MAXIMUM LIKELIHOOD ESTIMATION IN PRINCIPAL COMPONENT ANALYSIS : A NEW METHOD IN GAUSSIAN MODEL

Jean-Louis PHILOCHE

Groupe des Ecoles des Télécommunications (GET)

In the Gaussian form of Principal Component Analysis, one starts with a vector Y whose distribution is $N_p(m,\Sigma)$. It is well known that the p principal components can be deduced from S. They are independent, with variances $l_1 \geq l_2 \cdots \geq l_p \geq 0$. We study the restricted model, where :

$$l_1 > l_2 > \cdots > l_q > l_{q+1} = \cdots = l_p > 0$$

An explicit expression for the maximum likelihood estimate (MLE) is given, its uniqueness is established and we give an explicit sequence relying only the data and converging to the unknown parameters. The maximization relies on differential geometry, which uses the simplifying action of the orthogonal group $SO(p)$.

1. Introduction and Main Theorem

In this paper, we propose to study the behaviour of the M.L.E out of an i.i.d. sample of size $n : (y_1,...,y_n)$, where each y_i follows $N_p(m,\Sigma)$, we assume that the eigenvalues (l_i) of S satisfy :

$$l_1 > l_2 > \cdots > l_q > l_{q+1} = \cdots = l_p > 0 \tag{1}$$

We call this situation the *restricted* model.

Naturally this restricted model has been studied in detail by many statisticians. We remark that part of the results obtained are available only through arguments which either involve delicate combinatorial arguments or lack sufficient details. See Luirhead [8] and Seber [10], for example.

We provide here some improvements of formulations, proofs and results.

Considering the year of publishing, the article *Asymptotic theory for principal components* (th. 2 p. 130 in [1]) of Anderson is, nevertheless, outstanding. The same is true for Lawley [6].

We prove in Theorem I, below, that (essentially), from a sample of n i.i.d. y_i with $N_p(m,\Sigma)$ as distribution, satisfying (1), we can find a *unique* ‘and precisely defined) M.L.E. of the *true* parameters. We also give an explic it sequence relying only on the data and converging to the unknown parameters.

The subset U of M^+ (M^+ *denotes the set of symmetric and positive definite operators of* $\mathbf{R}^p$) *is defined by the fact that a typical element* S of U *has eigenvalues* $(l_i)_{1\le i\le p}$ *which can be ordered so that (1) holds.*

The data are $y_1,\dots,y_n$, i.i.d. from $N_p(m,\Sigma)$. We easily get the loglikelihood $L_n(m,\Sigma)$ of $(y_1,\dots,y_n)^{\mathrm{T}}$, which is :

$$L_n(m,\Sigma) - \ln(\det(S) - \mathrm{Tr}(S\Sigma^{-1}) - \langle \bar{y}-m, \Sigma^{-1}(\bar{y}-m)\rangle .$$

$\langle .,.\rangle$ denotes the usual inner product of $\mathbf{R}^p$ and we denote : $\bar{y} := n^{-1}\sum_{i=1}^{n} y_i$, (the sample mean), and $S := n^{-1}\sum_{i=1}^{n}(y_i-\bar{y})(y_i-\bar{y})^{\mathrm{T}}$ (the sample covariance). It is clear that the eigenvalues of S are distinct and strictly positive (a.s), so we denote them (after ordering) : $s_1 > s_2 > \cdots > s_p > 0$.

Since $\langle \bar{y}-m, \Sigma^{-1}(\bar{y}-m)\rangle \ge 0$, *for each* Σ, $L_n(m,\Sigma)$ has a unique maximum : $\hat{m} = \bar{y}$. Therefore, we limit ourselves to finding the maximum of the function L, with :

$$L(\Sigma) := -\ln(\det(S) - \mathrm{Tr}(S\Sigma^{-1}), \text{ for } \Sigma \in U$$

Before stating the main theorem, it is important to give a hint on the proof : with the change of variables $\Theta := \Sigma^{-1}$, $F(\Theta) := L(\Theta^{-1})$, we obtain the strict concavity of F on M^+. Using the action of $SO(p)$, we observe that V image of F, is an union of orbits. So we first find the maximum of F on any orbit O, then the maximum of the maxima.

It is well-known that Σ can be written :

$$\Sigma = \sum_{j=1}^{q} l_j P_j + l_p \sum_{j=q+1}^{p} P_j ,$$

where each P_j is an orthogonal projection $P_j = P_j^T = P_j^2$, of rank 1, such that $P_j P_{j'} = 0$, $(j \ne j')$ and $\sum_j P_j = \mathrm{Id}(\mathbf{R}^{\mathbf{P}})$. (This is the classical spectral decomposition of Σ, satisfying (1)).

THEOREM. (I) *On* U, L *attains a maximum at a unique point* $\hat{\Sigma}$, *where the corresponding spectral decomposition is :*

$$\hat{\Sigma} = \sum_{j=1}^{q} s_j P_j + \frac{1}{p-q}\left(\sum_{k=q+1}^{p} s_k\right) Q ,$$

with $Q = \sum_{j=q+1}^{p} P_j$. The P_j are the same projections as the ones of the decomposition of Σ.

(II) *Moreover, any maximizing sequence* $(\Sigma_n)_{n\in\mathbf{N}}$ in U, *i.e. such that:*
$\lim_{n\to\infty} L(\Sigma_n) = L(\hat{\Sigma})$, *converge to* $\hat{\Sigma}$.

2. Classical Case

It might be useful to recall the well-known result concerning the *unrestricted* case : n i.i.d. observations $y_i \in \mathbf{R}^p$ are assumed to follow the Gaussian distribution $N_p(m,\Sigma)$; the likelihood is (up to multiplication by a constant) :

$$L(m,\Sigma) = (\det\Sigma)^{n/2} \exp\left[-\frac{1}{2} \sum_{i=1}^{n} (y_i - m)^{\mathrm{T}} \Sigma^{-1} (y_i - m) \right]$$

The M.L.E. method finds a couple (m,Σ) maximizing L_n, or, equivalently, $\ln L_n$. By setting the partial derivatives equal to zero, we get :

$$\hat{m} = \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \text{ , and}$$

$$\hat{\Sigma} = \frac{1}{n} S \text{, with } S = \frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})(y_i - \overline{y})^{\mathrm{T}} .$$

Of course, to establish also that this estimate is *the maximum* needs further arguments; each author gives a different proof. Good references are Anderson [2] (§3.2), or Seber [10] (§3.2).

In this specific (classical) case, the theorem similar to the result above is :

THEOREM. *Let $y_1,\dots,y_n$, i.i.d., following $N_p(m,\Sigma)$; assume that the eigenvalues l_j of Σ are distinct and strictly positive :*

$$l_1 > l_2 > \cdots > l_p > 0 , \qquad m \in \mathbf{R}^p ,$$

the (l_j) and m being unknown. Then the M.L.E. of the (l_j) and m is :

$$\hat{l}_k = s_k , 1 \le k \le p \text{, and } \hat{m} = \overline{y} ,$$

the s_k are the eigenvalues of the sample covariance :

$$S := n^{-1} \sum_{j=1}^{n} (y_j - \overline{y})(y_j - \overline{y})^{\mathrm{T}} .$$

In all this, we suppose, of course, $n > p$.

3. Principal component

(From a practical point of view.)

Let Y be a random vector in $\mathbf{R}^p$, with $E(Y) = m, \operatorname{cov} Y = \Sigma$, we suppose that Σ in not degenerate. So we can find an orthogonal transformation of $\mathbf{R}^p$ (let us denote it U), such that :

$$U^{\mathrm{T}} \Sigma U = \Lambda = \operatorname{diag}(l_1, \cdots, l_p) .$$

(with $l_1 \ge l_2 \ge \cdots \ge l_p > 0$).

The *principal component* are obtained from the original Y, using the linear transform U^{T}, we get new variables $\begin{pmatrix} U_1 \\ \vdots \\ U_p \end{pmatrix}$, called the principal components, $V(U_1) = l_1, V(U_2 = l_2, \cdots, V(U_p) = l_p$ (supposing all l_i distinct, $\mathrm{cov}(U_i, U_j) = 0$, for $i \neq j$).

For a concrete analysis, it might be useful to visualize (approximately) a sample of n realization $y_1, \ldots, y_n$ in $\mathbf{R}^p$, getting scatter plots (e.g. on the two-dimensional space generated by the first two principal component). Clearly, the situation is more fruitful when the data are assumed to be a sample from the gaussian distribution $N_p(m, \Sigma)$.

4. Proof of the Theorem

Befor beginning the proof of our theorem, it is important to introduce a *change of variables* :

$$\Theta := \Sigma^{-1} \qquad V := \left\{ \Theta \in M^+ : \Theta^{-1} \in U \right\},$$

and $F(\Theta) = L(\Theta^{-1}) := \ln(\det \Theta) - \mathrm{tr}(S\Theta)$.

This change of variables is convenient for the remainder of this paper. We shall establish that F is strictly concave on M^+; this property will contribute to simplify the argumentation to come.

Another useful tool will be *the action by conjugation of the group* $SO(p)$ on $L(p)$; this conjugation is defined by :

$$(U, \Sigma) \to U \Sigma U^{-1}, \text{ with } U \in SO(p) \text{ and } \Sigma \in L(p).$$

We recall that $SO(p)$ denotes the special orthogonal group of $\mathbf{R}^p$, and $L(p)$ denotes the set of all linear applications of $\mathbf{R}^p$.

We will also use, later on, the set V_S, *consisting in the operators of* V commuting with S.

Lemma 1. *F is strictly concave on M^+.*

Proof : First, we give the expression of the first two derivatives of the function F.

We observe that, for every X tangent to M^+ (thus symmetric), we get[1] :

$$\left\langle F'(\Theta), X \right\rangle = \mathrm{Tr}\left\{ \Theta^{-1} X \right\} - \mathrm{Tr}\left\{ SX \right\}.$$

[1] To establish this, we might use the formula : $\det(\Theta) = \exp[\mathrm{tr}(\ln \Theta)]$

Now, we turn to the proof that the function $F : M^{+} \longrightarrow \mathbf{R}$ is strictly concave. In fact, if X_1, X_2 are symmetric (i.e. tangent to M^{+}), the same reasoning gives :

$$F''(\Theta)(X_1, X_2) = -\mathrm{Tr}\left\{\Theta^{-1} X_1 \Theta^{-1} X_2\right\},$$

($F''(\Theta) \in L_2(E \times E, \mathbf{R})$, he set of bilinear forms on E), so :

$$F''(\Theta)(X, X) = -\mathrm{Tr}\left\{\left(\Theta^{-1} X\right)^2\right\} = -\mathrm{Tr}\left\{\left(\Theta^{-1/2} X \,\Theta^{-1/2}\right)^2\right\} < 0$$

for $X \neq 0$.

Lemma 2. *Let O (contained in V) be an orbit of the action of $SO(p)$ by conjugation. The critical points of the restriction of F to O necessarily belong to $O \cap V_S$. In particular, because O is compact, the points where the restriction of F to O attains its maximum belong to $O \cap V_S$. Moreover, the corresponding eigenvalues are ordered in the opposite sense as those of S.*

4.1 Proof of lemma 2

Let O be an orbit of the action of $SO(p)$, $O \subset V$. The critical points of the restriction of L to O are necessarily in $O \cap U_S$.

First, we identify the tangent space at $\Theta \in O$: it exactly consists in the communicators :

$$[\Theta, A] = (\Theta A - A\Theta), \ A \text{ antisymmetric.}$$

This result is classical.

We then notice that any critical point $\tilde{\Theta} \in O$ admits a simultaneous diagonalization with S. This is equivalent to commuting with S.

For a critical point $\tilde{\Theta}$ on O, the condition $\left\langle F'(\Theta), (\tilde{\Theta} A - A\tilde{\Theta})\right\rangle >= 0$ is equivalent to : $-\mathrm{Tr}\left\{S(\tilde{\Theta} A - A\tilde{\Theta})\right\} = \mathrm{Tr}\left\{(SA - AS)\tilde{\Theta}\right\} = 0$.

Representing operators by theirs matrix in an (orthonormal) basis, where S is diagonal, we find that $\tilde{\Theta}_{ij} = 0 \quad (i \neq j)$,

$$\left[\text{as} \quad \mathrm{Tr}\left\{\Theta^{-1}(\Theta A - A\Theta)\right\} = 0\right].$$

So, for an adequate change of basis in $\mathbf{R}^p$, both operators S and $\tilde{\Theta}$ can each be reduced into a diagonal matrix, with strictly positive elements on the diagonal.

The first matrix has as eigenvalues (when ordered) :

$$s_1 > s_2 > \cdots > s_p > 0$$

The matrix associated to $\tilde{\Theta}$ has eigenvalues : $\tilde{q}_1, \cdots, \tilde{q}_l$, with multiplicities $k_1, \cdots, k_l$ (of course $k_1 + \cdots + k_l = p$); we may suppose that the $(\tilde{q}_i)$ are chosen in increasing order (except for multiplicities).

We shall prove that $F(\Theta)$ Is maximum on O when the (s_i) and the $(\tilde{q}_i)$ are monotonic in opposite senses. This is a direct consequence of a Theorem of Hardy, Littlewood and Polya [4] (page 261, Theorem 368) which states[2] :

Inequalities. *If $(a_i)_{1\le i\le p}$ and $(b_i)_{1\le i\le p}$ are given, with the exception of a permutation $s \in G_p$, then $\sum_{i=1}^{p} a_i b_{s_i}$ is maximum then the (a_i) and (b_i) are monotonic in the same sense, and minimum in opposite sense.*

In our situation, $F(\tilde{\Theta}) = \ln(\det\tilde{\Theta}) - \mathrm{Tr}(S\tilde{\Theta})$; in the chosen basis, the first term is constant, while $\mathrm{Tr}(S\tilde{\Theta}) = \sum_{i=1}^{p} s_i \tilde{q}_{s_i}$ is minimum, according to result above if the permutation s is such that :

$$0 < \tilde{q}_1 < \tilde{q}_2 < \cdots < \tilde{q}_{q+1} = \cdots \tilde{q}_p .$$

This is the result we were looking for.
(Here, we observe that the maximum is attained for $(p-q)!$ different permutations.)

Lemma 3. The restriction of F to V_S *attains its maximum at a unique point $\tilde{\Theta}$, the eigenvalues $\tilde{q}_i$ of which satisfy :*

$$\tilde{q}_j = s_j^{-1},\ (1 \le j \le q) \quad \text{and} \quad \tilde{q}_j = \frac{p-q}{\sum_{k=q+1}^{p} s_k},\ (q < j \le p).$$

Proof of lemma 3 : From lemma 2, the maximum of F on an orbit O is reached for any $\tilde{\Theta}$, the eigenvalues $(\tilde{q}_j)$ of which are *disposed in the inverse order* as those of S.

As $\tilde{\Theta} \in V$, i.e. $\tilde{q}_1 = \cdots = \tilde{q}_{p-q1} > \tilde{q}_{p-q+1} > \cdots > \tilde{q}_p > 0$, we can choose an adequate parametrization of V , our goal being to maximise $F(\Theta)$, for Θ belonging to following open set D of $\mathbf{R}^{q+1}$, defined by :

$$0 < x_p < x_{p-1} < \cdots < x_{p-q} .$$

We already know that F is strictly concave (lemma 2), as it is easily seen that the equation $F''(\Theta) = 0$ has a unique maximum in the open set D, this solution is, necessarily, the unique maximum M in this domain.

Making explicit the coordinates of this point M , we obtain :

$$M_i = s_i^{-1},\ (1 \le i \le q) \quad \text{and} \quad M_i = \frac{p-q}{\sum_{k=q+1}^{p} s_k},\ q < j \le p$$

[2] Our formulation tries to make the statement more explicit

Lemma 4. *Let C be a finite dimensional convex set, and $\overset{\circ}{C}$ its relative interior. Let $j : C \longrightarrow \mathbf{R}$ a strictly convex function which attains a minimum at a point $m \in \overset{\circ}{C}$, then there exists $a > j(m)$ such that $\{x \in C; j(x) \le a\}$ is compact.*

Proof If $B \subset \overset{\circ}{C}$ is a compact ball centered at m, and $S = B \setminus \overset{\circ}{B}$, the corresponding sphere; j is continuous in $\overset{\circ}{C}$ and so attains a maximum on S : then, for $x \notin B, j(x) > a$ holds.

4.2 Proof of Part I

If $\Theta \in V \setminus V_S$, the restriction of F to the orbit of $SO(p)$ containing Θ does not admit Θ as a critical point (see lemma 2), so F cannot reach its maximum at this point. Moreover, using lemma 2 again, there exists Θ' in the same orbit, also element of V_S, such that $F(\Theta') < F(\Theta)$.

Then, the points where the restriction of F to V attains its maximum are necessarily in V_S : they are also the points where the restriction of F to V_S reaches its maximum.

Using lemma3, we see that there is a unique element $\tilde{\Theta}$ in V_S where the restriction of F to V_S attains its minimum. This element, characterized in lemma 3, is thus the only element of V at which the restriction of F to V attains its minimum. This element is completely characterized in lemma 3; the statement of Theorem (I) (giving the spectral decomposition of $\tilde{\Theta}$) is only another way of writing the same facts.

As above, expressing the result in terms of (Σ, l_j) gives theorem (I).

4.3 Proof of Part II

Let us denote by M_S^+ the set $\{\Theta \in M^+; \Theta S = S\Theta\}$: this set is an affine manifold[3] (closed in M^+). The affine manifold $M_S^+ \cap V$ is not convex, but each of its connected components is. We denote by V_S° the connected component containing the maximum $\hat{\Theta}$.

According to lemma 4 above, there exists a $(a < F(\tilde{\Theta}))$, such that :

$$K : \{\Theta \in V_S^\circ; F(\Theta) \ge a\}$$

is a compact set. Let $K' \subset V$ be the union of K under $SO(p)$. K' is clearly a compact set (as image of $K \times SO(p)$) and $\{\Theta; \Theta \in V; F(\Theta) \ge a\} \subset K'$.

The reason for this is that, if $F(\Theta) \ge a$, F attains, on the orbit of Θ, a maximum at a point $\tilde{\Theta} \in V_S^\circ$, then $F(\tilde{\Theta}) \ge F(\Theta) \ge a$, thus $\tilde{\Theta} \in K$, so $\tilde{\Theta} \in K'$.

[3] We recall that such a manifold is nothing but an open set in an affine set.

Let $\{\Theta_n \in V, n \in \mathbf{N}\}$ be a maximizing sequence; it is clear that $\Theta_n \in K'$, for n large enough. This ends the proof.

From this, we deduce directly the theorem (II).

5. Acknowledgments

I would like to thank Gérard Letac, Professor at Paul Sabatier University (Toulouse, France), Pierre Renouard, Member of the Theoretical Physics Center of the Ecole Polytechnique (Palaiseau, France), Peter Jupp, Reader at School of Mathematics and Statistics (St Andrews, Scotland) for many constructive discussions on the subject, and Damien Wyart for pertinent remarks and useful computer help.

References

[1] Anderson, T.W. (1963). *Asymptotic theory for principal components*. In Annals of Mathematicals Statistics (p. 122-148).

[2] Anderson, T.W. (1984). An *Introduction to Multivariate Analysis*, second edition, John Wiley.

[3] Fine, J. (1992). *Modèles pour l'analyse des données multidimensionnelles*. In Droesbeke, J.J. Fichet, B., and Tassi, P., Economica, Paris.

[4] Hardy, G.H., Littlewood J.E., Polya G. (1967). *Inequalities*, second edition, Cambridge University Press.

[5] Helgason, S. (1962). *Differential Geometry and Symmetric Spaces*, Academic Press

[6] Lawley, D.N. (1940). *The estimation of Factor loading by the Method of Maximum Likelihood*, in Proceedings of the Royal Society of Edinburgh, section A 60, p. 64-82.

[7] Lebart, L., Morineau, A., Tabard, N. (1977). *Technique de la description statistique*. Dunod, Paris.

[8] Muihead, R.J. (1982). *Aspects of Multivariate Analysis. Academic Press*, London.

[9] Philoche J.-L., Mitteau, J.-C. (1996). *Utilisation des groupes de Lie pour l'estimation des paramètres d'un modèle normal multidimensionnel bruité*. In CRAS, Paris, 323, Série 1, p. 939-942, Statistics.

[10] Seber, G.A.F. (1984). *Multivariate Observations*, John Wiley.

[11] Thurstone, L.L. (1947). *Multiple-Factor Analysis*, University of Chicago Press.

[12] Weyl, H. (1973). eighth edition, *The classical Groups*, Princeton University Press.