Kernel Feature Selection with Side Data using a Spectral Approach

Amnon Shashua[1] and Lior Wolf[2]

[1] School of Engineering and Computer Science, The Hebrew University, Jerusalem
shashua@cs.huji.ac.il,
[2] Center for Biological and Computational Lerning, MIT
liorwolf@mit.edu

Abstract. We address the problem of selecting a subset of the most relevant features from a set of sample data in cases where there are multiple (equally reasonable) solutions. In particular, this topic includes on one hand the introduction of hand-crafted kernels which emphasize certain desirable aspects of the data and, on the other hand, the suppression of one of the solutions given "side" data, i.e., when one is given information about undesired aspects of the data. Such situations often arise when there are several, even conflicting, dimensions to the data. For example, documents can be clustered based on topic, authorship or writing style; images of human faces can be clustered based on illumination conditions, facial expressions or by person identity, and so forth.
Starting from a spectral method for feature selection, known as $Q - \alpha$, we introduce first a kernel version of the approach thereby adding the power of non-linearity to the underlying representations and the choice to emphasize certain kernel-dependent aspects of the data. As an alternative to the use of a kernel we introduce a principled manner for making use of auxiliary data within a spectral approach for handling situations where multiple subsets of relevant features exist in the data. The algorithm we will introduce allows for inhibition of relevant features of the auxiliary dataset and allows for creating a topological model of all relevant feature subsets in the dataset.
To evaluate the effectiveness of our approach we have conducted experiments both on real-images of human faces under varying illumination, facial expressions and person identity and on general machine learning tasks taken from the UC Irvine repository. The performance of our algorithm for selecting features with side information is generally superior to current methods we tested (PCA,OPCA,CPCA and SDR-SI).

1 Introduction

The problem of focusing on the most relevant measurements in a potentially overwhelming quantity of data is fundamental in machine vision and learning. Seeking out the relevant coordinates of a measurement vector is essential for making useful predictions as prediction accuracy drops significantly and training set size might grow exponentially with the growth of irrelevant features. To

add complexity to what already is non-trivial, natural data sets may contain multiple solutions, i.e., valid alternatives for relevant coordinate sets, depending on the task at hand. For example, documents can be analyzed based on topic, authorship or writing style; face images can be classified based on illumination conditions, facial expressions or by person identity; gene expressions levels can be clustered by pathologies or by correlations that also exist in other conditions.

The main running example that we will use in this paper is that of selecting features from an unlabeled (unsupervised) dataset consisting of human frontal faces where the desired features are relevant for inter-person variability. The face images we will use vary along four dimensions; (i) people identity, (ii) facial expressions, (iii) illumination conditions, and (iv) occlusions (see Fig. 1). One could possibly select relevant features for each of the three dimensions of relevance — *the challenge is how to perform the feature selection process on unlabeled data given that there are multiple solutions (in this case four different ones)*?

There are two principal ways to handle this problem. First is by embedding the feature selection algorithm into a higher dimensional space using a hand-crafted kernel function (the so called "kernel design" effort [11]). By selecting the right kernel function it may be possible to emphasize certain aspects of the data and de-emphasize others. Alternatively, the second approach is to introduce the notion of side information which is to provide auxiliary data in the form of an additional dataset which contains only the undesired dimensions of relevance. The feature selection process would then proceed by selecting features that enhance general dimensions of relevancy in the main dataset while inhibiting the dimensions of relevance in the auxiliary dataset.

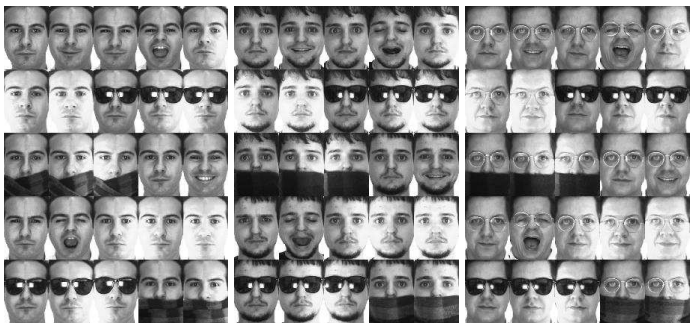


Fig. 1. 25 out of the 26 images in the AR dataset for three different persons. Images vary not only in person identity but also in illumination, facial expression, and amount and type of occlusion.

In this work we address both approaches. We start with the principle of spectral-based feature selection (introduced by [19]) and modify it to serve two new purposes: (i) endowing the approach with the power of kernel functions, satisfying the first approach for enriching the vector representation, and (ii) making use of auxiliary data for situations in which multiple subsets of relevant features exist in the data. The algorithm we will introduce allows for inhibition of relevant features of the auxiliary dataset and allows for creating a topological model of all relevant feature subsets in the dataset. The auxiliary dataset we consider could come in two different forms: the first being additional data points

which represent undesired variability of the data, while the second form of side data consists of pairs of points which belong to different classes of variability, i.e., are considered far away from each other in the space of selected coordinates.

Side information (a.k.a "irrelevant statistics" or "background information") appears in various contexts in the literature — clustering [20, 1, 14] and continuous dimensionality reduction [15, 5]. In this paper we address the use of side information in the context of a hard selection of a feature subset. Feature selection from unlabeled data differs from dimensionality reduction in that it only selects a handful of features which are "relevant" with respect to some inference task. Dimensionality reduction algorithms, for example PCA, generate a small number of features each of which is a combination of all of the original features. In many situations of interest, in visual analysis in particular but also in other application domains such as Genomics for instance, it is assumed that each process being studied involves only a limited number of features taken from a pool of a very large set of measurements. For this reason feature combination methods are not as desirable as methods that extract a small subset of features. The challenge in the selection process is to overcome the computational burden of pruning an exponential amount of feature subsets. The $Q-\alpha$ algorithm [19] which we propose using as a basis for our approach handles the exponential search space by harnessing the spectral information in such a manner where a computationally straightforward optimization guarantees a sparse solution, i.e., a selection of features rather than a combination of the original features.

In the subsection below we will describe the $Q-\alpha$ algorithm which forms the background for the work presented in this paper. In Section 2 we derive a kernel method version of the $Q-\alpha$ algorithm which enables the representation of high order cumulants among the entries of the feature vectors thereby considerably strengthening the feature selection methodology. In Section 3 we introduce the auxiliary data matrix as a side data and derive the optimization for selecting relevant features using the main dataset while inhibiting relevant features from the auxiliary dataset. In Section 4 we take the notion of auxiliary dataset a step further and form a complete topographical model of the relevant feature subsets. The general idea is based on rounds where the relevant features selected in previous rounds form "side" information for subsequent rounds. In this manner a hierarchical modeling the feature subsets becomes feasible and can be used for visualization and data modeling. In Section 5 we make use of another form of side information where the auxiliary data consists of pairs of points which belong to different classes of variability, i.e., are considered far away from each other in the space of selected coordinates. In Section 6 we evaluate the effectiveness of our algorithms by experiments on various datasets including real-image experiments on our main running example, and also running examples on general machine learning tasks taken from the UC Irvine repository.

1.1 Selecting Relevant Features with the $Q-\alpha$ Algorithm

The $Q-\alpha$ algorithm for unsupervised feature selection is based on the assumption that the selection of the relevant features (coordinates) will result in a

coherent set of clusters formed by the input data points restricted to the selected coordinates. The clustering score in this approach is measured indirectly. Rather than explicitly performing a clustering phase per feature selection candidates, one employs spectral information in order to measure the cluster arrangement coherency. Spectral algorithms have been proven to be successful in clustering [16], manifold learning or dimensionality reduction [12], approximation methods for NP-hard graph theoretical questions. In a nutshell, given a selection of features, the strength (magnitude) of the leading k eigenvalues of the affinity matrix constructed from the corresponding feature values across the sample data are directly related to the coherence of the cluster arrangement induced by the subset of selected features. The scheme is described as follows:

Let the data matrix be denoted by M. The feature values form the rows of M denoted by $\mathbf{m}_1^\top, ..., \mathbf{m}_n^\top$ and normalized to unit norm $\|\mathbf{m}_i\| = 1$. Each row vector represents a feature (coordinate) sampled over the q trials. The column vectors of M represent the q samples (each sample is a vector in R^n). For example, a column can represent an image represented by its pixel values and a row can represent a specific pixel location whose value runs over the q images. As mentioned in the previous section, our goal is to select rows (features) from M such that the corresponding candidate data matrix (containing only the selected rows) consists of columns that are coherently clustered in k groups. The value of k is user dependent and is specific to the task at hand. The challenge in this approach is to avoid the exponential number of row selections and preferably avoid explicitly clustering the columns of the data matrix per each selection.

Mathematically, to obtain a clustering coherency score we compute the "affinity" matrix of the candidate data matrix defined as follows. Let $\alpha_i \in \{0, 1\}$ be the indicator value associated with the i'th feature, i.e., $\alpha_i = 1$ if the i'th feature is selected and zero otherwise. Let A_α be the corresponding affinity matrix whose (i, j) entries are the inner-product (correlation) between the i'th and j'th columns of the resulting candidate data matrix: $A_\alpha = \sum_{i=1}^n \alpha_i \mathbf{m}_i \mathbf{m}_i^\top$ (sum of rank-1 matrices). From algebraic graph theory, if the columns of the candidate data matrix are coherently grouped into k clusters, we should expect the leading k eigenvalues of A_α to be of high magnitude [8, 10, 2, 16]. The resulting scheme should therefore be to maximize the sum of eigenvalues of the candidate data matrix over all possible settings of the indicator variables α_i.

What is done in practice, in order to avoid the exponential growth of assigning binary values to n indicator variables, is to allow α_i to receive real values in an unconstrained manner. A least-squares energy function over the variables α_i is formed and its optimal value is sought after. What makes this approach different from the "garden variety" soft-decision-type algorithms is that this particular setup of optimizing over spectral properties guarantees that the α_i *always come out positive and sparse* over all local maxima of the energy function. This property is intrinsic rather than being the result of explicit constraints in the form of regularizers, priors or inequality constraints. We optimize the following:

$$\max_{Q,\alpha_i} \text{trace}(Q^\top A_\alpha^\top A_\alpha Q) \quad \text{subject to} \quad \sum_{i=1}^n \alpha_i^2 = 1, \;\; Q^\top Q = I \qquad (1)$$

Note that the matrix Q holds the first k eigenvectors of A_α and that $\text{trace}(Q^\top A_\alpha^\top A_\alpha Q)$ is equal to the sum of squares of the leading k eigenvalues: $\sum_{j=1}^{k} \lambda_j^2$. A local maximum of the energy function is achieved by interleaving the "orthogonal iteration" scheme [6] within the computation of α as follows:

Definition 1 ($Q - \alpha$ Method). *Let M be an $n \times q$ input matrix with rows $\mathbf{m}_1^\top, ..., \mathbf{m}_n^\top$, and some orthonormal $q \times k$ matrix $Q^{(0)}$, i.e., ${Q^{(0)}}^\top Q^{(0)} = I$. Perform the following steps through a cycle of iterations with index $r = 1, 2, ...$*

1. *Let $G^{(r)}$ be a matrix whose (i, j) components are*

$$(\mathbf{m}_i^\top \mathbf{m}_j)\mathbf{m}_i^\top Q^{(r-1)} {Q^{(r-1)}}^\top \mathbf{m}_j.$$

2. *Let $\alpha^{(r)}$ be the leading eigenvector of $G^{(r)}$.*
3. *Let $A^{(r)} = \sum_{i=1}^{n} \alpha_i^{(r)} \mathbf{m}_i \mathbf{m}_i^\top$.*
4. *Let $Z^{(r)} = A^{(r)} Q^{(r-1)}$.*
5. *$Z^{(r)} \xrightarrow{QR} Q^{(r)} R^{(r)}$, that is, $Q^{(r)}$ is determined by the "QR" factorization of $Z^{(r)}$.*
6. *Increment index r and go to step 1.*

Note that steps 4,5 of the algorithm consist of the "orthogonal iteration" module, i.e., if we were to repeat steps 4,5 *only* we would converge onto the eigenvectors of $A^{(r)}$. However, the algorithm does not repeat steps 4,5 in isolation and instead recomputes the weight vector α (steps 1,2,3) before applying another cycle of steps 4,5.

The algorithm would be meaningful provided that three conditions are met:

1. the algorithm converges to a local maximum,
2. at the local maximum $\alpha_i \geq 0$ (because negative weights are not admissible), and
3. the weight vector α is *sparse* (because without it the soft decision does not easily translate into a hard gene selection).

Conditions (2) and (3) are not readily apparent in the formulation of the algorithm (the energy function lacks the explicit inequality constraint $\alpha_i \geq 0$ and an explicit term to "encourage" sparse solutions) but are nevertheless satisfied. The key for having sparse and non-negative (same sign) weights is buried in the matrix G (step 1). Generally, the entries of G are not necessarily positive (otherwise α would have been non-negative due to the Perron-Frobenious theorem) — nevertheless due its makeup it can be shown that in a probabilistic manner the leading eigenvector of G is positive with probability $1 - o(1)$. In other words, as the number of features n grows larger the chances that the leading eigenvector of G is positive increases rapidly to unity. The details of why the makeup of G induces such a property, the convergence proof and the proof of the "Probabilistic Perron-Frobenious" claim can be found in [19].

Finally, it is worth noting that the scheme can be extended to handle the supervised situation (when class labels are provided); that the scheme can be applied also to the Laplacian affinity matrix; and that the scheme readily applies when the spectral gap $\sum_{i=1}^{k} \lambda_i^2 - \sum_{j=k+1}^{q} \lambda_j^2$ is maximized rather than $\sum_{i=1}^{k} \lambda_i^2$ alone. Details can be found in [19].

2 Representing Higher-order Cumulants using Kernel Methods

The information on which the $Q-\alpha$ method relies on to select features is contained in the matrix G. Recall that the criterion function underlying the $Q-\alpha$ algorithm is a sum over all pairwise feature vector relations:

$$trace(Q^\top A_\alpha^\top A_\alpha Q) = \alpha^\top G \alpha,$$

where G is defined such that $G_{ij} = (\mathbf{m}_i^\top \mathbf{m}_j)\mathbf{m}_i^\top Q Q^\top \mathbf{m}_j$. It is apparent that feature vectors interact in pairs and the interaction is *bilinear*. Consequently, cumulants of the original data matrix M which are of higher order than two are not being considered by the feature selection scheme. For example, if M were to be decorrelated (i.e., $MM^\top$ is diagonal) the matrix G would be diagonal and the feature selection scheme would select only a single feature.

In this section we employ the "kernel trick" to include cumulants of higher orders among the feature vectors in the feature selection process. This serves two purposes: On one hand the representation is enriched with non-linearities induced by the kernel, and on the other hand, given a successful choice of a kernel (so called Kernel Design effort [11]) one could possibly emphasize certain desirable aspects of the data while inhibiting others.

Kernel methods in general have been attracting much attention in the machine learning literature — initially with the support vector machines [13] and later took a life of their own (see [11]). Mathematically, the kernel approach is defined as follows: let $\mathbf{x}_1, ..., \mathbf{x}_l$ be vectors in the input space, say R^q, and consider a mapping $\phi(\mathbf{x}) : R^q \to \mathcal{F}$ where $\mathcal{F}$ is an inner-product space. The kernel-trick is to calculate the inner-product in $\mathcal{F}$ using a kernel function $k : R^q \times R^q \to R$, $k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$, while avoiding explicit mappings (evaluation of) $\phi()$. Common choices of kernel selection include the d'th order polynomial kernels $k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^\top \mathbf{x}_j + c)^d$ and the Gaussian RBF kernels $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{1}{2\sigma^2}\|\mathbf{x}_i - \mathbf{x}_j\|^2)$. If an algorithm can be restated such that the input vectors appear in terms of inner-products only, one can substitute the inner-products by such a kernel function. The resulting kernel algorithm can be interpreted as running the original algorithm on the space $\mathcal{F}$ of mapped objects $\phi(\mathbf{x})$. Kernel methods have been applied to the support vector machine (SVM), principal component analysis (PCA), ridge regression, canonical correlation analysis (CCA), QR factorization and the list goes on. We will focus below on deriving a kernel method for the $Q-\alpha$ algorithm.

2.1 Kernel $Q-\alpha$

We will consider mapping the rows $\mathbf{m}_i^\top$ of the data matrix M such that the rows of the mapped data matrix become $\phi(\mathbf{m}_1)^\top, ..., \phi(\mathbf{m}_n)^\top$. Since the entries of G consist of inner-products between pairs of mapped feature vectors, the interaction will be no longer bilinear and will contain higher-order cumulants whose nature depends on the choice of the kernel function.

Replacing the rows of M with their mapped version introduces some challenges before we could apply the kernel trick. The affinity matrix $A_\alpha = \sum_i \alpha_i \phi(\mathbf{m}_i)\phi(\mathbf{m}_i)^\top$ cannot be explicitly evaluated because A_α is defined by *outer-products* rather than inner-products of the mapped feature vectors $\phi(\mathbf{m}_i)$. The matrix Q holding the eigenvectors of A_α cannot be explicitly evaluated as well and likewise the matrix $Z = A_\alpha Q$ (in step 4). As a result, kernelizing the $Q-\alpha$ algorithm requires one to represent α without explicitly representing A_α and Q both of which were instrumental in the original algorithm. Moreover, the introduction of the kernel should be done in such a manner to preserve the key property of the original $Q-\alpha$ algorithm of producing a sparse solution.

Let $V = MM^\top$ be the $n \times n$ matrix whose entries are evaluated using the kernel $v_{ij} = k(\mathbf{m}_i, \mathbf{m}_j)$. Let $Q = M^\top E$ for some $n \times k$ (recall k being the number of clusters in the data) matrix E. Let $D_\alpha = diag(\alpha_1, ..., \alpha_n)$ and thus $A_\alpha = M^\top D_\alpha M$ and $Z = A_\alpha Q = M^\top D_\alpha V E$. The matrix Z cannot be explicitly evaluated but $Z^\top Z = E^\top V D_\alpha V D_\alpha V E$ can be evaluated. The matrix G can be expressed with regard to E instead of Q:

$$\begin{aligned} G_{ij} &= (\phi(\mathbf{m}_i)^\top \phi(\mathbf{m}_j))\phi(\mathbf{m}_i)^\top Q Q^\top \phi(\mathbf{m}_j) \\ &= k(\mathbf{m}_i, \mathbf{m}_j)\phi(\mathbf{m}_i)^\top (M^\top E)(M^\top E)^\top \phi(\mathbf{m}_j) \\ &= k(\mathbf{m}_i, \mathbf{m}_j)\mathbf{v}_i^\top E E^\top \mathbf{v}_j \end{aligned}$$

where $\mathbf{v}_1, ..., \mathbf{v}_n$ are the columns of V. Step 5 of the $Q-\alpha$ algorithm consists of a QR factorization of Z. Although Z is uncomputable it is possible to compute R and R^{-1} directly from the entries of $Z^\top Z$ without computing Q using the Kernel Gram-Schmidt described in [18]. Since $Q = ZR^{-1} = M^\top D_\alpha V E R^{-1}$ the update step is simply to replace E with ER^{-1} and start the cycle again. In other words, rather than updating Q we update E and from E we obtain G and from there the newly updated α. The kernel $Q-\alpha$ is summarized below:

Definition 2 (Kernel $Q-\alpha$). *Let M be an uncomputable matrix with rows $\phi(\mathbf{m}_1)^\top, ..., \phi(\mathbf{m}_n)^\top$. The kernel function is given by $\phi(\mathbf{m}_i)^\top \phi(\mathbf{m}_j) = k(\mathbf{m}_i, \mathbf{m}_j)$. The matrix $V = MM^\top$ is a computable $n \times n$ matrix. Let $E^{(0)}$ be an $n \times k$ matrix selected such that $M^\top E^{(0)}$ has orthonormal columns. Iterate over the steps below, with the index $r = 1, 2, ...$*

1. *Let $G^{(r)}$ be a $n \times n$ matrix whose (i,j) components are $k(\mathbf{m}_i, \mathbf{m}_j)\mathbf{v}_i^\top E^{(r-1)} {E^{(r-1)}}^\top \mathbf{v}_j$.*
2. *Let $\alpha^{(r)}$ be the largest eigenvector of $G^{(r)}$, and let $D^{(r)} = diag(\alpha_1^{(r)}, ..., \alpha_n^{(r)})$.*
3. *Let $Z^{(r)}$ be an uncomputable matrix*

$$Z^{(r)} = (M^\top D^{(r)} M)(M^\top E^{(r-1)}) = M^\top D^{(r)} V E^{(r-1)}.$$

4. *$Z^{(r)} \xrightarrow{QR} QR$. It is possible to compute directly R, R^{-1} from the entries of the computable matrix ${Z^{(r)}}^\top Z^{(r)}$ without explicitly computing the matrix Q (see [18]).*
5. *Let $E^{(r)} = E^{(r-1)} R^{-1}$.*
6. *Increment index r and go to step 1.*

The result of the algorithm is the weight vector α and the design matrix G which contains all the data about the features. The drawback of the kernel approach for handling multiple structures of the data is that the successful choice

of a kernel depends on the user and is largely an open problem. For example, with regard to our main running example it is unclear which kernel to choose that will strengthen the clusters induced by inter-personal variation and inhibit the clusters induced by lighting facial expressions. We therefore move our attention to the alternative approach using the notion of side data.

3 $Q-\alpha$ with Side Information

Consider the $n \times q$ data matrix M defined above as the "main" data. We are given an auxiliary $n \times p$ data matrix W with rows $\mathbf{w}_1^\top, ..., \mathbf{w}_n^\top$ representing p data points comprising the "side" information. Our goal is to select a subset of coordinates, namely, determine the weight vector α such the affinity matrix $\sum_i \alpha_i \mathbf{m}_i \mathbf{m}_i^\top$ has coherent k clusters (measured by the sum of squares of the first k eigenvalues) whereas $\sum_i \alpha_i \mathbf{w}_i \mathbf{w}_i^\top$ has low cluster coherence. The desire for low cluster coherence for the side information can be represented by small variance of each coordinate value along the p samples. Namely, if $\mathbf{m}_i$ is selected as a relevant feature of the main data, we should expect that the corresponding side feature vector $\mathbf{w}_i$ will have a small variance. Small variance of the selected rows of W means that the corresponding affinity matrix $\sum_i \alpha_i \mathbf{w}_i \mathbf{w}_i^\top$ represents a single cluster (whether coherent or not is immaterial).

To clarify the logic behind our approach, consider the scenario presented in [5]. Assume we are given face images of 5 individuals covering variability of illumination and facial expressions — a total of 26 images per individual. The main data matrix M will contain therefore 130 columns. We wish to select relevant features (rows of M), however, there are three dimensions of relevancy: (i) person identity, (ii) illumination direction, and (iii) facial expressions. One could possibly select relevant features for each dimension of relevance and obtain a coherent clustering in that dimension. Say we are interested in the person identity dimension of relevance. In that case the auxiliary matrix W will contain 26 images of a 6th individual (covering facial expressions and illumination conditions). Features selected along the dimensions of facial expression or illumination will induce coherent clusters in the side data, whereas features selected along the person identity dimension will induce a single cluster (or no structure at all) in the side data — and low variance of the feature vectors is indicative to single cluster or no structure at all. In formal notations we have the following:

Let $D = diag(var(\mathbf{w}_1^\top), ..., var(\mathbf{w}_n^\top))$ be a diagonal matrix with the variance of the rows of W. The low coherence desire over the side data translates to minimization of $\alpha^\top D \alpha$. Taken together, we have a Rayleigh quotient type of energy function to maximize:

$$\max_{Q,\alpha_i} \frac{trace(Q^\top A_\alpha^\top A_\alpha Q)}{\alpha^\top (D + \lambda I)\alpha} = \frac{\alpha^\top G \alpha}{\alpha^\top (D + \lambda I)\alpha} \qquad (2)$$

$$subject\ to \quad \sum_{i=1}^{n} \alpha_i^2 = 1, \quad Q^\top Q = I$$

where G is the matrix defined above whose entries are: $G_{ij} = (\mathbf{m}_i^\top \mathbf{m}_j)\mathbf{m}_i^\top QQ^\top \mathbf{m}_j$. The scalar $\lambda \geq 0$ is user-setabale with the purpose of providing the tradeoff between the main data and the side data. Large values of λ translates to low weight for the side information in the feature selection scheme. A vanishing value $\lambda = 0$ is admissible provided that none of the variances vanishes (D has no vanishing entries along its diagonal) — in that case equal weight is given to the two sources of data. The $Q - \alpha$ with side information algorithm becomes:

Definition 3 ($Q-\alpha$ with Side Information). *Let M be an $n \times q$ input matrix with rows $\mathbf{m}_1^\top, ..., \mathbf{m}_n^\top$, W be an $n \times p$ "side" matrix where the variance of its rows form a diagonal matrix D, and $Q^{(0)}$ is some orthonormal $q \times k$ matrix, i.e., ${Q^{(0)}}^\top Q^{(0)} = I$. Perform the following steps through a cycle of iterations with index $r = 1, 2, ...$*

1. *Let $G^{(r)}$ be a matrix whose (i, j) components are $(\mathbf{m}_i^\top \mathbf{m}_j)\mathbf{m}_i^\top Q^{(r-1)}{Q^{(r-1)}}^\top \mathbf{m}_j$.*
2. *Let $\alpha^{(r)}$ be the largest eigenvector of $(D + \lambda I)^{-1}G^{(r)}$.*
3. *Let $A^{(r)} = \sum_{i=1}^n \alpha_i^{(r)} \mathbf{m}_i \mathbf{m}_i^\top$.*
4. *Let $Z^{(r)} = A^{(r)}Q^{(r-1)}$.*
5. *$Z^{(r)} \xrightarrow{QR} Q^{(r)}R^{(r)}$.*
6. *Increment index r and go to step 1.*

Note the change in step 2 compared to the $Q - \alpha$ algorithm. Since $D + \lambda I$ is a diagonal positive matrix, its inverse is also positive therefore the positivity of G is not affected. In other words, the properties of G which induce a positive (and sparse) solution for the weight vector α (see [19]) are not negatively affected when G is multiplied with a positive diagonal matrix. If D were not diagonal, then D^{-1} would not have been positive and the optimized α values would not come out positive and sparse.

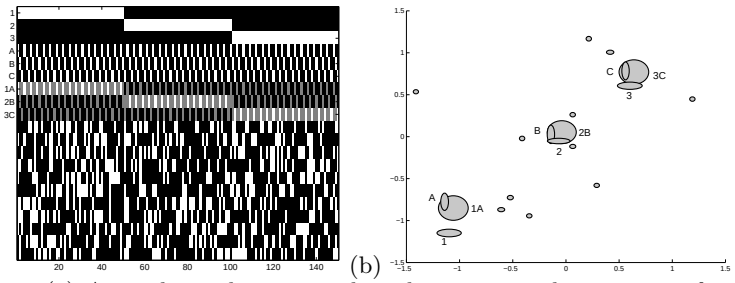
4 Topographical Model of all Relevant Feature Subsets

We can further extend the notion of "negative variability" embedded in the side information to a wider perspective of representing a hierarchy of feature subsets extracted iteratively. The general idea is to treat the weight vector α (which determines the feature selection as it is a sparse positive vector) as representing axes of negative variability for subsequent rounds. Let α be the feature selection solution given by running the $Q-\alpha$ algorithm. We wish to run $Q-\alpha$ again while looking for an alternative solution along a different dimension of variability We construct a "side information" matrix D whose diagonal is $D = diag(\alpha_1^2, ..., \alpha_n^2)$ and run the $Q - \alpha$-with-SI algorithm. The new weight vector α' will be encouraged to have high values in coordinates where α has low values. This is applied iteratively where in each round $Q - \alpha$-with-SI is executed with the matrix D containing the sum of square α values summed over all previous rounds.

Furthermore, the G matrix resulting from each round of the above scheme can be used for generating a coordinization of the features as a function of the implicit clustering of the (projected) data. The weight vector α is the largest eigenvector

of G, but as in Multi-Dimensional-Scaling (MDS), the first largest eigenvectors of G form a coordinate frame. Assume we wish to represent the selected features by a 1D coordinate. This can be achieved by taking the first two largest eigenvectors of G thereby each feature is represented by two coordinates. A 1D representation is made by normalizing the coordinate-pair (i.e., each feature is represented by a direction in the 2D MDS frame). Given r rounds, each feature is represented by r coordinates which can be used for visualization and data modeling.

An example of such a topographical map is shown in figure 2. The data matrix consists of 150 data points each described by 20 features out of which 9 are relevant. The relevant features form two possible solution sets where each solution induces three clusters of data points. The first set consists of three features marked by "1,2,3", while the second set consists of three different features marked by "A,B,C". Three additional features marked by "1A,2B,3C" were constructed by summing the corresponding feature vectors 1,2,3 and A,B,C, respectively. The remaining 11 (irrelevant) features were constructed by randomly permuting the entries of the first feature vector. We ran $Q-\alpha$ twice creating for each feature two coordinates (one per each run) as described above. In addition to the coordinization of each feature we have associated the corresponding α value as a measure of "relevancy" of the feature per solution. Taken together, each feature is represented by a position in the 2D topographical map and a 2D magnitude represented by an ellipse whose major axes capture the respective α values. The horizontal axis in Fig. 2(b) is associated with the solution set of features "1,2,3" and the vertical axis with the solution set "A,B,C". We see that the hybrid features 1A,2B,3C, which are relevant to both cluster configurations, have equal (high) relevancy in both sets (large circles in the topographical map).



(a) (b)

Fig. 2. (a) A synthetic dataset used to demonstrate the creation of a topographical model of the features (b) The resulting topographical model (see text).

5 Pairwise Side Information

Another possible variant of $Q-\alpha$-SI is when the side information is given over pairs of "negative" data points. Consider the (adapted) problem raised by [20] in the context of distance metric learning for clustering: we are given a set of data points forming a data matrix M (the "main" data). As side information we are given pairs of points $\mathbf{x}_i, \mathbf{x}_j$ which are *known* to be part of different clusters. We wish to select features (coordinates) such that the main data contains maximally

coherent clusters while obeying the side information (i.e., features are selected such that for each of the "side" pairs $(\mathbf{x}_i^\top \mathbf{x}_j)^2$ is small).

We can incorporate the side information by constructing a side matrix B which functions similarly to the diagonal matrix D we constructed in the previous sections. The difference here would be that B is not diagonal and therefore needs to be handled differently. Consider a pair of side points $\mathbf{x}, \mathbf{y}$. We wish to find the weight vector α such that: $(\mathbf{x}^\top \mathbf{y})^2 = (\sum_i \alpha_i x_i y_i)^2 = \alpha^\top F \alpha$ is small, where $F_{rs} = x_r y_r x_s y_s$. Denote by F^{ij} the matrix corresponding to the pair of side points $\mathbf{x}_i, \mathbf{x}_j$ and let $B = \sum_i \sum_j F^{ij}$.

Our goal is to maximize the spectral information coming from the main data (as before) while minimizing $\alpha^\top B \alpha$. We are back to the same framework as in Sections 3 and 4 with the difference that B is not diagonal therefore the product $B^{-1}G$ is not guaranteed to obey the properties necessary for the weight vector α to come out positive and sparse. Instead, we define an additive energy function:

$$\max_{Q,\alpha_i} trace(Q^\top A_\alpha^\top A_\alpha Q) - \lambda \alpha^\top B \alpha \tag{3}$$

$$subject\ to \qquad \sum_{i=1}^{n} \alpha_i^2 = 1, \quad Q^\top Q = I$$

This energy function is equivalent to $\alpha^\top (G - \lambda B)\alpha$ where λ tradeoffs the weight given to the side data. The algorithm follows the steps of the $Q - \alpha$ algorithm with the difference in step 2: "$\alpha^{(r)}$ is the largest eigenvector of $G^{(r)} - \lambda B$."

There is an open issue of showing that α comes out positive and sparse. The matrix G is "dominantly positive", i.e., when treated as a random matrix each entry has a positive mean and thus it can be shown that the probability of a positive α asymptotes at unity very fast with n [19]. The question what happens to the mean when one subtracts λB from G. Our working assumption is that the entries of B are significantly smaller than the corresponding entries of G because the inner-products of the side points should be small — otherwise they wouldn't have been supplied as side points. Empirical studies on this algorithm validate this assumption and indeed α maintains the positivity and sparsity properties in our experiments.

6 Experiments

We present below three types of experiments (i) simulations on synthetic data for the purpose of studying the effects of different weightings of the side data, (ii) our main running example on the AR face dataset, and (iii) various examples taken from the UC Irvine repository of data sets. Due to space constraints the synthetic simulations are given only in the technical report [17].

Face images. Our main running example is the selection of features from an unlabeled data set of face images taken from the AR dataset [7]. The dataset consists of 100 people with 26 images per person varying according to lighting direction and facial expressions. Our task is to select those features which are relevant for distinguishing between people identities only. The dataset contains

three dimensions of relevancy, and the use of side data is crucial for inhibiting the unwanted dimensions of facial expressions and lighting variations. Following [5] we adopted the setting where the main data set contained the images of 5 randomly chosen men (out of the 50 men) totaling 130 images. The side dataset consisted of the 26 images of a random sixth man. The feature selection process $Q-\alpha-SI$ looks for coordinates which maximize the cluster coherence of the main dataset while minimizing the variance of the coordinate vectors of the side data. As a result, the selected coordinates are relevant for separating among person identities while being invariant to the other dimensions of variability. The task of clustering those images into the five *correct* clusters is hard since the nuisance structures (such as those generated by variation of lighting and facial expressions) are far more dominant than the structure of person variability.

The feature values we use as a representation of the image is designed to capture the relationship between average intensities of neighboring regions. This suggests the use of a family of basis functions, like the Haar wavelets, which encode such relationships along different orientations (see [9, 4]). In our implementation the Haar wavelet transform is run over an image and results in a set of 5227 coefficients at several scales that indicate the response of the wavelets over the entire image. Many of the coefficients are irrelevant for the task of separating between facial identities and it is therefore the goal of the $Q-\alpha-SI$ to find those coefficients that represent the relevant regions.

To quantify the performance of our algorithm in a comparative manner we used the normalized precision score introduced in [5, 15] which measures the average purity of the k-Nearest Neighbors for varying values of k. We compared the performance to four methods: PCA which is the most popular technique for dimensionality reduction, Constrained PCA (CPCA) and Oriented PCA (OPCA) [3], and Sufficient Dimensionality Reduction with Side Information (SDR-SI) [5]. All but the first method (PCA) utilize the same side data as the $Q-\alpha-SI$. Also worth noting that all the methods we compared to extract features by combinations of the original features rather than just select features.

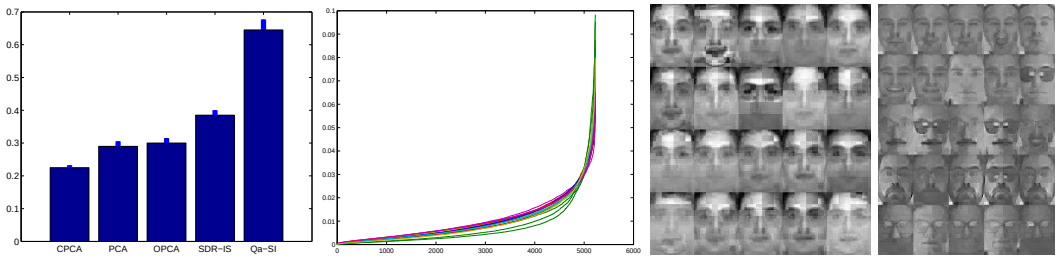
Optimal parameters (dimensionality and λ) for all methods were chosen to maximize the precision index for a training set. The wavelet decomposition was not optimal for the other methods and therefore the raw image intensities were used instead. Reported results were obtained on a separate test set. The entire procedure was repeated 20 times on randomly chosen subsets of the AR database.

Fig. 3a shows the results averaged over 20 runs. The precision index is normalized between 0 to 1 where 0 is obtained with random neighboring and 1 when all nearest neighbors are of the same class. Note that the precision index of $Q-\alpha-SI$ is 0.64 which is significantly higher than 0.39 obtained by the next best method (SDR-SI). Fig. 3(b) shows the resulting α values sorted separately at each one of the 20 runs. As can be seen those values are extremely sparse - having only few of the feature weights above a very clear threshold at each run.

Fig. 3(c) illustrates the selected features by the $Q-\alpha-SI$ at each run. This is done by synthesizing (reconstructing) the images from their wavelet coefficients weighted by the α values. What is shown per run is the average male image.

Fig. 3(d) shows the projection of random faces from a specific run to the weighted features space. Each row contains images of one person. In both figures (c,d) some characteristic features of each individual (beard, dark glasses frame, distinctive hair line) are highlighted, while the illumination differences are reduced.

Finally, it is worth noting that our attempts to find an appropriate kernel which will perform as well as the side data approach were unsuccessful. Our experiments show that the kernel $Q-\alpha$ has significant advantages over $Q-\alpha$ in general, but selecting an appropriate kernel for the multiple structure paradigm is a hard problem and is left open (see [11] for work on kernel design).



(a) (b) (c) (d)

Fig. 3. (a) Comparison of the normalized precision index between CPCA,PCA,OPCA,SDR-IS, and $Q\alpha - SI$ on the AR dataset. (b) Sorted feature weights (α values) for each of the 20 runs showing the sparsity of the feature selection (c) The average image of all men in the AR dataset projected to the selected features for each one of the 20 runs. (d) For a specific run: each row contains the images of one person projected onto the selected feature space.

UC Irvine Repository Tests. We also applied our method to several datasets from the UC Irvine repository. On each dataset we applied k-means clustering on the raw data and on features provided by PCA, OPCA and CPCA. An accuracy score was computed for each clustering result similarly to what was done in [20]. The results are shown for the dermatology, segmentation, wine and ecoli datasets. We also tested the algorithm on the glass, Boston-housing and arrhythmia datasets where non of the algorithms were significantly better than chance. The results are summarized in the table below. Each report result is an average of several experiments where, at turns, each class served as side information and the other classes were taken to be the main dataset. The features were weighted, combined or selected according to the algorithm in question, and then the data points were clustered by k-means. Each result shown in the table was averaged over 20 runs. The number of features used for each PCA variants was the one which gave the best average accuracy. The parameter λ used in the $Q-\alpha$ with side information was fixed at $\lambda = 0.1$.

Dataset	raw data	$Q-\alpha$ SI	PCA	CPCA	OPCA
dermatology	0.5197	0.8816	0.5197	0.6074	0.8050
ecoli	0.6889	0.7059	0.6953	0.6973	0.5620
segmentation	0.7157	0.7817	0.7208	0.7089	0.7110
wine	0.7280	0.9635	0.7280	0.7280	0.9493

The $Q-\alpha$-SI performed the best over all the experiments we conducted. In some of the datasets constrained PCA or oriented PCA performed only slightly worse, but non of these methods gave good results consistently in all four datasets. Unlike PCA and its variants, the $Q-\alpha$ algorithm tends to produce a sparse selection of features, showing a large preference toward a small number of features. For example, in the wine dataset the α values corresponding to the features Alcohol and Proline were three times larger than the rest.

References

1. G. Chechik and N. Tishby. Extracting relevant structures with side information. In *NIPS 2002*.
2. F.R.K. Chung. *Spectral Graph Theory*. AMS, 1998.
3. K.I. Diamantaras and S.Y. Kung *Principal Component Neural Networks: Theory and Applications* NY: Wiley, 1996.
4. C. E. Jacobs, A. Finkelstein, and D. H. Salesin. Fast multiresolution image querying. In SIGGRAPH 1995
5. A. Globerson, G. Chechik, and N. Tishby. Sufficient dimensionality reduction with irrelevance statistics. In *UAI-2003*.
6. G.H. Golub and C.F. Van Loan. *Matrix computations*. , 1989.
7. A.M. Martinez and R. Benavente. The AR face database. Tech. Rep. 24, CVC, 1998.
8. T.S. Motzkin and E.G. Straus. Maxima for graphs and a new proof of a theorem by turan. *Canadian Journal of Math.*, 17:533–540, 1965.
9. M. Oren, C. Papageorgiou, P. Sinha, E. Osuna and T. Poggio. Pedestrian Detection Using Wavelet Templates. In *CVPR 1997*.
10. M. Pavan and M. Pelillo. A new graph-theoretic approach to clustering and segmentation. In *CVPR*, 2003.
11. B. Scholkopf and A.J. Smola. *Learning with Kernels* The MIT press, 2002.
12. Joshua B. Tenenbaum A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, (2000).
13. V. N. Vapnik. *The nature of statistical learning*. Springer, 2nd edition, 1998.
14. K. Wagstaf, A. Cardie, S. Rogers, and S. Schroedl. Constrained K-means clustering with background knowledge. In *ICML-2001*.
15. D. Weinshall, N. Shental, T. Hertz, and M. Pavel. Adjustment learning and relevant component analysis. In *ECCV*, 2002.
16. A.Y. Ng, M.I. Jordan and Y. Weiss. On Spectral Clustering: Analysis and an algorithm. *NIPS*, 2001.
17. A. Shashua amd L. Wolf Sprse Spectral-based Feature Selection with Side Information. TR 2003-57, Leibniz Center for Research, HUJI, 2003.
18. L. Wolf and A. Shashua. Kernel principal angles for classification machines with applications to image sequence interpretation. In *CVPR*, 2003.
19. L. Wolf and A. Shashua. Direct feature selection with implicit inference. *ICCV'03*.
20. E.P. Xing, A.Y. Ng, M.I. Jordan, and S. Russel. Distance metric learning, with applications to clustering with side information. In *NIPS 2002*.