Principal Component Analysis Over Continuous Subspaces and Intersection of Half-spaces*

Anat Levin and Amnon Shashua

Computer Science Department,
Stanford University,
Stanford, CA 94305

Abstract. Principal Component Analysis (PCA) is one of the most popular techniques for dimensionality reduction of multivariate data points with application areas covering many branches of science. However, conventional PCA handles the multivariate data in a discrete manner only, i.e., the covariance matrix represents only sample data points rather than higher-order data representations.
In this paper we extend conventional PCA by proposing techniques for constructing the covariance matrix of uniformly sampled continuous regions in parameter space. These regions include polytops defined by convex combinations of sample data, and polyhedral regions defined by intersection of half spaces. The applications of these ideas in practice are simple and shown to be very effective in providing much superior generalization properties than conventional PCA for appearance-based recognition applications.

1 Introduction

Principal Component Analysis (PCA)[12], also known as Karhunen-Loeve transform, has proven to be an exceedingly useful tool for dimensionality reduction of multivariate data with many application areas in image analysis, pattern recognition and appearance-based visual recognition, data compression, time series prediction, and analysis of biological data — to mention a few.

The typical definition of PCA calls for a given set of vectors $\mathbf{a}_1, ..., \mathbf{a}_k$ in an n-dimensional space, with zero mean, arranged as the columns of an $n \times k$ matrix A. The output set of principal vectors $\mathbf{u}_1, ..., \boldsymbol{u}_q$ are an orthonormal set of vectors representing the eigenvectors of the sample covariance matrix $AA^\top$ associated with the $q < n$ largest eigenvalues. The matrix $UU^\top$ is a projection onto the principal components space with the property that (i) the projection of the original sample is "faithful" in a least-square sense, i.e.,

$$\min \sum_{i=1}^{k} \mid \mathbf{a}_i - UU^\top \mathbf{a}_i \mid^2,$$

* This work was done while A.S. was a visiting faculty at Stanford University. The permanent address of the authors is at the Hebrew University of Jerusalem, Israel.

(ii) equivalently, that the projection of the sample set onto the lower dimensional space maximally retains the variance, i.e., the first principal vector $\mathbf{u}_1$ maximizes $\sum_i \mid A^\top \mathbf{u}_1 \mid^2$, and so forth. The representation of a sample point $\mathbf{a}_i$ in the lower dimensional feature space is defined by $\mathbf{x}_i = U^\top \mathbf{a}_i$, and (iii) the covariance matrix $Q = \sum_i \mathbf{x}_i \mathbf{x}_i^\top$ of the reduced dimension representation is diagonal, i.e., PCA *decorrelates* the sample data (thus, if the sample data is drawn from a normal distribution then the variables in the feature space are statistically independent).

The strength of PCA for data analysis comes from its efficient computational mechanism, the fact that it is well understood, and from its general applicability. For example, a sample of applications in computer vision includes the representation and recognition of faces [25, 26, 16, 3], recognition of 3D objects under varying pose [17], tracking of deformable objects [6] and for representations of 3D range data of heads [1].

Over the years there have been many extensions to conventional PCA. For example, Independent Component Analysis (ICA) [8, 5] is the attempt to extend PCA to go beyond decorrelation and to perform a dimension reduction onto a feature space with statistically independent variables. Other extensions address the situation where the sample data live in a low-dimensional (non-linear) manifold in an effort to retain a greater proportion of the variance using fewer components (cf. [7, 11, 10, 13, 27, 21]); and yet other (related) extensions derive PCA from the perspective of density estimation (which facilitate modeling non-linearities in the sample data) and the use of Bayesian formulation for modeling the complexity of the sample data manifold [28].

In this paper we propose a different kind of extension to conventional PCA, which is orthogonal to the extensions proposed in the past, i.e., what we propose could be easily retrofitted to most of the PCA-extensions. The extension we propose has to do with the *representation* of the sample data. Rather than the data consist of *points* alone we allow for representation of continuous *regions* described by (i) convex combinations of sample points (polytops), and (ii) convex regions defined by intersection of half-spaces. In other words, we show how to construct the covariance matrix of a uniformly sampled polytop described by a finite set of sampled points (the generators of the polytop), and a uniformly sampled polyhedral defined by intersection of half-spaces. In the former case, the integration over the polytop region boils down to a very simple modification of the original covariance matrix of the sampled point set: replace $AA^\top$ with $A\Phi A^\top$ where Φ is a symmetric positive definite matrix which is described analytically per region. We show that despite the simplicity of the result the approach has a significant effect on the generalization properties of PCA in appearance-based recognition — especially when the raw data is not uniformly sampled. In the case of polyhedral regions described by intersections of half-spaces we show that although the concept of integration over the bounded region is not obvious it can be done at a cost of $O(n^3)$ in certain cases where the half-spaces are defined by inequalities over pairs of variables forming a tree structure. We demonstrate the application of this concept to intensity-ratio representations in appearance-based recognition and show a much superior generalization over conventional PCA.

1.1 Region-based PCA — Definition and Motivation

We will start with a simple example. Given two arbitrary points $\mathbf{a}_1, \mathbf{a}_2$ in the two-dimensional plane, the first principal component $\mathbf{u}$ (a unit vector) maximizes the scatter: $(\mathbf{a}_1^\top \mathbf{u})^2 + (\mathbf{a}_2^\top \mathbf{u})^2$, which is the first eigenvector (associated with the largest eigenvalue) of the 2×2 matrix $\mathbf{a}_1 \mathbf{a}_1^\top + \mathbf{a}_2 \mathbf{a}_2^\top$. Consider the case where we would like the entire line segment $\lambda \mathbf{a}_1 + (1-\lambda)\mathbf{a}_2$, $0 \leq \lambda \leq 1$ to be sampled, how would that change the direction of the principal component $\mathbf{u}$? In other words, if W is a polytop defined by the convex combination of a set of points (in this case W is a line segment), one is looking for the evaluation of the integral:

$$\max_{|u|=1} \int_{\mathbf{a} \in W} \mid \mathbf{a}^\top \mathbf{u} \mid^2 d\mathbf{a} = \mathbf{u}^\top \left[\int_{\mathbf{a} \in W} \mathbf{a}\mathbf{a}^\top d\mathbf{a} \right] \mathbf{u} \tag{1}$$

By substituting $\lambda \mathbf{a}_1 + (1-\lambda)\mathbf{a}_2$ for $\mathbf{a}$ in the integral $\int \mathbf{a}\mathbf{a}^\top$ and noting that

$$\int_0^1 \lambda^2 d\lambda = \int_0^1 (1-\lambda)^2 d\lambda = \frac{1}{3}, \qquad \int_0^1 \lambda(1-\lambda) d\lambda = \frac{1}{6},$$

we obtain the optimization problem:

$$\max_{|u|=1} \mathbf{u}^\top [\mathbf{a}_1 \mathbf{a}_1^\top + \mathbf{a}_2 \mathbf{a}_2^\top + \frac{1}{2}(\mathbf{a}_1 \mathbf{a}_2^\top + \mathbf{a}_2 \mathbf{a}_1^\top)] \mathbf{u}$$

Therefore, the first principal component $\mathbf{u}$ is the largest[1] eigenvector of the matrix:

$$A \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} A^\top = A \Phi A^\top \tag{2}$$

where $A = [\mathbf{a}_1, \mathbf{a}_2]$ the matrix whose columns consists of the sample points. In the following section we will generalize this simple example and handle any polytop (represented by its vertices). Once we have that at hand it is simple to accommodate a collection of polytops or a combination of points and polytops as input to the PCA process.

The motivation for representing ploytopes in a PCA framework arises from the fact that in many instances in visual recognition it is known apriori that the data resides in polytopes: probably the most well known example is the case of varying illumination over Lambertian surfaces. There are both empirical and analytic justifications to the fact that a relatively small number of images are necessary to model the image variations of human faces under different lighting conditions. In this context, a number of researchers have raised the issue of how to optimally construct the subspace using a sample of images which may be biased. Integration over the polyhedral defined by a sample, even though it

[1] We mean by "largest" the eigenvector associated with the largest eigenvalue.

is a biased sample, would be a way to construct the image subspace. This is addressed in Section 2.1 of this paper.

In the second part of the paper we consider polyhedrals defined by the intersection of half-spaces. Let the variables of a data point be denoted by $x_1, ..., x_n$ where the range of each variable is finite (say, denote pixel values) and consider the polyhedral defined by the relations:

$$\alpha_{ij} x_j < x_i < \beta_{ij} x_j$$

for a number of pairs of variables. Each inequality defines a pair of half-spaces (area on side of a hyperplane) thus a (sufficient and consistent) collection of inequalities will define a polyhedral cone whose apex is at the origin. As before, we would like to represent the entire bounded region in the PCA analysis — and we will show how this could be done in the sequel.

Our motivation for considering regions defined by inequalities comes from studies in visual recognition showing that the ratio alone between pre-selected image patches provides a very effective mechanism for matching under variability of illumination. For example, [24, 18] propose a graph representation (see Fig. 3b for an example) where regions in the image correspond to nodes in the graph and the edges connect neighboring regions which have a consistent relation "the average image intensity of node i is between 0.35 and 0.75 of node j", for instance. A match between a test image and the model reduces to a graph matching problem. In this paper we show that the idea of a graph representation could be embedded into the PCA framework by looking for the principal components which best describe the region in space bounded by the hyperplanes associated with those inequalities. In other words, it is possible to recast data analysis problems defined by inequalities within a PCA framework — whatever the application may be.

2 PCA over Polytops Defined by Convex Combinations

Let W denote a polytop defined by the convex combinations of a set of d points $\mathbf{a}_1, ..., \mathbf{a}_d$, such that each subset of size $d-1$ is linearly independent. Let $\mathcal{D}_d$ be the $d-1$ dimensional manifold

$$\mathcal{D}_d = \{\mu = (\mu_1, ..., \mu_d\} \in R^d \mid \sum_i \mu_i = 1, \quad \mu_i \geq 0\},$$

and let $V(\mathcal{D}_d) = \int_{\mathcal{D}_d} 1 d\mu$ be the volume of $\mathcal{D}_d$. The principal components are the eigenvectors of the covariance matrix:

$$Cov(W) = \frac{1}{V(W)} \int_{\mathbf{a} \in W} \mathbf{a}\mathbf{a}^\top d\mathbf{a},$$

where $V(W)$ denotes the volume of the polytop W, and the inverse volume outside the integral indicates that we have assumed a uniform density function when sampling the points $\mathbf{a} \in W$. Let $A = [\mathbf{a}_1, ..., \mathbf{a}_d]$ be the $n \times d$ matrix whose columns are the generators of W, then for every vector $\mu \in \mathcal{D}_d$ we have that $A\mu \in W$. Therefore, the covariance matrix representing the dense (uniform) sampling of the polytop W takes the form:

$$Cov(W) = \frac{1}{V(\mathcal{D}_d)} A \left[\int_{\mu \in \mathcal{D}_d} \mu \mu^\top d\mu \right] A^\top = A \Phi_d A^\top. \tag{3}$$

Note that the matrix $\Phi_d = (1/V(\mathcal{D}_d)) \int \mu\mu^\top d\mu$ does not depend on the choice of the generators $\mathbf{a}_1, ..., \mathbf{a}_d$, thus the integral needs to be evaluated only once for every choice of d.

Note that since $\mathcal{D}_d$ is symmetric under the group of permutations of d letters, then $\int \mu_i \mu_j d\mu = \int \mu_1 \mu_2 d\mu$ for all $i \neq j$ and $\int \mu_i^2 d\mu = \int \mu_1^2 d\mu$. Thus, the matrix Φ_d has the form:

$$\Phi_d = \frac{1}{V(\mathcal{D}_d)} \begin{bmatrix} \alpha_d & \beta_d & \cdots & \beta_d \\ \beta_d & \alpha_d & \cdots & \beta_d \\ \cdot & \cdot & \cdots\cdot & \cdot \\ \cdot & \cdot & \cdots\cdot & \cdot \\ \beta_d & \beta_d & \cdots & \alpha_d \end{bmatrix},$$

where $\alpha_d = \int \mu_1^2 d\mu$ and $\beta_d = \int \mu_1 \mu_2 d\mu$. We are therefore left with the task of evaluating three integrals α_d, β_d and $V(\mathcal{D}_d)$. By induction on d one can evaluate those integrals (derivation is omitted due to lack of space) and obtain:

$$V(\mathcal{D}_d) = \int_{\mathcal{D}_d} 1 d\mu = \frac{1}{(d-1)!}, \quad \alpha_d = \int_{\mathcal{D}_d} \mu_1^2 d\mu = \frac{2}{(d+1)!}, \quad \beta_d = \int_{\mathcal{D}_d} \mu_1 \mu_2 d\mu = \frac{1}{(d+1)!}$$

We therefore obtain the following result:

Theorem 1. *The covariance matrix of the uniformly sampled polytop W defined by the convex linear combinations of a set of d points $\mathbf{a}_1, ..., \mathbf{a}_d$ (such that each subset of size $d-1$ is linearly independent), arranged as the columns of a $n \times d$ matrix A, has the form:*

$$Cov(W) = \frac{1}{V(W)} \int_{\mathbf{a} \in W} \mathbf{a}\mathbf{a}^\top d\mathbf{a} = \frac{1}{d(d+1)} A(I + \mathbf{e}\mathbf{e}^\top) A^\top, \tag{4}$$

where $\mathbf{e} = (1, 1, ..., 1)$ and "I" is the identity matrix.

There are few points worth noting. First, when the data consists of a single polytop, then centering the data, i.e., subtracting the mean such that $Ae = 0$, then the discrete covariance matrix $AA^\top$ over the vertices of W is the same the covariance matrix $cov(W)$ of the entire polytop. Therefore, the integration makes a difference only when there are a number of polytops.

Second, the matrix $\Phi_d = I + \mathbf{e}\mathbf{e}^\top$ can be factored as $Q_d Q_d^\top$:

$$I + \mathbf{e}\mathbf{e}^\top = (I + c\mathbf{e}\mathbf{e}^\top)(I + c\mathbf{e}\mathbf{e}^\top)^\top = Q_d Q_d^\top,$$

where $c = \frac{\sqrt{d+1}-1}{d}$. This property can be used to perform PCA on a $d \times d$ matrix instead of the $n \times n$ covariance matrix when $d << n$, as follows. Denote

$\hat{A} = AQ_d$ (an $n \times d$ matrix), thus $\hat{A}\hat{A}^\top = A\Phi_d A^\top$. In case $d << n$ then let $\mathbf{y}$ be an eigenvector of $\hat{A}^\top \hat{A}$ ($d \times d$ matrix), then $\hat{A}\mathbf{y}$ is the corresponding eigenvector of $\hat{A}\hat{A}^\top$. The importance of this property is that the computational complexity of recovering the principal vectors is proportional to the dimension of the polytop rather than the dimension n of the vector space.

The third point to note is that the covariance matrix of two uniformly sampled polytops of dimensions $d_1 - 1$ and $d_2 - 1$ is the sum of the covariance matrices corresponding to each polytop separately. In other words, let $\mathbf{a}_1, ..., \mathbf{a}_{d_1}$, arranged as columns of a matrix A, be the generators of the first polytop, and let $\mathbf{b}_1, ..., \mathbf{b}_{d_2}$, arranged as columns of a matrix B, be the generators of the second polytop. The covariance matrix of the data covering the uniform sampling of both polytops is simply $A\Phi_{d_1} A^\top + B\Phi_{d_2} B^\top$. Thus, for example, given a collection of triplets of images of a class of 3D objects (say, frontally viewed human faces) where the i'th triplet, represented by a $n \times 3$ matrix A_i, spans the illumination cone of the i'th object, then the covariance matrix of the entire class is simply $\sum_i A_i \Phi_3 A_i^\top$. In case the number of triplets $k << n$ the cost of computing the principal vectors is proportional to $3k$ rather than to n by noting that one can compute the eigenvectors of the $3k \times 3k$ matrix $B^\top B$ where $B = [A_1 Q_3, ..., A_k Q_3]$ is an $n \times 3k$ matrix.

2.1 Demonstrating the Strength of Results

In this section we will provide one example to illustrate the strength of our result. The simplicity of our result, replace AA^T by $A\Phi A^\top$, may be somewhat misleading — the procedure is indeed trivial, but its effects could be very significant as we show next.

As mentioned in the introduction, empirical and analytic observations on the class of human faces have shown that a relatively small number of sample images of a Lambertian object are sufficient to model the image space of the object under varying lighting conditions. Early work showed that when no surface point is shadowed, as little as three images suffice [23, 22]. Empirical results have shown that even with cast and attached shadows, the set of images is still well approximated by a low dimensional space [9]. Later work has shown that the set of all images form a polyhedral cone which is well approximated (for human faces) by a low dimensional linear subspace [4], and more recently that the illumination cone (for convex Lambertian objects) can be represented by a 9-dimensional linear subspace [2, 19]. In this context, researchers [29, 14] have also addressed the issue of how to construct the illumination cone from sample images, i.e., what would be the best representative sample?. A biased set of samples would produce a PCA space which is not effective for recognition. Closest to our work, [29] proposed a technique for integration over any three images employing a spherical parameterization of the 3D space, which in turn is specific to 3-dimensional subspaces.

Therefore, the problem of "relighting" provides a good testing grounds for the integration over polytops idea. The integration can turn a biased sample into non-biased — and this is exactly the nature of the experiment below.

Consider a training set of images of human frontal faces covering different people and covering various illumination conditions (direction of light sources). One would like to represent the training set by a small number of principal components [26, 9]. The fact that the image space of a 3D object with matte (Lambertian) surface properties is known to occupy a small dimensional subspace suggests that the collection of training images per person forms a polytop which will be uniformly sampled when creating the covariance matrix of the entire training set. In other words, the training set would consist of a collection of polytops — one per person; where each polytop is defined by the convex combinations of the set of images of that person.

In order to appreciate the difference between representing the polytops versus representing the sample data points alone, we constructed a *biased* set of images with respect to the illumination. We used the training set provided by Yale University's "illumination dome" where for each of the 38 objects we sampled 42 images: 40 of them illuminated from light sources in the left hemisphere and only 2 from light sources located at the right hemisphere.

Since each of the faces is represented by a (biased) sample of 42 images, whereas the polytop we wish to construct is only 3-dimensional we do the following. For each PCA-plane corresponding to the sample of a face, we reparameterize the space such that all the 42 images are represented with respect to their first 3 principal components, i.e., each image is represented by a 3-coordinate vector. Those vectors are normalized thus they reside on a sphere. The normalized 2D coordinates then undergo a triangulation procedure. Let A_i be the $n \times 3$ matrix representing the i'th triangle, then $A_i \Phi_3 A_i^\top$ represents to covariance matrix of the uniformed sampled triangle. As a result, $\sum_i A_i \Phi_3 A_i^\top$ represents the covariance matrix of the continuous space defined by the 42 sample images of the face. This is done for each of the 38 faces and the final covariance matrix is the sum of all the individual covariance matrices.

Fig. 1a.1-6 shows a sample of images of a person in the training set. In row c.1-3 we show the first three principal vectors when the covariance matrix is constructed in the conventional way (i.e., AA^T where the columns of A contain the entire training set). Note that the first principal vector (which typically represents the average data) has a strong shadow on the right hand side of the face. In row c.4-6, on the other hand, we show the corresponding principal vectors when the covariance matrix is constructed by summing up the individual covariance matrices one for each polytop (as described above). Note that the principal vectors represent an *un-baised* illumination coverage. The effect of representing the polytops is very noticeable when we look at the projection of novel images onto the principal vector set. In row b.1-4 we consider four novel images, two images per person. The projections of the novel images onto the subspace spanned by the first 40 principal vectors are shown in row b.5-8 for conventional PCA and in row b.9-12 when polytops are represented. The difference is especially striking when the illumination of the novel image is to the right (where the original training sample was very small). One can clearly see that the region-based PCA has a much superior generalization property than the conventional approach — de-

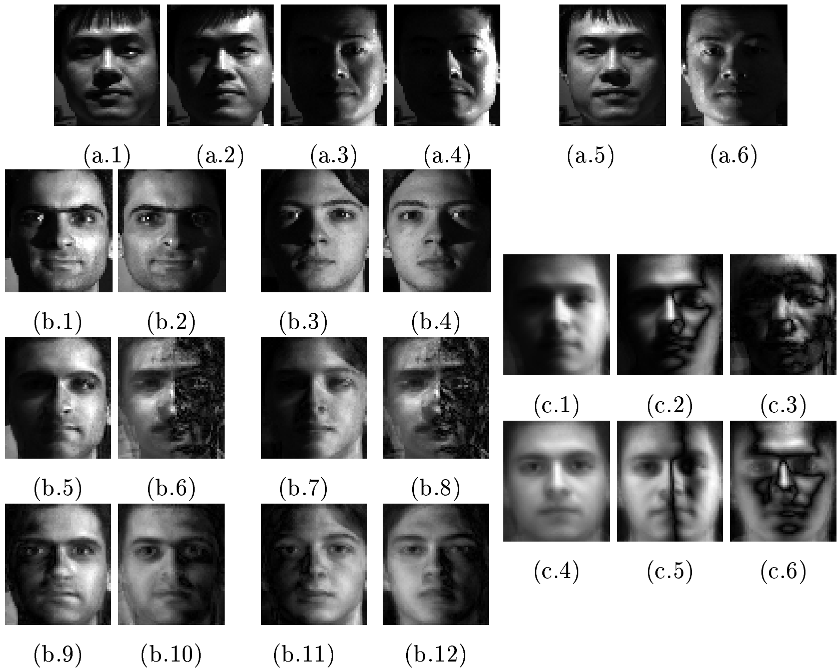

Fig. 1. Representing polytops constructed by images of different people, each person sampled by a biased collection of light source directions. (a.1-6) a sample of training images, (c.1-3) the first three principal vectors of the raw data — notice that the light source is biased compared to the principal vectors in (c.4-6) computed from the polytop representations. (b.1-4) novel images of two persons (test images), and their projections on the principal space: (b.5-8) when computed from the raw data, and (b.9-12) when computed from the polytop representation. Note that the latter has much superior generalization performance.

spite the fact that a very simple modification was performed to the conventional construction of the covariance matrix.

3 PCA over Polyhedrals Defined by Inequalities

In this section we turn our attention to polyhedrals defined by intersection of half spaces. Specifically, we will focus on the family of polyhedrals defined by the inequalities:

$$\alpha_{ij} x_j \leq x_i \leq \beta_{ij} x_j, \tag{5}$$

where $x_1, ..., x_n$ are the coordinates of the vector space representing the input parameter space, and α_{ij}, β_{ij} are scalars. Each such inequality between a pair of variables x_i, x_j represents a pair of hyperplanes, passing through the origin, bounding the possible data points in the desired (convex) polyhedral. These type of inequalities arise in appearance-based visual recognition where only *ratios* among certain key regions are considered when matching an image to a class of objects. Fig. 3b illustrates a sketch of a face image with the key regions marked and the pairs of regions for which the ratio of the average grey value is being considered for matching. In the following section we will discuss in more detail this type of application.

In order to make the relation between a set of inequalities, as defined above, and polyhedrals bounding a finite volume in parameter space, it would be useful to consider the following graph structure. Let the graph $G(V, E)$ be defined such that the set of vertices $V = \{x_1, ..., x_n\}$ represent the coordinates of the parameter space, and for each inequality relation between x_i, x_j there is a corresponding edge $e_{ij} \in E$ coincident with vertices x_i, x_j in the graph. Tree structures (connected graph with $n-1$ edges) are of particular interest because they correspond to a convex polyhedral:

Claim. Let the associated graph of the set of inequalities of the type $\alpha_{ij} x_j \le x_i \le \beta_{ij} x_j$ form a tree. Then, given that an arbitrary variable x_1 is bounded $0 \le x_1 \le 1$, then all other variables are defined in a finite interval, and as a result the collection of resulting hyperplanes bound a finite volume in space.

Proof: Due to the connectivity and lack of cycles in the graph, one can chain the inequality relations along paths of the graph leading to x_1 and obtain the set of new inequalities of the form $\delta_j x_1 \le x_j \le \gamma_j x_1$ for some scalars δ_j, γ_j. Therefore, the hyperplane $x_1 = constant$, which does not pass through the origin, bounds a finite volume because the range of all other variables is finite. $\square$

At this point, till the end of this section, we will assume that the associated graph representing the set of input inequalities is a connected tree (i.e., has n-1 edges and has no cycles). Our goal is to compute the integral:

$$\int_{\mathbf{x} \in W} \mathbf{x}\mathbf{x}^\top dx_1 \cdots dx_n,$$

where W is the region bounded by the hyperplanes corresponding to the inequalities and the additional hyperplane corresponding to $x_o = 1$ where x_o is one arbitrarily chosen variable (we will discuss how to choose x_o later in the implementation section). Since the entries of the matrix $\mathbf{x}\mathbf{x}^\top$ are bilinear products of the variables, we need to find a way of evaluating the integral on monomials $x_1^{\mu_1} \cdots x_n^{\mu_n}$ where μ_i are non-negative natural numbers. For a single constraint $\alpha_{ij} x_j \le x_i \le \beta_{ij} x_j$ the integration over dx_i is straightforward:

$$\int_{\alpha_{ij} x_j}^{\beta_{ij} x_j} x_1^{\mu_1} \cdots x_n^{\mu_n} dx_i = \frac{1}{\mu_i + 1} (\beta_{ij}^{\mu_i + 1} - \alpha_{ij}^{\mu_i + 1}) x_j^{\mu_i + \mu_j + 1} \prod_{k \ne i,j} x_k^{\mu_k} \quad (6)$$

For multiple constraints our challenge is to perform the integration without breaking W into sub-regions. For example, consider the two inequalities below:

$$\alpha_{ij} x_j \le x_i \le \beta_{ij} x_j$$
$$\alpha_{ik} x_k \le x_i \le \beta_{ik} x_k$$

Then, the integration over the variable x_i (which is bounded both by x_j and by x_k) takes the form:

$$\int_{\max\{a_{ij}x_j, a_{ik}x_k\}}^{\min\{b_{ij}x_j, b_{ik}x_k\}} x_1^{\mu_1} \cdots x_n^{\mu_n} dx_i$$

which requires breaking up the region W into 4 pieces. Alternatively, by noticing that $\alpha_{ij}x_j \le x_i \le \beta_{ij}x_j$ is equivalent to $\frac{1}{\beta_{ij}}x_i \le x_j \le \frac{1}{\alpha_{ij}}x_i$ the integration would take the form:

$$\int_{\alpha_{ik}x_k}^{\beta_{ik}x_k} \int_{\frac{1}{\beta_{ij}}x_i}^{\frac{1}{\alpha_{ij}}x_i} x_1^{\mu_1} \cdots x_n^{\mu_n} dx_j dx_i.$$

Therefore, in order to simplify the complexity of the integration process one must permute the variables $i \to \pi(i)$ and switch the variables inside the inequalities such that after the re-ordering we have the following condition: for every i, there exist at most a single constraint $\alpha_{i\rho(i)}x_{\rho(i)} \le x_i \le \beta_{i\rho(i)}x_{\rho(i)}$ where $\pi(\rho(i)) > \pi(i)$, i.e., the integration over x_i is performed before the integration over $x_{\rho(i)}$. In this case the integration over the region W takes the form:

$$\int_0^1 \int_{\alpha_{\pi_{n-1},\rho(\pi_{n-1})}}^{\beta_{\pi_{n-1},\rho(\pi_{n-1})}} \cdots \int_{\alpha_{\pi_1,\rho(\pi_1)}}^{\beta_{\pi_1,\rho(\pi_1)}} \mathbf{x}\mathbf{x}^\top dx_{\pi_1} \cdots dx_{\pi_n} \tag{7}$$

where π_i stands for $\pi(i)$. Before we explain how this could be achieved via the associated graph, consider the following example for clarification. Let $n = 4$ and we are given the following inequalities:

$$x_2 \le x_1 \le 2x_2$$
$$x_3 \le x_1 \le 3x_3$$
$$x_1 \le x_4 \le 2x_1$$

Since x_1 is bounded twice, we replace the first inequality with its equivalent form:

$$\frac{1}{2}x_1 \le x_2 \le x_1.$$

We therefore have $\rho(1) = 3, \rho(2) = 1$ and $\rho(4) = 1$. Select the permutation (143), i.e., $\pi(1) = 4, \pi(2) = 2, \pi(3) = 1$ and $\pi(4) = 3$. The integration of the monomial x_3^2 (for instance) over the bounded region W is therefore:

$$\int_0^1 \int_{x_3}^{3x_3} \int_{\frac{1}{2}x_1}^{x_1} \int_{x_1}^{2x_1} x_3^2 dx_4 dx_2 dx_1 dx_3.$$

The integration over x_4 is performed first: $\int_{x_1}^{2x_1} x_3^2 dx_4 = x_1 x_3^2$ (according to eqn. 6), then the integration over x_2 is performed: $\int_{0.5x_1}^{x_1} x_1 x_3^2 dx_2 = \frac{1}{2}x_1^2 x_3^2$, followed by the integration over x_1, $\int_{x_3}^{3x_3} \frac{1}{2}x_1^2 x_3^2 dx_1 = \frac{14}{3}x_3^5$ and finally the integration over x_3 (the free variable), $\int_0^1 \frac{14}{3}x_3^5 dx_3 = \frac{7}{9}$.

The decision of which inequality to "turn around" and how to select the order of integration (the permutation $\pi(i)$) can be made through simple graph algorithms, as follows. We will assign directions to the edges of the graph G with the convention that a directed edge $x_i \rightarrow x_j$ represents the inequality $\alpha_{ij} x_j \leq x_i \leq \beta_{ij} x_j$. The condition that for every i there should exist at most a single inequality $\alpha_{ij} x_j \leq x_i \leq \beta_{ij} x_j$ is equivalent to the condition that the associated directed graph will have at most one outgoing edge for every node. The algorithm for directing the edges of the undirected graph would start from some degree-1 node (a node with a single incident edge) and trace a path until a degree-1 node is reached again. The direction of edges would then follow the path. The process repeats itself with a new degree-1 node until no new nodes remain. Since G is a tree this process is well defined.

The selection of the order of integration is then simply obtained by a *topological sort* procedure. The reason for that is that one can view every pair $x_i \rightarrow x_j$ as a partial ordering (x_i comes before x_j). The topological sort provides a complete ordering (which is not necessarily unique) which satisfies the partial orderings. The complete order is the desired permutation. The example above is displayed graphically in Fig. 3a where the directed 4-node graph is shown and the topological sort result x_4, x_2, x_1, x_3 (note that x_2, x_4, x_1, x_3 is also a complete ordering which yields the same integration result).

To summarize, given a set of $n-1$ inequalities that form a connected tree, the covariance matrix of the resulting polyhedral is computed as follows.

1. Direct the edges of the associated graph so that there would be at most a single outgoing edge from each node.
2. "turn around" inequalities which do not conform to the edge direction convention.
3. Perform a topological sort on the resulting directed tree.
4. Evaluate the integral in eqn. 7 where the complete ordering from the topological sort is substituted for the permutation $\pi(i)$.

The complexity of this procedure is $O(n)$ for every entry of the $n \times n$ matrix $\mathbf{x}\mathbf{x}^\top$.

3.1 Experimental Details

In this section we illustrate the application of principal vectors defined by a set of inequalities in the domain of representing a class of images by intensity ratios — an idea first introduced by [24, 18]. Consider a training set of human frontal faces, roughly aligned, where certain key regions have been identified. For example, [24] illustrates a manual selection of key regions and a manual determination of the inequalities on the average intensity of the key regions. The associated graph becomes the model of the class of objects and the matching against a novel image is reduced to a graph matching procedure.

In this section we will re-implement the intensity-ratio inequality approach, but instead of using a graph matching procedure we will apply a PCA representation on the resulting polyhedral defined by the associated tree. There are a number of advantages of doing so: for example, the PCA approach allows us

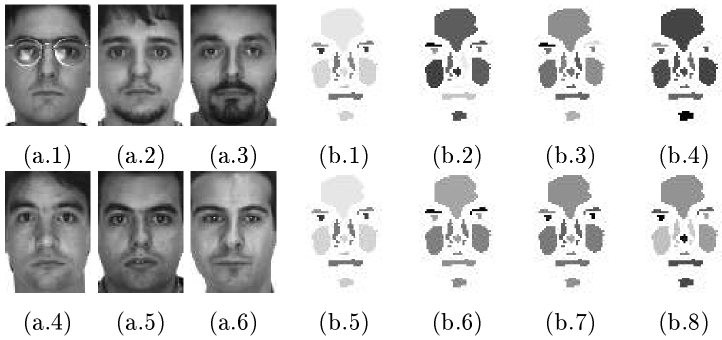

Fig. 2. (a.1-6) a sample of the training set from the AR dataset. (b.1-4) the first four principal vectors computed by integrating over the polyhedral region defined by the inequalities, and (b.5-8) are the principal vectors computed from the raw point data (in feature space).

to combine both raw data points, polytops defined by convex combinations of raw data points, and the polyhedrals defined by the inequalities. In other words, rather than viewing the intensity ratio approach as *the* engine for classification it could be just another cue integrated in the global covariance matrix. Second, by representing the polyhedral by its principal vectors one can make “soft” decisions based on the projection onto the reduced space, which is less natural to have in a graph matching approach.

As for training set, we used 100 images from the AR set [15] representing aligned frontal human faces (see Fig. 2a). The key regions were determined by applying a K-means clustering algorithm on the covariance matrix; five clusters were found and those were broken down based on connectivity to 13 key regions. The average intensity value was recorded per region thus creating the vector $\mathbf{x} = (x_1, ..., x_{13})$ as the feature representation of the original raw images. For every pair of variables x_i, x_j we recorded the sine of the angle between the vectors x_i recorded over the entire training set and the vector x_j over the training set — thus defining a complete associated graph with weights inversely proportional to the correlation between the pairs of variables. The *minimal spanning tree* of this graph was selected as the associated tree. Fig. 3b shows the key regions and the edges of the associated tree. Finally, for every pair of variables x_i, x_j which has an incident edge in the associated tree we determined the upper and lower bounds of the inequality by constructing the histogram of x_i/x_j and selected a_{ij} to be at the lower 30% point of the histogram and b_{ij} to be at the upper 30% of the histogram. This completes the data preparation phase for the region-based PCA applied to the polyhedral region defined by the associated tree.

Fig. 2b.1-4 shows the first four principal vectors of the region-based PCA using the integration formulas described in the previous section, while Fig. 2b.5-8 show the principal vectors using conventional PCA on the feature space vectors. One can see that the first principal vector (b.1 and b.5) are very similar, yet

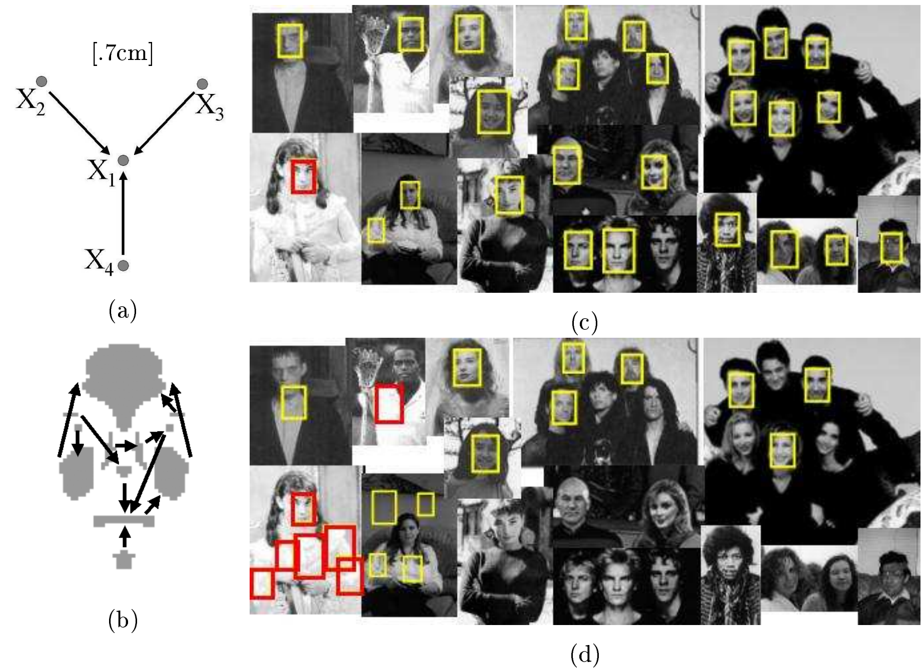

Fig. 3. (a) The associated tree of the $n = 4$ example. (b) A graphical description of the associated tree on the face detection experiment using inequalities. (c) Typical examples of true and false positives and negative detections on the leading technique (first row in table 4) (d) Typical examples of the worst technique (third row in table 4)

the remaining principal vectors are quite different. In table 4 we compare the performance over various uses of PCA on the CMU [20] test set of faces (which constitute postcards of people). The best technique was the product of the conventional PCA score on the raw image representation and the region-based PCA score. The results are displayed in the first row of the table. The false detections (false positives) are measured as a fraction of the total number of faces in the CMU test set. The miss-detections (false negatives) are measured as the percentage of the total number of true faces in the test set. Each column in the table represents a different tradeoff between the false positives and negatives — the better detection performance is at the expense of false positives. Thus, for example, when the detection rate was set to 96% (the highest possible in this technique) the false detection rate was 1.7 the amount of the total number of faces in the training set, whereas when the detection rate was set to 89% the false detection rate went down to 0.67 of the total number of faces. In the second row we use only conventional PCA: the score on the raw image representation multiplied with the score on the clustered image (feature vector of 13 dimensions).

The reduced performance is noticeable and significant. The worst performance is presented in the third row where only conventional PCA was used on the raw image representation. The region-based PCA performance is shown in the 4'th row: the performance is lower than the leading approach, but not much lower. And finally, conventional PCA on the clustered representation (13 dimensional feature vector) is shown in the 5'th row: note that the performance compared to the 4'th row is significantly reduced. Taken together, the region-PCA approach provides significant superiority in generalization properties compared to the conventional PCA - despite the fact that it is essentially a PCA approach. The fact that the relevant region of the parameter space is sampled correctly is the key factor behind the superior performance.

In Fig. 3c-d we show some typical examples of detections which contain true detections, false positives and negatives on the leading technique (first row in the table) and the worst technique (third row in table).

False detections	1.7	1.1	0.67
raw-PCA & region-PCA	96%	91%	89%
raw-PCA & PCA(13-dim)	80%	76%	75%
raw-PCA	60%	52%	54%
region-PCA	90%	86%	83%
conventional-PCA(13-dim)	79%	76%	72%

Fig. 4. Comparison of detection performance. The false detections (false positives) are measured as a fraction of the total number of faces in the CMU test set. The mis-detections (false negatives) are measured as the percentage of the total number of true faces in the test set. Each column in the table represents a different tradeoff between the false positives and negatives — the better detection performance is at the expense of false positives. The rows in the table represent the different techniques being used. See text for further details.

4 Summary

The paper makes a number of statements which include: (i) in some data analysis applications it becomes important to represent (uniform sampling of) continuous regions of the parameter space as part of the global covariance matrix of the data, (ii) in case where the continuous regions are polytops, defined by the convex combinations of sample data, the construction of the covariance matrix is extremely simple: replace the conventional $AA^\top$ covariance matrix with $A\Phi A^\top$ where Φ is described analytically in this paper, and (iii) the general idea extends to challenging regions such as those defined by intersections of half spaces — there we have derived the equations for constructing the covariance matrix where the regions are formed by $n-1$ inequalities on pairs of variables forming an associated tree structure.

The concepts laid down in this paper are not restricted to computer vision applications and have possibly a wider range of applications — just as the conventional PCA is widely applicable. In the computer vision domain we have shown that these concepts are effective in the domains of appearance-based visual recognition where continuous regions are defined by the illumination space (Section 2) — which are known to occupy low-dimensional subspaces — and in intensity-ratio representations. In the former case the regions form polytops and we have seen that the representation of those polytops make a big effect in the generalization properties of the principal vectors (Fig. 1), yet the price of applying the proposed approach is minimal. In the case of intensity-ratio representations, the notion of representing bounded spaces, defined by inequalities, by integration over the bounded region is not obvious, but is possible and at a low cost of $O(n^3)$. We have shown that the application of this concept provides much superior generalization properties compared to conventional PCA (Table 4).

Future work on these ideas include non-uniform sampling of regions in the case of polytops, handling the integration for general associated graphs (although in general the amount of work is exponential with the size and number of cycles in the graph) and exploring more applications for these basic concepts.

Acknowledgements

A.S. thanks Leo Guibas for hosting him during the academic year 2001/2. We thank Michael Elad and Gene Golub for insightful comments on the draft of this paper.

References

1. J.J. Atick, P.A. Griffin, and N.A. Redlich. Statistical approach to shape-from-shading: deriving 3d face surfaces from single 2d images. *Neural Computation*, 1997.
2. R. Basri and D. Jacobs. Photometric stereo with general, unknown lighting. In *iccv*, Vancouver, Canada, July 2001.
3. P.N Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. In *Proceedings of the European Conference on Computer Vision*, 1996.
4. P.N Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. In *Proceedings of the European Conference on Computer Vision*, 1996.
5. A.J. Bell and T.J. Sejnowski. An information maximization approach to blind separation and blind deconvolution. *Neural Computation 7(6)*, pages 1129–1159, 1995.
6. Michael J. Black and D. Jepson. Eigen-tracking: Robust matching and tracking of articulated objects using a view-based representation. In *Proceedings of the European Conference on Computer Vision* (ECCV), pages 329–342, Cambridge, England, 1996.
7. C. Bregler and S.M. Omohundro. Nonlinear manifold learning for visual speech recognition. In *iccv*, Boston, Jun 1995.

8. P. Comon. Independent component analysis, a new concept? *Signal processing 36(3)*, pages 11–20, 1994.
9. P. Hallinan. A low-dimentional representation of human faces for arbitrary lightening conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 995–999, 1994.
10. T. Hastie and W. Stuetzle. Principal curves. *Journal of Americam Statistical Association 84*, pages 502–516, 1989.
11. T. Heap and D. Hogg. Wormholes in shape space: Tracking through discontinuous changes in shape. In *iccv*, 1998.
12. I.T. Jolliffe. *Principal Component Analysis.* Springer-Verlag, New York, 1986.
13. M.A. Kramer. Non linear principal component analysis using autoassociative neural networks. *AI Journal 37(2)*, pages 233–243, 1991.
14. K.C. Lee, J. Ho and D. Kriegman. Nine points of light: Acquiring subspaces for face recognition under variable lighting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2001.
15. A.M. Martinez and A.C. Kak. PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(1):228–233, 2001.
16. B. Moghaddam A. Pentland and B. Starner. View-based and modular eigenspaoes for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), pages 84–91, 1994.
17. H. Murase and S.K. Nayar. Learning and recognition of 3D objects from appearance. In *IEEE 2nd Qualitative Vision Workshop*, pages 39–50, New York, NY, June 1993.
18. E. Grimson P. Lipson and P. Sinha. Configuration based scene classification and image-indexing. In *cvpr*, San Juan, Puerto Rico, 1997.
19. R. Ramamoorthi and P. Hanrahan. On the relationship between Radiance and Irradiance: Determining the illumination from images of a convex Lambertian object. In *Journal of the Optical Society of America* (JOSA A), Oct. 2001, pages 2448-2459.
20. H. Schneiderman and T. Kanade. A statistical model for 3d object detection applied to faces and cars. In *cvpr*, South Carolina, June 2000.
21. V. Silva, J.B. Tenenbaum and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290, December 2000.
22. A. Shashua. Illumination and view position in 3D visual recognition. In *Proceedings of the conference on Neural Information Processing Systems (NIPS)*, Denver, CO, December 1991.
23. A. Shashua. On photometric issues in 3D visual recognition from a single 2D image. *International Journal of Computer Vision*, 21:99–122, 1997.
24. P. Sinha. Object recognition via image invariances. *Investigative Ophthalmology and Visual Science 35/4:#1735*, 1994.
25. L. Sirovich and M. Kirby. Low dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America*, 4(3):519–524, 1987.
26. M. Turk and A. Pentland. Eigen faces for recognition. *J. of Cognitive Neuroscience*, 3(1), 1991.
27. A.R. Webb. An approach to nonlinear principal components-analysis using radially symmetrical kernel functions. *Statistics and computing 6(2)*, pages 159–168, 1996.
28. J.M. Winn C.M. Bishop. Non-linear bayesian image modelling. In *Proceedings of the European Conference on Computer Vision*, Dublin, Ireland, June 2000.
29. L. Zhao and Y.H. Yang. Theoretical analysis of illumination in PCA-based vision systems. Pattern Recognition, 32:547-564, 1999.