Electronic Journal of Linear Algebra ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 10, pp. 201-211, July 2003

ELA
http://math.technion.ac.il/iic/ela

OPTIMAL SUBSPACES AND CONSTRAINED PRINCIPAL COMPONENT ANALYSIS*

JEAN-DANIEL ROLLE†

Abstract. The main result of this article allows formulas of analytic geometry to be elegantly unified, by readily providing parametric as well as cartesian systems of equations. These systems characterize affine subspaces in $\mathbb{R}^p$ passing through specified affine subspaces of lower dimension. The problem solved is closely related to constrained principal component analysis. A few interesting applications are pointed out, notably a measure of the relative loss of optimality due to the constraints. The results pave the way for further research.

Key words. Optimal subspace problem, Constrained principal component analysis, Affine subspace.

AMS subject classifications. 15A03.

1. Introduction. The purpose of this paper is to provide results in analytical geometry that generalize and clarify the principle of constrained principal component analysis. Various forms of constrained principal component analysis have already been treated in the literature, notably in [8] and [9]. In [8], constraints are considered for both variables and individuals. These authors give several references to earlier applications of linear constraints in principal component analysis. The optimal subspace problem (see Theorem 3.1 below) is as follows: for a given $n \times p$ matrix Z, and a given d-dimensional subspace $\mathcal{D}$ of $\mathbb{R}^p$, find the subspace $\mathcal{H} = \mathcal{M}(H)$ that minimizes $\mathrm{tr}[(Z - ZP_H)(Z' - P_H Z')]$ subject to the condition that $\mathcal{H}$ contains the subspace $\mathcal{D}$.

Some secondary results follow from the specific way in which the main result is presented.

2. Tools and definitions. To present our results, we shall use the following notations: for a matrix B, B^+ will denote the Moore-Penrose inverse of B. Note that $B^+ = (B'B)^{-1}B'$ if B has full column rank and $B^+ = B'(BB')^{-1}$ if B has full row rank. We define $P_B = BB^+$ and $M_B = I - BB^+$, where I is the identity matrix of appropriate order. Thus P_B is the matrix of the orthogonal projection onto the column space of B, which we denote by $\mathcal{M}(B)$. We shall write $\mathrm{r}(B)$ for the rank of B, $\mathrm{tr}(B)$ for its trace, $[A|B]$ for the compound matrix formed by the blocks A and B horizontally displayed. All projections occurring in this article are orthogonal, so 'projection' will here always refer to 'orthogonal projection', with regard to the standard scalar product in $\mathbb{R}^p$. Finally, we shall use the notation $\mathcal{E} < \mathbb{R}^p$ for a linear subspace $\mathcal{E}$ of $\mathbb{R}^p$.

Lemma 2.1. *Define $G = [G_1|G_2]$. Then*
(a) $M_G M_{G_1} = M_{G_1} M_G = M_G = M_G M_{G_2} = M_{G_2} M_G$.

* Received by the editors on 11 April 2002. Accepted for publication on 23 June 2003. Handling editor: Ravindra B. Bapat.

†Haute Ecole de Gestion, University of Applied Sciences, Fribourg, Switzerland (Jean-Daniel.Rolle@eif.ch, rolle@freesurf.ch). Supported by a generous grant from the Haute Ecole de Gestion, Fribourg.

(b) $P_G P_{G_1} = P_{G_1} P_G = P_{G_1}$ *and* $P_G P_{G_2} = P_{G_2} P_G = P_{G_2}$.
(c) *If* $G_2 G_2^+ G_1 = 0$*, then* $M_{G_1} M_{G_2} = M_G$*, and* $\mathrm{r}(G) = \mathrm{r}(G_1) + \mathrm{r}(G_2)$.

Proof. The proof of this lemma appears in [7]. □

Definition 2.2. Let $\mathcal{E}$ be a linear subspace in $\mathbb{R}^p$, with $0 \leq s \leq p$. We call *affine subspace* of dimension s in $\mathbb{R}^p$ a subset $\mathcal{H} = \{c + y; y \in \mathcal{E}\}$ of $\mathbb{R}^p$, where c is a vector of $\mathbb{R}^p$. It is convenient to simply write $\mathcal{H} = c + \mathcal{E}$.

It should be clear that a linear subspace is also an affine subspace, and that the affine subspaces of dimension 0 are points, those of dimension 1 are lines, etc. We shall be interested later in (orthogonal) projection of vectors of $\mathbb{R}^p$ onto affine subspaces.

Let Z be a $n \times p$ matrix (e.g., a matrix of data in multivariate analysis). Denote by z_i' the rows of Z, so that $Z = \begin{bmatrix} z_1' \\ z_2' \\ \vdots \\ z_n' \end{bmatrix}$. The z_i, of dimension $p \times 1$, are therefore vectors of $\mathbb{R}^p$. We shall often refer to p-vectors as 'points' in p-space.

The solution of the optimal subspace problem will be given in parametric form. Lemma 2.3 will allow us to readily obtain the cartesian form of the solution.

Lemma 2.3. *Let v denote a $p \times q$ matrix of rank q, $1 \leq q \leq p-1$. Let w denote a $p \times (p-q)$ semi-orthogonal matrix (i.e., $w'w = I_{p-q}$). Moreover, assume that $w'v = 0$. Then*

$$\mathcal{M}(v) = \{x \in \mathbb{R}^p; w'x = 0\} \qquad (i.e., \mathcal{M}(v) = \ker(w')). \tag{2.1}$$

Proof. Lemma 2.3 can be proved using standard linear algebra results. □

Formally, even the case $q = 0$ and $q = p$ may be included. If $q = 0$, define v as the null vector of dimension $p \times 1$. If $q = 0$, w is an orthogonal matrix. If $q = p$, define w as the null vector of dimension $p \times 1$. Let $c \in \mathbb{R}^p$. As a corollary of Lemma 2.3, and with the same notations and assumptions, write $\mathcal{H} = c + \mathcal{M}(v)$ for a q-dimensional affine subspace of $\mathbb{R}^p$, $0 \leq q \leq p$. The set $\mathcal{H}$ passes through $c \in \mathbb{R}^p$, and has direction $\mathcal{M}(v)$. This is the *parametric* form of $\mathcal{H}$. Lemma 2.3 tells us that

$$\mathcal{H} = \{x \in \mathbb{R}^p; w'(x - c) = 0\}. \tag{2.2}$$

We call (2.2) the *cartesian* form of $\mathcal{H}$.

It is useful to recall what is meant by *inertia* and *dispersion.* To this aim, let Z denote a $n \times p$ matrix. Define

$$\mathcal{C} = \{z_i \in \mathbb{R}^p; z_i' = i^{\text{th}} \text{ row of } Z, i = 1, \ldots, n\}. \tag{2.3}$$

For a $p \times k$ matrix H and for $c \in \mathbb{R}^p$, let $\Delta = c + \mathcal{M}(H)$ denote an affine subspace of dimension $\mathrm{r}(H)$ in $\mathbb{R}^p$. Let p_H be the projection onto Δ (notice the lower-case letter used for p), that is

$$p_H(z) = c + P_H(z - c) \tag{2.4}$$

for z in $\mathbb{R}^p$.

DEFINITION 2.4. The *inertia* of $\mathcal{C}$ with regard to Δ is defined by

$$I(\mathcal{C},\Delta)=\frac{1}{n-1}\sum_{i=1}^{n}||z_i-p_H(z_i)||^2. \tag{2.5}$$

If we omit dividing by $n-1$ in (2.5), we speak of *dispersion* rather than of *inertia.* Note that when $\mathrm{r}(H)=0$, $I(\mathcal{C},\Delta)$ in (2.5) is the inertia of $\mathcal{C}$ with regard to a point. When $\mathrm{r}(H)=p$, p_H in (2.4) is the identity, i.e., $p_H(z)=z$ for all $z\in\mathbb{R}^p$.

3. The main results. We are now ready to give an important optimization result of geometrical nature, (Theorem 3.1), of which constrained principal component analysis is a direct consequence. Incidentally, note that a corollary of Theorem 3.1 (see the corollary below) allows the elegant solution of classical problems of analytical geometry, such as finding the parametric and cartesian form of the k-dimensional affine subspace of $\mathbb{R}^p$ passing through $k+1$ given points, or through specified lines or affine subspaces of dimension smaller than k.

3.1. Optimal subspace. As specified in the introduction, the object of Theorem 3.1 is to find the k-dimensional linear subspace $\mathcal{H}$ of $\mathbb{R}^p$ containing a given linear subspace $\mathcal{D}$ of lower dimension, where $\mathcal{H}$ is such that it is 'as close as possible' to a cloud $\mathcal{C}$ of points in p-space, i.e., where the inertia of $\mathcal{C}$ with regard to a k-dimensional linear subspace of $\mathbb{R}^p$ containing $\mathcal{D}$ is minimal for $\mathcal{H}$.

THEOREM 3.1. *Let $\mathcal{D}$ be a linear d-dimensional subspace of $\mathbb{R}^p$, $0\le d\le p$. If $d\ge 1$, let $D=[d_1|\dots|d_d]$ denote a $p\times d$ matrix of rank d (a full rank property that eases the presentation but that can be relaxed), whose columns form a basis of $\mathcal{D}$. If $d=0$, define D as the null vector of dimension $p\times 1$. Let Z be the $n\times p$ matrix $\begin{bmatrix} z_1' \\ \vdots \\ z_n' \end{bmatrix}$ and, for $M_D=I-P_D$, define the matrix $V_D=M_DZ'ZM_D$, assumed to be of rank r. Let s be a whole number such that $0\le s\le r$. If $s\ge 1$, let $\lambda_1\ge\ldots\ge\lambda_s$ denote the s largest eigenvalues of V_D (with $\lambda_s>\lambda_{s+1}$), and let $u_1,\ldots,u_s$ denote orthogonal eigenvectors corresponding to them. Write $U=[u_1|\dots|u_s]$ for the matrix collecting these vectors. If $s=0$, define U as the null vector of dimension $p\times 1$. Write $H=[U|D]$ for the matrix formed with the blocks U et D. Then the linear $(s+d)$-dimensional subspace $\mathcal{H}$ of $\mathbb{R}^p$ containing $\mathcal{D}$, and with respect to which the inertia of the cloud $\mathcal{C}$ is minimal, is given by $\mathcal{H}=\mathcal{M}(H)$. Moreover, $\mathcal{C}\subset\mathcal{H}$ if and only if $s=r$.*

The condition $\lambda_s>\lambda_{s+1}$ insures unicity of $\mathcal{H}$, and this even though some of the s eigenvalues are multiple. For potential applications in data analysis, the positive eigenvalues will generally be all distinct. This is notably the case (with probability one) when the z_i's are realizations of a continuous p-variate distribution.

Besides, notice that the columns of H are eigenvectors corresponding to the s largest positive eigenvalues, and to d zero eigenvalues of V_D. Since $V_DD=0$, at least d eigenvalues of V_D are zero. The columns of U (corresponding to positive eigenvalues) and the columns of D (corresponding to zero eigenvalues) are orthogonal, since V_D is symmetric, and since for symmetric matrices eigenvectors associated with distinct eigenvalues are orthogonal. Note also that the following properties (i) $\mathcal{C} \subset \mathcal{D}$, (ii) $V_D = 0$, and (iii) $r = 0$ are equivalent.

Let us give now the proof of Theorem 3.1.

Proof. Let $H = [U|D]$. Without loss of generality, we can assume that the columns of the $p \times s$ matrix U are orthonormal and $U'D = 0$. The problem is to find such a $p \times s$ matrix U such that U minimizes $\mathrm{tr}[(Z - ZP_H)(Z' - P_H Z')] = \mathrm{tr}[Z(I - P_H)Z']$, where Z is a given $n \times p$ matrix. Note that since $U'D = 0$, $P_H = P_D + P_U$ and $P_D P_U = 0$, i.e., $M_D P_U = P_U$. Using these properties, we get

$$\mathrm{tr}[Z(I - P_H)Z'] = \mathrm{tr}[Z(M_D - P_U)Z'] = \mathrm{tr}[M_D Z'Z] - \mathrm{tr}[P_U Z'Z]. \tag{3.1}$$

Thus we have to choose U to maximize $\mathrm{tr}[P_U Z'Z]$ subject to $M_D P_U = P_U$. However, $P_U = UU'$ (since $U'U = I_s$) and $M_D P_U = P_U$ is equivalent to $M_D U = U$. Hence

$$\mathrm{tr}[P_U Z'Z] = \mathrm{tr}[U'Z'ZU] = \mathrm{tr}[U'M_D Z'ZM_D U], \tag{3.2}$$

which is to be maximized subject to $U'U = I_s$. It is a standard linear algebra result that this maximum is the sum of the s largest eigenvalues of $M_D Z'ZM_D$, and the columns of the maximizing U are the corresponding eigenvectors. Note that since $s \leq r = \mathrm{r}(M_D Z'ZM_D)$, such a U will also satisfy $M_D U = U$.

To show that $\mathcal{C} \subset \mathcal{H}$ if and only if $s = r$, note that $\mathcal{C} \subset \mathcal{H}$ if and only if $Z' = P_H Z'$, or equivalently, $M_D Z' = P_U Z'$. Thus we have to show that $s = r$ if and only if $M_D Z' = P_U Z'$.

Note that when $r = s$, the r columns of U are the orthonormal eigenvectors corresponding to the r nonzero eigenvalues of V_D, where r is also the rank of V_D. Thus $s = r$ if and only if $\mathcal{M}(V_D) = \mathcal{M}(P_U)$, if and only if $\mathcal{M}(M_D Z') = \mathcal{M}(P_U)$, (since $\mathcal{M}(V_D) = \mathcal{M}(M_D Z')$), if and only if $M_D Z' = P_U M_D Z'$, if and only if $M_D Z' = P_U Z'$ (since $M_D P_U = P_U$).

We can conclude "$\mathcal{M}(M_D Z') = \mathcal{M}(P_U)$ if and only if $M_D Z' = P_U M_D Z'$" since we already know that $\mathcal{M}(P_U) \subset \mathcal{M}(M_D Z')$. □

Theorem 3.1 can be generalized to the affine subspaces. The object of the corollary thereafter is to find the k-dimensional affine subspace $\mathcal{H}$ of $\mathbb{R}^p$ containing a given affine subspace $\mathcal{D}$ of lower dimension, where $\mathcal{H}$ is such that it is 'as close as possible' to a cloud $\mathcal{C} = \{z_i \in \mathbb{R}^p; i = 1, \ldots, n\}$, i.e., where the inertia of $\mathcal{C}$ with regard to a k-dimensional affine subspace of $\mathbb{R}^p$ containing $\mathcal{D}$ is minimal.

Corollary 3.2. *Let $\mathcal{D}$ be a d-dimensional affine subspace of $\mathbb{R}^p$, $0 \leq d \leq p$, and let d_0 be any vector of $\mathcal{D}$. If $d \geq 1$, let $D = [d_1| \ldots |d_d]$ denote a $p \times d$ matrix of rank d whose columns form a basis for the linear subspace $\mathcal{D} - d_0$. If $d = 0$, define D as the null vector of dimension $p \times 1$. Let Z be the $n \times p$ matrix $Z = \begin{bmatrix} z_1' - d_0' \\ \vdots \\ z_n' - d_0' \end{bmatrix}$, and, for $M_D = I - P_D$, define the matrix $V_D = M_D Z'ZM_D$, assumed to be of rank r. Let s be a whole number such that $0 \leq s \leq r$. If $s \geq 1$, let $\lambda_1 \geq \ldots \geq \lambda_s$ denote the s largest eigenvalues of V_D (with $\lambda_s > \lambda_{s+1}$), and let $u_1, \ldots, u_s$ denote corresponding orthogonal eigenvectors. Write $U = [u_1| \ldots |u_s]$ for the matrix collecting these vectors.*

Electronic Journal of Linear Algebra ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 10, pp. 201-211, July 2003

ELA
http://math.technion.ac.il/iic/ela

If $s = 0$, define U as the null vector of dimension $p \times 1$. Finally, define $H = [U|D]$, the matrix formed by the blocks U and D. Then the $(s+d)$-dimensional affine subspace $\mathcal{H}$ of $\mathbb{R}^p$ containing $\mathcal{D}$ and with respect to which the inertia of the cloud $\mathcal{C}$ is minimal, is given by $\mathcal{H} = d_0 + \mathcal{M}(H)$. Moreover, $\mathcal{C} \subset \mathcal{H}$ if and only if $s = r$.

Proof. A simple line of arguments shows that the choice of d_0 in $\mathcal{D}$ does not matter. The rest of the proof is direct, if one follows the following three steps: (i) perform a rigid translation by removing d_0 from the z_i's and from $\mathcal{D}$, in such a way that the hypotheses of Theorem 3.1 are valid, (ii) apply Theorem 3.1 to find $\mathcal{M}(H)$, and (iii) perform the inverse translation by adding d_0 to $\mathcal{M}(H)$. □ The matrix M_D in the corollary has rank $p - d$. The eigenvalues of V_D, apart from d zeros, are $\lambda_1 \geq \ldots \geq \lambda_{p-d}$. Write $\eta_1 \geq \ldots \geq \eta_p$ for the eigenvalues of $Z'Z$, and $v_1, \ldots, v_p$, for corresponding eigenvectors. Then the Poincaré separation theorem (see [5]) implies $\lambda_j \leq \eta_j$, $j = 1, \ldots, p-d$. Moreover, it may be readily shown that $\lambda_j = \eta_j$, $j = 1, \ldots, p-d$ if $D = [v_{p-d+1}|v_{p-d+2}|\ldots|v_p]$.

As a special case, if we set $\mathcal{D} = \{\overline{z}\}$ in the corollary, we have $d = 0$, and the optimal s-dimensional affine subspace $\mathcal{H}$ containing $\mathcal{D}$ is generated by the first s principal axes of the (unconstrained) principal component analysis performed on the sample covariance matrix (in fact, we may replace $Z'Z$ in the corollary by the sample covariance matrix $C = Z'Z/(n-1)$ without changing the eigenvectors; only the eigenvalues would be divided by $n-1$).

The corollary tells us that, as the method completes $\mathcal{D}$ to provide $\mathcal{H}$, the optimal subspace is always obtained by "adding" another orthogonal eigenvector. That is, u_1 is orthogonal to the columns of the matrix D, u_2 is orthogonal to the columns of D and to u_1, and so on.

We now give the usual decomposition formula for dispersion, in our framework or space extension.

Decomposition of dispersion:

Lemma 3.3. *Let Z, $\mathcal{C}$, $\mathcal{D}$ and $\mathcal{H}$ be as in Theorem 3.1. Then one has the following identity (decomposition of dispersion):*

$$\sum_{i=1}^{n} ||z_i - P_D z_i||^2 = \sum_{i=1}^{n} ||P_H z_i - P_D P_H z_i||^2 + \sum_{i=1}^{n} ||z_i - P_H z_i||^2. \tag{3.3}$$

Proof. Use similar arguments as in the proof of Theorem 3.1. □ If we divide each term of (3.3) by $n-1$, we have the formula of decomposition of inertia. The left term of (3.3) is the dispersion of $\mathcal{C}$ with regard to $\mathcal{D}$. The second term on the right of (3.3) is the dispersion of $\mathcal{C}$ with regard to $\mathcal{H}$. Let $\mathcal{C}^*$ denote the cloud obtained by projecting $\mathcal{C}$ onto $\mathcal{H}$. The first term on the right of (3.3) is the dispersion of the cloud $\mathcal{C}^*$ with regard to $\mathcal{D}$; it measures the part of the inertia of $\mathcal{C}$ with regard to $\mathcal{D}$ due to the extension of $\mathcal{D}$ to $\mathcal{H}$. Note that the identity (3.3) is valid whatever linear subspace $\mathcal{H}$ containing $\mathcal{D}$ we take (i.e., $\mathcal{H}$ is not necessarily the optimal subspace of Theorem 3.1).

Let $\mathcal{C} = \{z_i \in \mathbb{R}^p; i = 1, \ldots, n\}$ denote a cloud of points and, for affine subspaces $\mathcal{D}$ and $\mathcal{H}$, $(\mathcal{D} \subset \mathcal{H})$, let p_D and p_H denote the orthogonal projectors onto $\mathcal{D}$ and

Electronic Journal of Linear Algebra ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 10, pp. 201-211, July 2003

ELA
http://math.technion.ac.il/iic/ela

$\mathcal{H}$, respectively (for the exact definition of p_H or p_D, see (2.4)). Then the following identity

$$\sum_{i=1}^{n} ||z_i - p_D(z_i)||^2 = \sum_{i=1}^{n} ||p_H(z_i) - p_D(p_H(z_i))||^2 + \sum_{i=1}^{n} ||z_i - p_H(z_i)||^2 \tag{3.4}$$

is a direct consequence of (3.3).

3.2. Application in analytic geometry. The corollary was not primarily intended for analytic geometry. However, it is general enough to be an interesting tool in this area. To illustrate its potential applications, suppose, for example, that we are interested in finding the parametric (or the cartesian) form of a r-dimensional affine subspace $\mathcal{H}$ $(1 \leq r \leq p-1)$ passing through $d_0 \in \mathbb{R}^p$ and through the set of points $\mathcal{C} = \{z_1, \ldots, z_r\} \subset \mathbb{R}^p$ (think of a line passing through two points (d_0 and z_1) in $\mathbb{R}^p$, or a plane passing through three points (d_0, z_1 and z_2) in $\mathbb{R}^p$). Let us assume that $\mathrm{r}(Z) = r$, where $Z = \begin{bmatrix} z_1' - d_0' \\ \vdots \\ z_r' - d_0' \end{bmatrix}$; that is, $d_0, z_1, \ldots, z_r$ generate indeed an r-dimensional affine subspace. To find $\mathcal{H}$, we set $\mathcal{D} = \{d_0\}$ (i.e., $V_D = Z'Z$) in the corollary. This implies that the affine subspace of dimension r passing (exactly) through the $r+1$ points $d_0, z_1, \ldots, z_r$ is

$$\mathcal{H} = d_0 + \mathcal{M}(U), \tag{3.5}$$

where the columns of the $p \times r$ matrix U are orthonormal eigenvectors of $Z'Z$ associated with the r positive (some of them possibly multiple) eigenvalues of this matrix. Let w be a $p \times (p-r)$ matrix, whose columns are orthonormal eigenvectors of $Z'Z$ associated with the zero eigenvalue (of multiplicity $p-r$) of this matrix. Since for a symmetric matrix eigenvectors associated with distinct eigenvalues are orthogonal, we have $w'U = 0$. Lemma 2.3 and (2.2) yield the cartesian form of $\mathcal{H}$:

$$\mathcal{H} = \{x \in \mathbb{R}^p; w'(x - d_0) = 0\}. \tag{3.6}$$

Let us illustrate this by two simple examples.

Example 3.4. Our first example concerns the plane analytic geometry. We use (3.5) and (3.6) to give the explicit parametric (resp. cartesian) form of the line passing through points $P_1 = (x_1, y_1)'$ and $P_2 = (x_2, y_2)'$, $P_1 \neq P_2$. Setting $d_0 = P_1$ and $Z = P_2' - d_0'$ yields, by spectral decomposition of $Z'Z$,
$U = ((x_1 - x_2)/(y_1 - y_2), 1)'$ and $w = ((y_2 - y_1)/(x_1 - x_2), 1)'$. From (3.5), the parametric form of the line is

$$\mathcal{H} = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \alpha \begin{pmatrix} \frac{x_1 - x_2}{y_1 - y_2} \\ 1 \end{pmatrix}, \qquad \alpha \in \mathbb{R}, \tag{3.7}$$

whereas from (3.6), the cartesian form of the line is

$$\mathcal{H} = \{\begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2;\ \frac{y_2 - y_1}{x_1 - x_2}(x - x_1) + (y - y_1) = 0\}. \tag{3.8}$$

EXAMPLE 3.5. Our second example concerns the solid analytic geometry) We use (3.6) to give the cartesian form of the plane passing through points $P_1 = (x_1, y_1, z_1)'$, $P_2 = (x_2, y_2, z_2)'$ and $P_3 = (x_3, y_3, z_3)'$. Setting $d_0 = P_1$ and $Z = \begin{pmatrix} P_2' - d_0' \\ P_3' - d_0' \end{pmatrix}$ yields

$$w = \begin{bmatrix} (z_3 - z_1)(y_2 - y_1) - (z_2 - z_1)(y_3 - y_1) \\ (x_3 - x_1)(z_2 - z_1) - (z_3 - z_1)(x_2 - x_1) \\ (x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1) \end{bmatrix}, \tag{3.9}$$

and the cartesian form of the plane is

$$\mathcal{H} = \{ \begin{pmatrix} x \\ y \\ z \end{pmatrix} \in \mathbb{R}^3; w' \begin{pmatrix} x - x_1 \\ y - y_1 \\ z - z_1 \end{pmatrix} = 0 \}. \tag{3.10}$$

3.3. Application in statistics. Let $\overline{z} = \sum z_i/n$ be the mean of the z_i's (the latter being as in the corollary). When $\overline{z} \in \mathcal{D}$ is assumed, the corollary may be seen as a *sample* version of a constrained principal component analysis. In this section, we complete the picture by considering the question at the *population* level.

Although geometric by nature, the corollary has natural developments in multivariate statistics. Let X denote a p-variate random vector with mathematical expectation $\mathrm{E}(X) = \mu$ and covariance matrix $\mathrm{cov}(X) = \Psi$ (i.e., existence of second moments is assumed). Let $\mathcal{D}$ be an affine subspace of dimension d in $\mathbb{R}^p$, $0 \leq d \leq p$, and let d_0 be any vector of $\mathcal{D}$. If $d \geq 1$, let $D = [d_1|\ldots|d_d]$ denote a $p \times d$ matrix of rank d whose columns form a basis for the linear subspace $\mathcal{D} - d_0$. If $d = 0$, define D as the null vector of dimension $p \times 1$. In the following, we assume that $\mu \in \mathcal{D}$. Define the matrix

$$V_D = M_D \Psi M_D. \tag{3.11}$$

Assume that the rank of V_D is r, and let s be an integer such that $0 \leq s \leq r$. If $s \geq 1$, write $\lambda_1 \geq \ldots \geq \lambda_s$ for the s largest eigenvalues of V_D, assuming moreover $\lambda_s > \lambda_{s+1}$ (this will ensure that $\mathcal{M}(H)$ and P_H below are unique), and write $u_1, \ldots, u_s$ for orthonormal eigenvectors associated with $\lambda_1 \geq \ldots \geq \lambda_s$. Collect these vectors in the matrix $U = [u_1|\ldots|u_s]$. If $s = 0$, define U as the null vector of dimension $p \times 1$. Let $H = [U|D]$ denote the matrix formed by the blocks U and D. For $d_0 \in \mathcal{D}$, the projection of X onto the affine subspace $\mathcal{H} = d_0 + \mathcal{M}(H)$ is given by

$$Y = d_0 + P_H(X - d_0). \tag{3.12}$$

Importantly, note that Y does not depend on d_0. As by assumption $\mu \in \mathcal{D}$, we can define without loss of generality

$$Y = \mu + P_H(X - \mu). \tag{3.13}$$

To emphasize the fact that Y is the projection of X onto an affine subspace of dimension $s+d$, we will write $Y_{(s+d)}$ instead of Y.

Definition 3.6. We call $Y_{(s+d)}$ *the $(s+d)$-dimensional constrained principal component approximation* of X. One has $\mathrm{E}(Y_{(s+d)}) = \mu$, and $\mathrm{Cov}(Y_{(s+d)}) = P_H \Psi P_H$. Special cases: 1. The p-dimensional constrained principal approximation of X is X itself. 2. If $\mathcal{D} = \{\mu\}$, then $d = 0$, and the random vector

$$Y_{(s)} = \mu + P_U(X - \mu) \tag{3.14}$$

is called the *s-dimensional principal component approximation* of X, see [1].

We will now show the optimality of $Y_{(s+d)}$ in terms of the so-called mean squared difference.

Definition 3.7. Let X and Y denote two jointly distributed, p-variate random vectors. Then the *mean squared difference* between X and Y is defined by

$$\mathrm{MSD}(X, Y) = \mathrm{E}(||X - Y||^2). \tag{3.15}$$

Theorem 3.8. *Let X denote a p-variate random vector with mathematical expectation $\mathrm{E}(X) = \mu$. Let $\mathcal{D} = \mu + \mathcal{M}(D)$ denote an affine subspace of dimension d in $\mathbb{R}^p$, with $0 \leq d \leq p$. Suppose that $\mathrm{cov}(X) = \Psi$ exists, and define the matrix $V_D = M_D \Psi M_D$, that is assumed of rank r. Let s be an integer such that $0 \leq s \leq r$, and let the p-variate random vector Y denote the projection of X onto an affine subspace of dimension $s+d$ containing $\mathcal{D}$. Then $Y_{(s+d)}$ is optimal in the sense that*

$$\mathrm{MSD}(X, Y_{(s+d)}) \leq \mathrm{MSD}(X, Y) \tag{3.16}$$

for all Y.

Proof. Since Y is a projection of X onto an affine subspace $\mathcal{H}$, say, and since $\mu \in \mathcal{D} \subset \mathcal{H}$, Y can be written $Y = \mu + P_H(X - \mu)$, where $H = [U_{p\times s}|D_{p\times d}]$ (recall that D is a given matrix). Then $\mathrm{MSD}(X, Y) = \mathrm{E}(||X - Y||^2)$
$= \mathrm{E}(||X - \mu - P_H(X - \mu)||^2) = \mathrm{E}(||M_H(X - \mu)||^2) = \mathrm{E}[(X - \mu)' M_H (X - \mu)] = \mathrm{E}[\mathrm{tr}(X - \mu)' M_H (X - \mu)] = \mathrm{tr}(M_H \Psi)$. It is easy to show (using Theorem 3.1 that $\mathrm{MSD}(X, Y)$ is minimal for $H = [U^*|D]$, where the columns $u_1^*, \ldots, u_s^*$ of U^* are orthonormal eigenvectors associated with the eigenvalues $\lambda_1 \geq \ldots \geq \lambda_s$ of the matrix $V_D = M_D \Psi M_D$. This means that $\mathrm{MSD}(X, Y)$ is minimal for $Y = Y_{(s+d)}$. □

In the rest of this section, we make further assumptions for expositional convenience only (these assumptions can be easily lifted). Consider $V_D = M_D \Psi M_D$. First, assume that $\mathrm{r}(\Psi) = p$. Therefore $\mathrm{r}(V_D) \leq p - d$. Moreover, we assume that $\mathrm{r}(V_D) = p - d$. Define

$$H = [U|D], \tag{3.17}$$

where U is the $p \times (p-d)$ matrix whose columns $u_1, \ldots, u_{p-d}$ are orthonormal eigenvectors associated with the positive eigenvalues $\lambda_1 \geq \ldots \geq \lambda_{p-d}$ of V_D, and D is a (known) $p \times d$ matrix of rank d. Without loss of generality, we assume that the

Electronic Journal of Linear Algebra ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 10, pp. 201-211, July 2003

columns of D are orthonormal. Thus H is an orthogonal $p \times p$ matrix. We are now ready for the following definition:

DEFINITION 3.9. Let X denote a p-variate random vector with mathematical expectation $\mathrm{E}(X) = \mu$ and covariance matrix $\mathrm{cov}(X) = \Psi$. We call the p jointly distributed random variables

$$W = \begin{bmatrix} W_U \\ W_D \end{bmatrix} = H'(X - \mu) \tag{3.18}$$

the *constrained linear principal components* of X, where $W_U = U'(X - \mu) = (w_1, \ldots, w_{p-d})'$, and where $W_D = D'(X - \mu) = (w_{p-d+1}, \ldots, w_p)'$ is the vector of the constraints. The sample version of (3.18) is obtained by substituting the sample mean vector and the sample covariance matrix for their theoretical counterparts. By definition, the constraints will be retained in any analysis. The components that will be discarded in a particular analysis are the last entries of W_U, i.e., the constrained principal components corresponding to small eigenvalues of V_D. Note that the constrained principal component analysis generalizes the use of the so-called rotated principal components (see [6]).

As (unconstrained) linear principal component analysis is optimal in terms of variance maximization, introduction of *a priori* constraints – that can be very useful, see examples below – will generally imply some loss of optimality. We now give a measure for the loss of optimality induced by the constraints. The following lemma, which may be proved by simple arguments, will help us to define such a measure.

LEMMA 3.10. *Let X denote a p-variate random vector with mathematical expectation $\mathrm{E}(X) = \mu$ and covariance matrix Ψ. Let the p-variate random vector Q denote the projection of X onto an affine v-dimensional subspace $\mathcal{V}$ $(0 \le v \le p)$ containing μ. Then*

$$\mathrm{MSD}(X, \mu) = \mathrm{MSD}(Q, \mu) + \mathrm{MSD}(X, Q). \tag{3.19}$$

Note that the match of (3.19) at the sample level is

$$I(\mathcal{C}_X, \overline{x}) = I(\mathcal{C}_X^*, \overline{x}) + I(\mathcal{C}_X, \mathcal{V}), \tag{3.20}$$

where $\mathcal{C}_X$ is the cloud of points x_i formed by the rows (x_i') of the $n \times p$ matrix collecting the values of $X_1, \ldots, X_p$ on n individuals, $\overline{x}$ is the sample mean $\sum_{i=1}^n x_i/n$, $\mathcal{V}$ is an affine v-dimensional subspace of $\mathbb{R}^p$ containing $\overline{x}$, and $\mathcal{C}_X^*$ is the cloud formed by the points of $\mathcal{C}_X$ projected onto $\mathcal{V}$.

Recall that we have assumed here, for expositional convenience, that $\mathrm{r}(V_D) = p - d$. Suppose that we keep s constrained linear principal components of X in a particular analysis, where $1 \le s \le p - d$. Define $H = [U_{p\times s} | D_{p\times d}]$, where the columns of U are orthonormal eigenvectors of V_D associated with the s largest eigenvalues of this matrix. On the other hand, define $V = [v_1 | \ldots | v_{s+d}]$, where $v_1, \ldots, v_{s+d}$ are orthonormal eigenvectors of Ψ associated with the eigenvalues $\eta_1 \ge \ldots \ge \eta_{s+d}$ of Ψ. To ensure unicity of P_V, we assume moreover $\eta_{s+d} > \eta_{s+d+1}$. Let $Y_{(s+d)} = \mu + P_H(X - \mu)$ denote the $(s+d)$-dimensional constrained principal component approximation of X, and let $Y = \mu + P_V(X - \mu)$ be the (unconstrained) $(s+d)$-dimensional principal component

Electronic Journal of Linear Algebra ISSN 1081-3810
A publication of the International Linear Algebra Society
Volume 10, pp. 201-211, July 2003

ELA
http://math.technion.ac.il/iic/ela

approximation of X. That is, $Y_{(s+d)}$ is the projection of X onto the affine subspace $\mathcal{H} = \mu + \mathcal{M}(H)$, whereas Y is the projection of X onto $\mathcal{V} = \mu + \mathcal{M}(V)$. Both $\mathcal{H}$ and $\mathcal{V}$ have dimension $s+d$. Consider the identity (3.19). It is well known, and it may be readily shown, that Y is optimal in the sense that $\mathrm{MSD}(Y,\mu) \geq \mathrm{MSD}(Q,\mu)$ for all projection Q of X onto a $(s+d)$-dimensional affine subspace of $\mathbb{R}^p$ containing μ. Notably, we see that $\mathrm{MSD}(Y,\mu) \geq \mathrm{MSD}(Y_{(s+d)},\mu)$, and that $\mathrm{MSD}(Y,\mu) - \mathrm{MSD}(Y_{(s+d)},\mu)$ measures the amount of variance lost when constrained principal component analysis is used instead of usual principal component analysis with $s+d$ principal axes. We define the relative loss of optimality due to the constraints as

$$L_{(s+d)} = \frac{\mathrm{MSD}(Y,\mu) - \mathrm{MSD}(Y_{(s+d)},\mu)}{\mathrm{MSD}(Y,\mu)} = \frac{\mathrm{tr}(P_V\Psi) - \mathrm{tr}(P_H\Psi)}{\mathrm{tr}(P_V\Psi)}. \tag{3.21}$$

A natural estimator $\hat{L}_{(s+d)}$ is obtained by replacing Ψ by its sample counterpart C, or by any other estimate of Ψ:

$$\hat{L}_{(s+d)} = \frac{\mathrm{tr}(P_V C) - \mathrm{tr}(P_H C)}{\mathrm{tr}(P_V C)}. \tag{3.22}$$

An example where $\hat{L}_{(s+d)} = 0$ is the *rotated principal components* technique, see [2], [3] and [4], a special case of constrained principal component analysis, in which W_U in (3.18) is invariant in the sense that it is the vector of usual principal components. There is no loss of optimality (i.e., $\hat{L}_{(s+d)} = 0$ in (3.22) when rotation of factors is performed, as noted by Rencher (1998, p. 360).

We do not propose here constrained principal component analysis as a monolithic method, but as a useful exploratory tool that can be applied in a variety of domains. For example, if some of the p-variate observations are known to belong to groups, we may use the group information by forcing an axis to discriminate between the groups (in that case $\mathcal{D}$ would be the first discriminant axis), and then perform the constrained principal component analysis described above, to grasp as much variance as possible in the remaining axes. Although we remain here mainly at the descriptive level, we believe that a test of the hypothesis $\mathrm{tr}(P_V - P_H)\Psi = 0$ may be developed (see equation (3.21)). Low value of $\hat{L}_{(s+d)}$ may mean that there is no significant contradiction between the constraints and the data, i.e., that sample variation is large enough to encompass the constraints. A related question it to test the hypothesis that an eigenvector of Ψ has a particular value, $H_0 : v_j = v_{j0}$, versus $H_1 : v_j \neq v_{j0}$. This can be used to test the hypothesis that an eigenvector does not differ significantly from some simplified form.

4. Conclusion.

In this paper, an optimal subspace problem directly related to constrained principal component analysis is solved. The result is interesting *per se*. The approach - mainly geometrical and directed towards data analysis - yields a natural exploratory tool and allows to define a measure of the relative loss of optimality following the introduction of constraints.

REFERENCES

[1] B. Flury. *A First Course in Multivariate Statistics.* New York, Springer, 1997.

[2] I.T. Jolliliffe. Rotation of principal components: some comments. *Journal of Climatology*, 7:507-510, 1987.

[3] I.T. Jolliliffe. Rotation of ill-defined principal components: some comments. *Applied Statistics*, 38:139-147, 1989.

[4] I.T. Jolliliffe. Rotation of principal components: choice of normalization constraints. *Journal of Applied Statistics*, 22(1):29-35, 1995.

[5] J.R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics.* New York, John Wiley, 1988.

[6] A.C. Rencher. *Multivariate Statistical Inference and Applications.* New York, John Wiley, 1998.

[7] J.-D. Rolle. Best nonnegative invariant partially orthogonal quadratic estimation in normal regression. *Journal of the American Statistical Association*, 428:1378-1385, 1994.

[8] Y. Takane and T. Shibayama. Principal component analysis with external information on both subjects and variables. *Psychometrika*, 56:97-120,1991.

[9] Y. Takane, H.A.L. Kiers and J. de Leeuw. Component analysis with different sets of constraints on different dimensions. *Psychometrika*, 60:259-280,1995.