrix, where N_c is the number of classes, rather than the solution of a convex quadratic optimisation problem [19] or an eigenvalue problem of size N (N is the number of examples) [3] in KLDA.

The result is a visualisation of the kernel-defined feature space, which can facilitate to find the most appropriate kernels for visualisation and, potentially, for kernel-based classification of a given data set. If we accept that the role of the kernel is to project the data onto a space where the projections are linearly separable, then it follows that measuring the extent of linear separation with the proposed method helps to short-list the most useful kernels for a given classification task. Since the proposed method can induce a linearly separable visualisation for data with non-linear decision boundaries between population cohorts, it provides a direct visualisation of complex data sets in a feature space that is relevant to their categorisation into labelled groups, be they clusters or classes.

The proposed method is illustrated with three cancer data sets. The visualisation plots give a good indication of the effect of the kernels on the data distribution vs. the classification label.

II. Cluster-based Linear Visualisation with Scatter Matrices

The method proposed in [16] is linear in nature and it is based on the decomposition of the invariant scatter matrix after projecting the data onto the subspace spanned by the class means. It is well-known that the overall variance of the data, $\mathbf{S}_T$, can be decomposed into the sum of scatter matrices calculated within and between labelled cohorts [7], [10] thus generating a within-cluster matrix, $\mathbf{S}_W$, and a between-cluster matrix, $\mathbf{S}_B$, such that $\mathbf{S}_T = \mathbf{S}_W + \mathbf{S}_B$. For a data matrix $\mathbf{X} = \{x_i\}_{i=1}^N$ comprising N rows with d-dimensional data points of overall mean m,

$$\mathbf{S}_T = \sum_{i=1}^{N} \{(x_i - m)^T (x_i - m)\}$$

$$\mathbf{S}_W = \sum_{j=1}^{N_c} \sum_{i=1}^{N_j} \{(x_i^j - m_j)^T (x_i^j - m_j)\}$$

$$\mathbf{S}_B = \sum_{j=1}^{N_c} N_j \{(m_j - m)^T (m_j - m)\},$$

where the data are partitioned into N_c groups, each with N_j points and mean m_j. Note that $\mathbf{S}_T$, $\mathbf{S}_W$, and $\mathbf{S}_B$ are $d \times d$ matrices. This decomposition generates a natural scalar index for the separation between the data cohorts by taking the trace of the scatter matrix $\mathbf{M} = \mathbf{S}_W^{-1}\mathbf{S}_B$ leading to the class separation index $J = tr(\mathbf{M})$.

A strength of the index J is its invariance to affine transformations of the data matrix, which makes it insensitive to co-linearities in the data and to changes in relative scaling of the covariates, both of which are useful properties for exploratory analysis of high dimensional data as in bioinformatics. Furthermore, if the covariance matrix is non-singular, then this invariance can be exploited by applying a Mahalanobis rotation to de-correlate the covariates, or sphering the data, thus rendering $\mathbf{S}_T$ is diagonal. The scatter matrix decomposition now indicates that the information contained in the within-cluster scatter matrix is implicit in the between-cluster matrix, which uses only the values of the class means $\{m_j\}_{j=1}^{N_c}$ as representatives for the classes.

This suggests that the class means form a natural basis to project the data with minimal loss in class separation as measured by second-order statistics. It is shown in [16] that the value of the separation index J is strictly invariant to a Mahalanobis transformation followed by a linear projection onto the space of class means. The paper also shows that when the covariance matrix is singular, then some loss is induced by this dimensionality reduction, but most of the class separation is maintained.

However, in cases where the classes are non-linearly separable, then the scatter-matrix based separation index is not a reliable measure of class separation since this is no longer well represented by the second-order statistics of the data. A better low-dimensional projection may then be obtained by resorting to non-linear features, for instance through the use of kernels. To see how this can be done we briefly review the linear projective method described in [16].

The compression onto the subspace of the class means is readily achieved by defining an orthonormal set of basis vectors $\mathbf{B}^T = \{b_j\}_{j=1}^{N_c}$, for instance by Gram-Schmidt orthogonalisation, generating the projection of $\mathbf{X}$ onto the space spanned by the set of orthonormalised cluster mean vectors

$$\mathbf{X}^c = \mathbf{X} \cdot \mathbf{B}.$$

Note that $\mathbf{X}^c$ and $\mathbf{B}$ are $N \times N_c$ and $d \times N_c$ matrices, respectively. Scatter matrices for $\mathbf{X}^c = \{x_i^c\}_{i=1}^N$ can be calculated in the space of class means, namely:

$$\mathbf{S}_W^c = \sum_{j=1}^{N_c} \sum_{i=1}^{N_j} \{(x_i^c - m_j^c)^T (x_i^c - m_j^c)\}$$

$$\mathbf{S}_B^c = \sum_{j=1}^{N_c} N_j \{(m_j^c - m^c)^T (m_j^c - m^c)\}$$

and, similarly, an invariant scatter matrix $\mathbf{M}^c = (\mathbf{S}_W^c)^{-1}\mathbf{S}_B^c$ and an invariant class separation index $J^c = tr(\mathbf{M}^c)$ can be defined. Note that $\mathbf{S}_W^c$ and $\mathbf{S}_B^c$ are $N_c \times N_c$ matrices, so that the computation of $(\mathbf{S}_W^c)^{-1}$ is computationally fast.

In [16] it is shown that the invariant separation measure J is exactly preserved (i.e., $J = J^c$) when the projection is preceded by a sphering of the data. A diagonalisation of the new scatter matrix $\mathbf{M}^c$ shows, typically, that the trace of the matrix is contained in the largest few eigenvalues. Their correspondent eigenvectors form the basis for a two- or three-dimensional visualisation of the data. The whole visualisation procedure can be summarised in figure 1. Note that the method is parameter-free.

```
Given $\mathbf{X} = \{x_i\}_{i=1}^{N}$,
    1. Optionally, sphere the data: $\mathbf{X} = \mathbf{X} \cdot \mathbf{\Sigma}^{-1/2}$
    2. Compute and orthonormalise the class means: $\mathbf{B}^T$
    3. Project data onto the orthonormalised class means:
           $\mathbf{X}^c = \mathbf{X} \cdot \mathbf{B}$
    4. Compute scatter matrices for $\mathbf{X}^c$: $\mathbf{S}_W^c$, $\mathbf{S}_B^c$ and $\mathbf{M}^c$
    5. Project $\mathbf{X}^c$ onto the eigenvectors of the largest
           eigenvalues of $\mathbf{M}^c$
```

Fig. 1. Linear visualisation algorithm proposed in [16].

As previously said, a natural extension of this approach is to investigate the use of kernel transformations to further separate the clusters, or class-labelled cohorts, in the low-dimensional projective space. The next section describes how this can be done.

III. Cohort-based Kernel Visualisation with Scatter Matrices

A deeper analysis of the method described in figure 1 reveals that, in order to construct $\mathbf{X}^c$, only inner products are needed. This allows us to develop a non-linear extension of the visualisation method in section II by employing the kernel trick, leading to what we term Cohort-based Kernel Visualisation with Scatter Matrices (CKVSM). The core idea is to map the data into an inner product space $\mathcal{F}$ corresponding to a high-dimensional (maybe infinite-dimensional) non-linear mapping $\boldsymbol{\phi}$, chosen *a priori*, where the method could be applied. A (positive definite) kernel function $K(\boldsymbol{u}, \boldsymbol{v})$ is used to evaluate the inner product between the mapped feature vectors $\langle \boldsymbol{\phi}(\boldsymbol{u}), \boldsymbol{\phi}(\boldsymbol{v}) \rangle = K(\boldsymbol{u}, \boldsymbol{v})$, so that the mapping function becomes implicit. This kernel-based procedure has been widely used to define non-linear versions of classical linear procedures, such as KPCA (for PCA) or KLDA (for LDA) [19].

As explained in section II, two levels of projection are pursued in the cluster-based visualisation: (1) projecting the sphered data points into the N_c-dimensional space spanned by the orthonormalised class means; (2) further projecting the obtained N_c-dimensional data points into an n-dimensional space spanned by the eigenvectors of the scatter matrix $\mathbf{M}_c$. In the kernelised version of the above approach, both the sphering procedure and the construction of the projection onto the space spanned by the orthonormalised cohort means in (1) are conducted in the kernel-based feature space, as explained in the next sections.

A. Sphering in the Feature Space

Given an $N \times d$ matrix of centered data points $\mathbf{X} = \{\boldsymbol{x}_i\}_{i=1}^{N}$, the covariance matrix can be defined as $\mathbf{\Sigma} = \frac{1}{N}\mathbf{X}^T\mathbf{X}$. Since $\mathbf{\Sigma}$ is symmetric, it can be decomposed as $\mathbf{\Sigma} = \mathbf{V}_\Sigma \mathbf{D}_\Sigma \mathbf{V}_\Sigma^T$, where $\mathbf{D}_\Sigma$ is a diagonal matrix with the non-zero eigenvalues of $\mathbf{\Sigma}$, and $\mathbf{V}_\Sigma$ is an orthonormal matrix whose columns are the corresponding eigenvectors of $\mathbf{\Sigma}$. The sphering of the data consists of a rotation of the data by applying a linear transformation $\mathbf{Y} = \mathbf{X}\mathbf{R}_\Sigma$, where $\mathbf{R}_\Sigma = \mathbf{\Sigma}^{-1/2} = \mathbf{V}_\Sigma \mathbf{D}_\Sigma^{-1/2} \mathbf{V}_\Sigma^T$, so that the covariance matrix of $\mathbf{Y}$ is the identity matrix. The inner product matrix between the transformed features is $\mathbf{Y}\mathbf{Y}^T = \mathbf{X}\mathbf{V}_\Sigma \mathbf{D}_\Sigma^{-1} \mathbf{V}_\Sigma^T \mathbf{X}^T = \mathbf{X}\mathbf{\Sigma}^{-1}\mathbf{X}^T$.

Suppose now that we have a set of centered data points $\mathbf{\Phi} = \{\boldsymbol{\phi}_i\}_{i=1}^{N}$ in the feature space induced by the kernel function $K(\boldsymbol{u}, \boldsymbol{v})$. The kernel trick cannot be directly applied for sphering the data with $\mathbf{R}_\Sigma$, because $\mathbf{D}_\Sigma$ and $\mathbf{V}_\Sigma$ represent matrices in the feature space, that may be unknown. Even if they were known, they could have infinite dimension (recall that $\mathbf{D}_\Sigma$ and $\mathbf{V}_\Sigma$ have d columns, where d is the dimension of the space).

As we will see, to work with sphered data in the inner product (kernel-based feature space) we only need to know how to compute the inner product of any two sphered data points, instead of the representation of the sphered data itself. The rest of the section explains how to compute the inner product matrix of sphered data in the feature space, by studying the relationship between the eigenvectors of the covariance matrix in the feature space and the eigenvectors of the kernel matrix.

1) Eigendecomposition of the Covariance Matrix in the Feature Space : We represent $\mathbf{\Phi}$ as an $N \times \infty$ matrix. The inner product matrix of $\mathbf{\Phi}$ can be computed as $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^T$, where $\mathbf{K}$ is the $n \times n$ kernel matrix. The covariance matrix in the feature space can be defined as usual $\hat{\mathbf{\Sigma}} = \frac{1}{N}\mathbf{\Phi}^T\mathbf{\Phi}$. Since $\mathbf{K}$ is symmetric, it can be decomposed as $\mathbf{K} = \mathbf{V}_K\ \mathbf{D}_K \mathbf{V}_K^T$ or, equivalently

$$\mathbf{K}\mathbf{V}_K = \mathbf{V}_K\ \mathbf{D}_K, \tag{1}$$

where $\mathbf{D}_K$ is a diagonal matrix with the non-zero eigenvalues of $\mathbf{K}$, and $\mathbf{V}_K$ is an orthonormal matrix whose columns are the corresponding eigenvectors of $\mathbf{K}$. The covariance matrix in the feature space $\hat{\mathbf{\Sigma}}$ can also be decomposed as

$$\hat{\mathbf{\Sigma}}\mathbf{V}_{\hat{\Sigma}} = \mathbf{V}_{\hat{\Sigma}}\mathbf{D}_{\hat{\Sigma}}, \tag{2}$$

where $\mathbf{D}_{\hat{\Sigma}}$ is a diagonal matrix with the non-zero eigenvalues of $\hat{\mathbf{\Sigma}}$, and $\mathbf{V}_{\hat{\Sigma}}$ is an orthonormal matrix whose columns are the corresponding eigenvectors of $\hat{\mathbf{\Sigma}}$.

In [20] it is proved that the eigenvectors of $\hat{\mathbf{\Sigma}}$ can be expressed as a function of the eigenvectors of $\mathbf{K}$ and vice-versa:

1) The eigenvectors $\mathbf{V}_{\hat{\Sigma}}$ of $\hat{\mathbf{\Sigma}}$ satisfy that $\mathbf{\Phi}\mathbf{V}_{\hat{\Sigma}}$ are eigenvectors of $\mathbf{K}$ with corresponding eigenvalues $N\mathbf{D}_{\hat{\Sigma}}$
2) the eigenvectors $\mathbf{V}_{_K}$ of $\mathbf{K}$ satisfy that $\mathbf{\Phi}^T\mathbf{V}_{_K}$ are eigenvectors of $\hat{\mathbf{\Sigma}}$ with corresponding eigenvalues $\frac{1}{N}\mathbf{D}_{_K}$

As a consequence, $\hat{\mathbf{\Sigma}}$ and $\mathbf{K}$ have the same number of eigenvectors with non-zero eigenvalue, and, in addition, $\mathbf{D}_K = N\mathbf{D}_{\hat{\Sigma}}$. Let $n \leqslant N$ be the number of eigenvectors of $\mathbf{K}$ with non-zero eigenvalue. Note that $\mathbf{V}_{\hat{\Sigma}}$ has n columns and the dimension of $\mathbf{D}_{\hat{\Sigma}}$ is n. We will express the eigenvectors of $\hat{\mathbf{\Sigma}}$ as a function of $\mathbf{V}_{_K}$ to compute the inner product of two sphered points in the feature space.

2) Orthogonality Condition : The orthogonality condition of the eigenvectors of the kernel matrix, given as $\mathbf{V}_K^T\mathbf{V}_K = \mathbf{I}$, is automatically preserved during the computation of the eigendecomposition of $\mathbf{K}$. However, when the eigenvectors of the covariance matrix $\hat{\Sigma}$ are computed from $\mathbf{V}_K$ instead of the direct computation of its eigendecomposition, the orthogonality condition, given as $\mathbf{V}_{\hat{\Sigma}}^T\mathbf{V}_{\hat{\Sigma}} = \mathbf{I}$, requires to be imposed. Let $\mathbf{S}$ be a $n \times n$ scaling diagonal matrix of $\mathbf{V}_{\hat{\Sigma}}$, such that $\mathbf{V}_{\hat{\Sigma}} = \mathbf{\Phi}^T\mathbf{V}_{_K}\mathbf{S}$. When the orthogonality condition is imposed, we have

$$\mathbf{I} = \mathbf{V}_{\hat{\Sigma}}^T\mathbf{V}_{\hat{\Sigma}} = \mathbf{S}^T\mathbf{V}_{_K}^T\mathbf{\Phi}\mathbf{\Phi}^T\mathbf{V}_{_K}\mathbf{S} = \mathbf{S}^T\mathbf{V}_K^T\mathbf{K}\mathbf{V}_K\mathbf{S}. \tag{3}$$

By incorporating Eq. (1) and $\mathbf{V}_K^T\mathbf{V}_K = \mathbf{I}$ into Eq. (3), we simply have $\mathbf{S}^T\mathbf{D}_K\mathbf{S} = \mathbf{I}$. Therefore, $\mathbf{S} = \mathbf{D}_K^{-1/2}$, and equivalently,

$$\mathbf{V}_{\hat{\Sigma}} = \mathbf{\Phi}^T\mathbf{V}_K\mathbf{D}_K^{-1/2}. \tag{4}$$

3) Sphered Inner Product Matrix in the Feature Space : Let us return to the problem of computing the inner product of two sphered data points in the feature space. Similar to the input space, we can define the rotation matrix $\mathbf{R}_{\hat{\Sigma}} = \mathbf{V}_{\hat{\Sigma}}\mathbf{D}_{\hat{\Sigma}}^{-1/2}\mathbf{V}_{\hat{\Sigma}}^T$, and apply the linear transformation $\hat{\mathbf{\Phi}} = \mathbf{\Phi}\mathbf{R}_{\hat{\Sigma}}$. By incorporating Eq. (4), the new kernel matrix of inner products after sphering in the feature space can be computed as $\hat{\mathbf{K}} = \hat{\mathbf{\Phi}}\hat{\mathbf{\Phi}}^T = \mathbf{\Phi}\mathbf{\Phi}^T\mathbf{V}_K\mathbf{D}_K^{-1/2}\mathbf{D}_{\hat{\Sigma}}^{-1}\mathbf{D}_K^{-1/2}\mathbf{V}_K^T\mathbf{\Phi}\mathbf{\Phi}^T$. Since $\mathbf{D}_K^{-1/2}\mathbf{D}_{\hat{\Sigma}}^{-1}\mathbf{D}_K^{-1/2} = N\mathbf{D}_K^{-2}$ and using Eq. (1), this leads to a sphered inner product matrix, finally computed as

$$\hat{\mathbf{K}} = N\mathbf{V}_K\mathbf{V}_K^T, \tag{5}$$

in the kernel-based feature space. If $n = N$, then $\hat{\mathbf{K}}$ is N times the identity matrix.

B. Projection onto the Space Spanned by the Orthonormalised Cohort Means in the Feature Space

1) Dual-form Representation: The kernel trick cannot be directly applied in the method described in figure 1, because $\mathbf{B}^T$ represent points in the feature space, that may be unknown. This affects to steps 2 and 3 of the algorithm in figure 1. However, this drawback can be avoided by representing the data in dual form, which is one of the key points of the proposed method: let $\boldsymbol{a} = (a_1, a_2, \ldots, a_N) \in \mathbb{R}^N$ represent the point $\hat{\boldsymbol{a}} = \sum_{i=1}^N a_i\boldsymbol{\phi}(\boldsymbol{x}_i)$ in the feature space $\mathcal{F}$ (recall that $\{\boldsymbol{x}_i\}_{i=1}^N$ is the original data). With this representation:

1) Vector space operations in $\mathbb{R}^N$ have a direct correspondence in $\mathcal{F}$.
2) Inner products in the feature space between two vectors $\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}} \in \mathcal{F}$ in dual form (without sphering) can be computed as usual:

$$\langle\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}\rangle = \sum_{i,j=1}^N a_i b_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{a}^T\mathbf{K}\boldsymbol{b}, \tag{6}$$

where $\mathbf{K}$ is the kernel matrix. Inner products between two vectors $\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}} \in \mathcal{F}$ in dual form after sphering can be computed as

$$\langle\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}\rangle = \sum_{i,j=1}^N a_i b_j \hat{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{a}^T\hat{\mathbf{K}}\boldsymbol{b}, \tag{7}$$

where $\hat{\mathbf{K}}$ is defined in Eq. (5). Let

$$\mathbf{\Upsilon} = \begin{cases} \mathbf{K} & \text{if data is not sphered in the feature space,} \\ \hat{\mathbf{K}} & \text{otherwise.} \end{cases} \tag{8}$$

3) The mean of class C_j in the feature space is represented as $\boldsymbol{m}_j = (m_{j1}, m_{j2}, \ldots, m_{jN})$, where

$$m_{ji} = \begin{cases} 1/N_j & \text{if } \boldsymbol{x}_i \text{ belongs to } C_j, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

4) The Gram-Schmidt orthonormalisation procedure can be applied as usual, since only inner products and vector spaces operations are needed. In this case, however, the orthonormal set of basis vectors $\hat{\mathbf{B}}$ is a $N \times N_c$ matrix (it is also represented in dual form).
5) Since $\hat{\mathbf{B}}$ represents the orthonormal set of basis vectors in dual form, and using Eq. (8), the projection of the data onto the orthonormalised cohort means in the feature space can be obtained as: $\mathbf{X}^c = \mathbf{\Upsilon}\hat{\mathbf{B}}$ (recall that the projection of $\boldsymbol{\phi}(\boldsymbol{x}_j)$ can be computed as $\boldsymbol{u}_j^T\mathbf{\Upsilon}\hat{\mathbf{B}}$, where $\boldsymbol{u}_j^T$ is the N-dimensional vector with a 1 in position j and 0 elsewhere).
6) Once $\mathbf{X}^c$ has been obtained, steps 4 and 5 of the algorithm in figure 1 can be performed.

2) Kernelised Gram-Schmidt Orthonormalisation: Here, we describe the extension of the Gram-Schmidt orthonormalisation in the feature space of vectors represented in dual form. Recall that inner products $\langle\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}\rangle$ in the feature space are computed with equations (6) or (7). Given a set $\{\boldsymbol{m}_j\}_{j=1}^M$ of vectors in $\mathbb{R}^N$ representing vectors in dual form $\{\hat{\boldsymbol{m}}_j\}_{j=1}^M$ in the feature space, the set $\{\boldsymbol{b}_j\}_{j=1}^M$ of vectors in $\mathbb{R}^N$ defined in the algorithm described in figure 2 represent a set of orthonormal vectors in dual form $\{\hat{\boldsymbol{b}}_j\}_{j=1}^M$ in the feature space that span the same subspace than $\{\hat{\boldsymbol{m}}_j\}_{j=1}^M$.

```
$\boldsymbol{b}_1 = \boldsymbol{m}_1/\sqrt{\langle\hat{\boldsymbol{m}}_1, \hat{\boldsymbol{m}}_1\rangle}$
for $j = 2 \ldots M$
    $\boldsymbol{b}_j = \boldsymbol{m}_j - \sum_{i=1}^{j-1}\langle\hat{\boldsymbol{m}}_j, \hat{\boldsymbol{b}}_i\rangle\boldsymbol{b}_i$
    $\boldsymbol{b}_j = \boldsymbol{b}_j/\sqrt{\langle\hat{\boldsymbol{b}}_j, \hat{\boldsymbol{b}}_j\rangle}$
end for
```

Fig. 2. Gram-Schmidt orthonormalisation algorithm in the feature space.

3) Non-centered Data: For the sake of simplicity, we have made the assumption that the data are centered. If this was not the case, the previously showed results are still valid changing $\mathbf{K}$ by $\overline{\mathbf{K}} = \mathbf{K} - \mathbf{K}\mathbf{1}_N - \mathbf{1}_N\ \mathbf{K} + \mathbf{1}_N\mathbf{K}\mathbf{1}_N$, where $\mathbf{1}_N$ is an $N \times N$ matrix such that $(\mathbf{1}_N)_{ij} = 1/N$ (see [20] for details).

C. Pseudo Code of the Proposed Algorithm

The whole cohort-based kernel visualisation algorithm is described in figure 3. After computing the orthonormalised cohort means in the feature space $\hat{\mathbf{B}}$, the projected data is then given as $\mathbf{X}^c = \mathbf{\Upsilon}\hat{\mathbf{B}}$, where $\mathbf{\Upsilon}$ is computed with Eq. (7)

```
Given $\mathbf{X} = \{\boldsymbol{x}_i\}_{i=1}^{N}$,
    0. Represent the mean of class $C_j$ in the feature space
           $m_j$ in dual form with Eq. (9)
    1. If sphering, compute the final inner product matrix
           $\boldsymbol{\Upsilon}$ with Eq. (7), otherwise Eq. (6)
    2. Obtain the orthonormalised class means in the
           feature space: $\hat{\mathbf{B}}^T$
    3. Project data onto the orthonormalised cohort means
           in the feature space: $\mathbf{X}^c = \boldsymbol{\Upsilon}\hat{\mathbf{B}}$
    4. Compute scatter matrices for $\mathbf{X}^c$: $\mathbf{S}_W^c$, $\mathbf{S}_B^c$ and $\mathbf{M}^c$
    5. Project $\mathbf{X}^c$ onto the eigenvectors of the largest
           eigenvalues of $\mathbf{M}^c$
```

Fig. 3. Cohort-based kernel visualisation algorithm.

or (6) depending on whether the data is sphered or not. Note that the only parameters of the method are the parameters of the kernel. The projection of a new point $\boldsymbol{y}$ (in a test set, for example) can be computed as follows: (1) compute a row vector $\boldsymbol{Y} = (K(\boldsymbol{x}_1, \boldsymbol{y}), K(\boldsymbol{x}_2, \boldsymbol{y}), \ldots, K(\boldsymbol{x}_N, \boldsymbol{y}))$; (2) compute $\boldsymbol{Y}^c = \boldsymbol{Y}\hat{\mathbf{B}}$; and (3) project $\boldsymbol{Y}^c$ onto the eigenvectors of the largest eigenvalues of $\mathbf{M}^c$.

IV. Experiments and Results

A. Data Sets

The data used in the experiments were: the Wisconsin Breast Cancer data set from the UCI repository [2], Magnetic Resonance Spectroscopy (MRS) data from the INTERPRET project [1], and breast cancer survival prognostic data ([9], [8]). A brief description of the data is provided in table I.

The MRS data analyzed in this study were extracted from an international and multi-centre web-accessible database resulting from the International Network for Pattern Recognition of Tumours Using Magnetic Resonance (INTERPRET) European research project [1]. These data correspond to 304 single voxel short echo time ^{1}H MR spectra acquired in vivo from brain tumour patients, out of which 115 are used in this study: glioblastomas (86) and astrocytomas grade II and III (29). For details on data acquisition and processing, see [12]. Class labelling was performed according to the World Health Organization (WHO) system for diagnosing brain tumours by histopathological analysis of a biopsy sample. The clinically-relevant regions of the spectra were sampled to obtain 195 frequency intensity values (data covariates), from 4.25 parts per million (ppm) down to 0.56 ppm. This data set will referred to as MRS-Ast-Gl.

For the Prognostic data, two data sets were used. The first one was obtained by survival analysis of 743 breast cancer case records from Christie Hospital (CH), Manchester, recruited between 1990-93, following the method described in [8]. The second one was a database of 4,016 cases acquired by the British Columbia Cancer Agency (BCCA), Vancouver, during the period 1989-93. The CH and BCCA data set were used as training and validation data sets, respectively. The CH data set includes 16 explanatory variables, in addition to outcome variables. The BCCA data set contains 10 explanatory variables, as well as the outcome variables. Model selection was carried out through Cox regression (proportional hazards) [9], where six predictive variables were identified: age at diagnosis, node stage, histological type, ratio of axillary nodes affected to axilar nodes removed, pathological size (i.e. tumour size in cm) and oestrogen receptor count. Using the CH data set and a prognostic index derived from the survival predictions for each patient at 5 years of follow-up, a stratification methodology was developed, based on regression trees, to group the patients into four different prognostic risk groups [8]. This obtained stratification methodology obtained with the CH data set was then applied to the BCCA data set. The covariates used for visualisation comprise the six original predictive variables. The class label was the prognostic index.

TABLE I

Description of the data sets.

Data Set	#Covariates	#Classes	#Examples
Wisconsin	9	2	699
MRS-Ast-Gl	195	2	115
Prognostic CH (training)	6	4	743
Prognostic BCCA (validation)	6	4	4,016

B. Experimental Setting

The Gaussian kernel $k(x, y) = e^{-\gamma\|x-y\|^2}$ was the non-linear kernel function used. In order to obtain the kernel parameters for the projection, a *grid search* was performed with γ ranging from 2^{-20} to 2^{10} and C ranging from 2^{-10} to 2^{10} for standard 1-norm soft margin Support Vector Machines (SVM) with the LIBSVM software [5]. The parameters corresponding to the best 10-fold cross-validation accuracy were kept to build the subsequent models. For the Prognostic data, only the CH data set was used to obtain the parameters.

Binary data (Wisconsin and MRS-Ast-Gl) was split into six categories, previous to the projection with CVSM and CKVSM, with the output values of the SVM model obtained:

1) New class 1: Misclassified points of original class 1
2) New class 2: Correctly classified of original class 1 under the median
3) New class 3: Correctly classified of original class 1 over the median
4) New class 4: Misclassified points of original class 2
5) New class 5: Correctly classified of original class 2 under the median
6) New class 6: Correctly classified of original class 2 over the median

For non-binary data (Prognostic), the original classes were used for the projection.

Data were projected with the algorithms for CVSM and CKVSM described in figures 1 and 3. The first two or three components were selected for visualisation. No sphering of the data was performed. For the Prognostic data, the directions of projection were computed with th CH data set and applied to the BCCA data set.

C. Results

For comparative purposes, a *grid search* was performed with linear SVM, with C ranging from 2^{-10} to 2^{10}. Table II shows the SVM accuracies for the selected data sets with the original classes. The visualisations of the projected data can be seen in figure 4. The relationship between classes and colors is: 1-red, 2-green, 3-blue, 4-yellow, 5-magenta and 6-cyan.

For the Wisconsin data, the Gaussian and linear kernels are similar for separating the data points by class label. It is interesting that one of the original binary groups has a much wider spread than the other. This would be apparent without a low-dimensional visualisation of the complete data set.

In the case of the MRS data, the linear kernel gives a better impression of the continuum of tumours. In fact, in the selected view the y-axis is a good classification direction, since allocating low and medium (astrocytomas) vs high grade (glioblastomas) astrocytic tumours either side of a threshold around 0 seems to have about the same classification performance as the original classifier.

In the case of the prognostic data set, the linear kernel shows the categorical nature of the data with some stratification of red-green-yellow, but the category blue does not separate from the others. Therefore the Gaussian kernel plots are better for visualisation with reference to membership of each risk groups, labelled 1:4. This is especially the case for 3D projections, where the data can be rotated to better place each individual case in the context of its neighbours and their risk group allocations.

V. Conclusions

This paper combines kernel methods with dimensionality reduction using class means, to visualise complete data bases of medical data. The visualisation plots give a good indication of the effect of the kernels on the data distribution vs. the classification label, being useful also to directly see how each individual case is located in relation to decision boundaries between classes or specific probabilistic bands. The banding of predicted class membership probabilities makes it possible to derive 3D views of the data in the case of binary classification.

It is proposed as a novel tool for the direct visualisation of complex data, in a way that is meaningful for clinical users. In particular, it provides linearly separable renditions of non-linearly separable data sets, opening a window into the classification of the data, which is not readily available by other methods. In addition, this methodology also permits the exploration of different classification kernels.

Acknowledgement

This work was supported in part by the Ministerio de Ciencia e Innovación (MICINN), under project TIN2009-13895-C02-01. Authors gratefully acknowledge the former INTERPRET (EU-IST-1999-10310) European project partners. Data providers: Dr. C. Majós (IDI), Dr. À.Moreno-Torres (CDP), Dr. F.A. Howe and Prof. J. Griffiths (SGUL), Prof. A. Heerschap (RU), Dr. W. Gajewicz (MUL) and Dr. J. Calvar (FLENI); data curators: Dr. A.P. Candiota, Ms. T. Delgado, Ms. J. Martín, Mr. I. Olier and Mr. A. Pérez (all from GABRMN-UAB). C. Arús and M. Julià-Sapé are funded by the CIBER of Bioengineering, Biomaterials and Nanomedicine, an initiative of the Instituto de Salud Carlos III (ISCIII) of Spain.

References

[1] International Network for Pattern Recognition of Tumours Using Magnetic Resonance (INTERPRET) project. URL: http://azizu.uab.es/INTERPRET.

[2] A. Asuncion and D. J. Newman. UCI machine learning repository, 2007. University of California, Irvine, School of Information and Computer Science. `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

[3] G. Baudat and F. Anouar. Generalized Discriminant Analysis using a Kernel Approach. *Neural Computation*, 11(2):483–497, 2000.

[4] C. M. Bishop, M. Svensén, and C. K. I. Williams. GTM: The Generative Topographic Mapping. *Neural Computation*, 10(1):215–234, 1998.

[5] C. C. Chang and C. J. Lin. LIBSVM: A Library for Support Vector Machines. http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[6] T. F. Cox and M. A. A. Cox. *Multidimensional Scaling*. Chapman and Hall, UK, 2001.

[7] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley, NY, 1973.

[8] A. S. Fernandes, D. Bacciu, I. H. Jarman, T. A. Etchells, J. M. Fonseca, and P. J. G. Lisboa. Different Methodologies for Patient Stratification using Survival Data. In *Conference on Computational Intelligence Methods for Bioinformatics and Biostatistics*, pages 989–996, 2009.

[9] A. S. Fernandes, I. H. Jarman, T. A. Etchells, J. M. Fonseca, E. Biganzoli, C. Bajdik, and P. J. G. Lisboa. Missing Data Imputation in Longitudinal Cohort Studies - Application of PLANN-ARD in Breast Cancer Survival. In *International Conference on Machine Learning and Applications*, pages 644–649, 2008.

[10] H. P. Friedman and J. Rubin. On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62(320):1159–1178, 1967.

[11] K. R. Gabriel. The biplot graphical display of matrices with applications to principal component analysis. *Biometrika*, 58(3):453–467, 1971.

[12] M. Julià-Sapé, D. Acosta, M. Mier, C. Arús, D. Watson, and the INTERPRET Consortium. A Multi-centre, Web-accessible and Quality Control-checked Database of in Vivo MR Spectra of Brain Tumour Patients. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 19(1):22–23, 2006.

[13] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, 1982.

[14] J. B. Kruskal. Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[15] J. A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Springer-Verlag, NY, 2007.

[16] P. J. G. Lisboa, I. O. Ellis, A. R. Green, F. Ambrogi, and M. B. Dias. Cluster-based Visualisation with Scatter Matrices. *Pattern Recognition Letters*, 29(13):1814–1823, 2008.

[17] E. Pękalska and B. Haasdonk. Kernel Discriminant Analysis for Positive Definite and Indefinite Kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(6):1017–1032, 2009.

[18] J. W. Sammon. A Non-linear Mapping for Data Structure Analysis. *IEEE Transactions on Computers*, C-18:401–408, 1969.

[19] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.

[20] B. Schölkopf, A. J. Smola, and K. R. Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10(5):1299–1319, 1998.

TABLE II

ACCURACIES AND CONFUSION MATRICES OF SVM WITH THE SELECTED DATA SETS (ORIGINAL CLASSES).

	Accuracy	Confusion Matrix
Wisconsin (linear SVM)	96.85	[446 12; 10 231]
Wisconsin (Gaussian SVM)	97.28	[445 13; 6 235]
MRS-Ast-Gl (linear SVM)	92.17	[25 4; 5 81]
MRS-Ast-Gl (Gaussian SVM)	91.30	[22 7; 3 83]
Prognostic BCCA (linear SVM)	83.42	[1451 59 0 0; 98 945 173 0; 0 111 521 0; 0 56 169 433]
Prognostic BCCA (Gaussian SVM)	99.15	[1508 2 0 0; 0 1216 0 0; 2 25 605 0; 0 4 1 653]

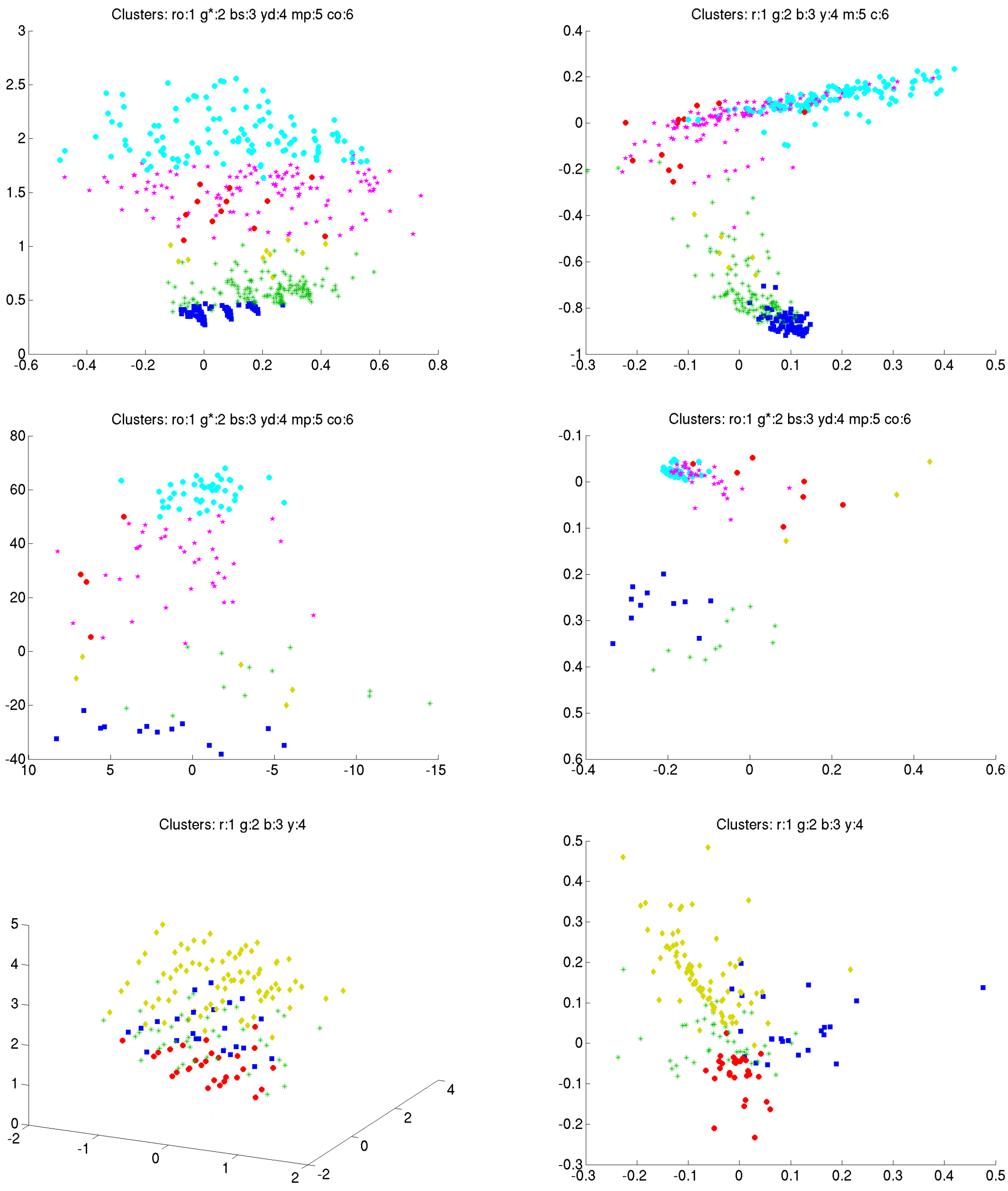

Fig. 4. 2D-visualisations of the 3D-projections for the Wisconsin (top row), MRS-Ast-Gl (middle row) and Prognostic BCCA (bottom row) data sets. Left plots correspond to linear projections and right plots correspond to Gaussian projections.

[21] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290:2319–2323, 2000.

[22] W. S. Torgerson. Multidimensional Scaling: I. Theory and Method. *Psychometrika*, 17(4):401–419, 1952.

[23] J. G. Wang, Y. S. Lin, W. K. Yang, and J. Y. Yang. Kernel Maximum Scatter Difference Based Feature Extraction and its Application to Face Recognition. *Pattern Recognition Letters*, 29(13):1832–1835, 2008.

[24] L. Wang, K. L. Chan, P. Xue, and L. Zhou. A Kernel-Induced Space Selection Approach to Model Selection in KLDA. *IEEE Transactions on Neural Networks*, 19(12):2116–2131, 2008.

[25] J. Yang, Z. Jin, J. Y. Yang, D. Zhang, and A. F. Frangi. Essence of Kernel Fisher Discriminant: KPCA plus LDA. *Pattern Recognition*, 37(10):2097–2100, 2004.