PSYCHOMETRIKA—VOL. 64, NO. 3, 317–324
SEPTEMBER 1999

RETRIEVING THE CORRELATION MATRIX FROM A TRUNCATED PCA SOLUTION: THE INVERSE PRINCIPAL COMPONENT PROBLEM

JOS M.F. TEN BERGE AND HENK A.L. KIERS

UNIVERSITY OF GRONINGEN

When r Principal Components are available for k variables, the correlation matrix is approximated in the least squares sense by the loading matrix times its transpose. The approximation is generally not perfect unless $r = k$. In the present paper it is shown that, when r is at or above the Ledermann bound, r principal components are enough to perfectly reconstruct the correlation matrix, albeit in a way more involved than taking the loading matrix times its transpose. In certain cases just below the Ledermann bound, recovery of the correlation matrix is still possible when the set of all eigenvalues of the correlation matrix is available as additional information.

Key words: Ledermann bound, principal component analysis, inverse eigenvalue problems

It is well-known that a $k \times k$ correlation matrix $\mathbf{R}$ can be retrieved from a $k \times r$ Principal Components loading matrix $\mathbf{A}$ as $\mathbf{R} = \mathbf{AA}'$, if and only if the number of components r is as high as the rank of the correlation matrix. In practice, this means that we need as many components as variables to have $\mathbf{AA}' = \mathbf{R}$. However, the need for retrieving the correlation matrix from $\mathbf{A}$ seems to arise exclusively in cases of incomplete principal components analysis (PCA), where $r < k$: One may wish to compare a published loading matrix with one to be obtained in a different manner, or in a replication study, while the correlation matrix is not available. For instance, Carroll (1993, p. 83) discusses the problem of missing correlation matrices hampering his attempt to replicate previous factor analysis studies. It is common belief that, in situations like these, the correlation matrix $\mathbf{R}$ cannot be determined when only $\mathbf{A}$ is available. The purpose of the present paper is to challenge this belief.

It is instructive to first consider the case $k = 3$, $r = 2$, where we have a 3×2 matrix $\mathbf{A}$ of loadings of three variables on two principal components, possibly rotated. In this case, the third eigenvector of $\mathbf{R}$ is known at once as the unique vector orthogonal to the two columns of $\mathbf{A}$. The third eigenvalue of $\mathbf{R}$ is also known, being 3 minus $\mathrm{tr}(\mathbf{A}'\mathbf{A})$. It follows that, in the 3×2 case, the PCA loading matrix $\mathbf{A}$ does determine $\mathbf{R}$. In fact, the same logic applies to any case where r, the number of principal components, is $k - 1$. The question is to what extent retrieving $\mathbf{R}$ from $\mathbf{A}$ is possible for smaller (less trivial) values of r. The present paper solves this "inverse principal component problem." The solution relies on the so-called Ledermann bound, which has a long history in the realm of factor analysis.

1. The Ledermann Bound

Ledermann (1937) was concerned with the question how many factors it takes to find a factor loading matrix $\mathbf{P}$ such that $\mathbf{PP}'$ has the same off-diagonal elements as a given $k \times k$ correlation matrix $\mathbf{R}$. From a count of the number of equations and the number of unknown parameters, he inferred that perfect fit in factor analysis is possible when the number of factors is at or above the function $\phi(k)$ of the number of variables k, defined as

$$\phi(k) = \frac{1}{2}[2k + 1 - (8k + 1)^{1/2}]. \tag{1}$$

Requests for reprints should be sent to Jos M.F. ten Berge, Heijmans Institute, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, THE NETHERLANDS. Email: j.m.f.ten.berge@ppsw.rug.nl

0033-3123/1999-3/1996-0532-A $00.75/0

This function has become known as the Ledermann bound. It was meant to be an *upper* bound to the number of factors, but this optimistic view was shattered by counterexamples by Wilson and Worcester (1939) and Guttman (1958). In fact, Shapiro (1982) has shown that Ledermann's bound is almost surely a *lower* bound to the number of factors. Regardless, Ledermann's bound has been a key concept in factor analytic theory. Apart from the number of factors issue, it appears in local and global uniqueness theorems for communality estimates, see Shapiro (1985) and Bekker and ten Berge (1997), respectively, and in a theorem on the stability of minimum rank (Shapiro, 1982).

So far, Ledermann's bound has never emerged in the realm of principal components analysis. However, the key result of the present paper is that retrieving the correlation matrix **R** from a $k \times r$ principal component loading matrix **A** is usually possible when r is at or above the Ledermann bound. Specifically, when k is

$$3\ 4\ 5\ 6\ 7\ 8\ 9\ 10 \ldots 15 \ldots 21,$$

respectively, then the smallest r at or above the Ledermann bound $\phi(k)$ is

$$1\ 2\ 3\ 3\ 4\ 5\ 6\ 6 \ldots 10 \ldots 15,$$

respectively. So the claim that "**A** determines **R** when r is at or above the Ledermann bound" means that it takes just one principal component to determine the correlation matrix of $k = 3$ variables; it takes two components when $k = 4$; it takes 3 components when $k = 5$ or 6, and so on. To be sure, **R** cannot simply be derived as $\mathbf{AA}'$ in these cases. More involved computations, to be explained below, will be needed.

Retrieving **R** from **A** when r is below the Ledermann bound is generally not possible. However, when additional information in the form of the complete set of k eigenvalues of **R** is available, we can retrieve **R** in certain cases where r is just below the Ledermann bound. Before explaining the details of this extension, we consider cases at or above the Ledermann bound.

2. Retrieving R from A at or Above the Ledermann Bound

Suppose we have a $k \times r$ unrotated or rotated PCA loadings matrix **A**, and we want to retrieve **R**. The problem will be approached as one of finding the set of $k \times q$ matrices $\mathbf{A}_2$, with $q = k - r$, such that the columns of $\mathbf{A}_2$ are orthogonal to those of **A**, and $\mathbf{AA}' + \mathbf{A}_2\mathbf{A}_2'$ has unit diagonal elements. This set, henceforth to be referred to as the feasible set, is nonempty, because the truncated part of the full $k \times k$ PCA loading matrix is an element of it, and so is any orthogonally rotated version of that loading matrix. When we have determined the entire feasible set, and it contains just one element (up to a rotation), we have retrieved **R** from **A**.

Without loss of generality, $\mathbf{A}_2$ can be written as **BW**, where **B** is any column-wise orthonormal $k \times q$ matrix satisfying $\mathbf{B}'\mathbf{A} = \mathbf{0}$ and **W** is a $q \times q$ transformation matrix. Let $\mathbf{b}_1', \ldots, \mathbf{b}_k'$ be the rows of **B**, and let $\delta_i = 1 - \mathbf{a}_i'\mathbf{a}_i$, where $\mathbf{a}_i'$ is row i of **A**, $i = 1, \ldots, k$. Searching for $\mathbf{A}_2$, such that $\mathbf{AA}' + \mathbf{A}_2\mathbf{A}_2'$ has unit diagonal elements, boils down to solving the system of equations

$$\mathbf{b}_i'\mathbf{WW}'\mathbf{b}_i = \delta_i, \tag{2}$$

$i = 1, \ldots, k$. Define $\mathbf{G} \equiv \mathbf{WW}'$, and write (2) in Vec-notation as

$$\text{Vec}(\mathbf{b}_i'\mathbf{G}\mathbf{b}_i) = (\mathbf{b}_i' \otimes \mathbf{b}_i')\,\text{Vec}(\mathbf{G}) = \delta_i, \tag{3}$$

where $\otimes$ is the Kronecker product and the result that $\text{Vec}(\mathbf{XYZ}) = (\mathbf{Z}' \otimes \mathbf{X})\,\text{Vec}(\mathbf{Y})$ is used. Upon removing the entries of Vec(**G**) that are redundant due to the symmetry of **G**, and the corresponding elements of $(\mathbf{b}_i' \otimes \mathbf{b}_i')$, (3) can be expressed equivalently as

$$\mathbf{Hg} = \mathbf{d}, \tag{4}$$

where $\mathbf{d}$ is the k-vector containing $\delta_1, \ldots, \delta_k$; $\mathbf{g}$ is the $\frac{1}{2}q(q+1)$-vector containing just the diagonal elements of $\mathbf{G}$ and those below the diagonal, arranged as $g_{11}, g_{21}, \ldots, g_{q1}, g_{22}, \ldots, g_{q2}, \ldots, g_{qq}$; and $\mathbf{H}$ contains, in the same order, the Hadamard (elementwise) products of the column pairs β_i and $\beta_j (i \geq j)$ of $\mathbf{B}$, the Hadamard products $\beta_i . \beta_j (i > j)$ being multiplied by 2, to make up for omitting those elements above the diagonal of $\mathbf{G}$ from $\mathbf{g}$. It remains to solve (4).

It is clear that $\mathbf{H}$ is a $k \times t$ matrix, with $t = \frac{1}{2}q(q+1)$. It is well-known (Shapiro, 1982: in particular Lemma 2.2 and the proof of Theorem 2.1) that the rank of $\mathbf{H}$ is a property of the column space of $\mathbf{B}$ alone, and that, when $k \geq t$, $\mathbf{H}$ is of full column rank t almost surely (the set of exceptions has Lebesgue measure zero). Therefore, when $k \geq t$, (4) has a unique solution almost surely, namely $\mathbf{g} = (\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{d}$. Whenever this is the unique solution, $\mathbf{A}_2$ can be retrieved as $\mathbf{B}\mathbf{G}^{1/2}$, up to a rotation, and we can retrieve $\mathbf{R}$ as $\mathbf{R} = \mathbf{A}\mathbf{A}' + \mathbf{A}_2\mathbf{A}_2'$. The requirement $k \geq t$ can be rewritten as $k \geq \frac{1}{2}(k-r)(k-r+1)$. It is well-known (e.g., Shapiro, 1982, p. 191) that this inequality is equivalent to $r \geq \phi(k)$, with $\phi(k)$ as defined in (1). It follows that we can almost surely retrieve $\mathbf{R}$ from $\mathbf{A}$ when r is at or above the Ledermann bound.

When the number of components is below the Ledermann bound, there is an infinite number of solutions for $\mathbf{g}$ in (4). Each of these implies a solution for $\mathbf{G}$ (symmetric), which may or may not be positive definite. As soon as two different solutions for $\mathbf{G}$ arise, both Gramian, it follows that the feasible set contains at least two distinct elements (differing by more than just an orthogonal rotation), and $\mathbf{R}$ cannot be retrieved from $\mathbf{A}$. In practice, this does indeed happen when r is below the Ledermann bound. Both retrieval and non-retrieval are demonstrated in Appendix A, where Harman's five socio-economic variables are analysed with $r = 3$ and $r = 2$, respectively.

3. Retrieving R From A and Eigenvalues

When r is below the Ledermann bound, having $\mathbf{A}$ is not enough to retrieve $\mathbf{R}$. However, when, in addition to $\mathbf{A}$, the full set of eigenvalues of $\mathbf{R}$ is also available (SPSS-PCA does print all eigenvalues, and some authors do report those), retrieval of $\mathbf{R}$ may be possible in certain cases. To examine this possibility, we shall use the property that the last q eigenvalues of $\mathbf{R}$ are the eigenvalues of $\mathbf{G}$. To verify this, note that $\mathbf{A}_2 = \mathbf{B}\mathbf{G}^{1/2}\mathbf{T}$, for some orthonormal $\mathbf{T}$, with $\mathbf{B}$ column-wise orthonormal and orthogonal to $\mathbf{A}$. So when $\mathbf{G}$ has the eigendecomposition $\mathbf{G} = \mathbf{L}\mathbf{\Lambda}\mathbf{L}'$ ($\mathbf{L}$ orthonormal; $\mathbf{\Lambda}$ diagonal), $\mathbf{A}_2'\mathbf{A}_2$ has the eigendecomposition $\mathbf{A}_2'\mathbf{A}_2 = \mathbf{T}'\mathbf{L}\mathbf{\Lambda}\mathbf{L}'\mathbf{T}$, hence it has the same eigenvalues as $\mathbf{G}$. It is obvious that the eigenvalues of $\mathbf{A}_2'\mathbf{A}_2$ are the last q eigenvalues of $\mathbf{R}$.

When the elements of $\mathbf{\Lambda}$ are known, we are concerned with a "feasible subset," containing all solutions for $\mathbf{G}$ that have the prescribed eigenvalues. If we can determine the entire feasible subset, and it contains just one element, $\mathbf{R}$ can be retrieved from $\mathbf{A}$ and the eigenvalues of $\mathbf{R}$. It will now be explored to what extent this approach is useful.

The general solution for (4), which generates all candidates for the feasible set, has the form

$$\mathbf{g} = \mathbf{H}^{+}\mathbf{d} + \mathbf{N}\mathbf{y}, \tag{5}$$

where $\mathbf{H}^{+}$ is the Moore–Penrose inverse of $\mathbf{H}$, $\mathbf{N}$ is a $k \times (t-k)$ orthonormal basis for the null space of $\mathbf{H}$, and $\mathbf{y}$ is an arbitrary vector. Upon writing $\mathbf{N}\mathbf{y}$ as $y_1\mathbf{n}_1 + y_2\mathbf{n}_2 \ldots + y_{t-k}\mathbf{n}_{t-k}$ and constructing symmetric matrices $\mathbf{G}_1, \mathbf{N}_1, \ldots, \mathbf{N}_{t-k}$ from the elements of $\mathbf{H}^{+}\mathbf{d}$ and $\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_{t-k}$, respectively, every element $\mathbf{G}$ of the feasible set can be written as a function of the scalars $y_1, y_2, \ldots, y_{t-k}$ as

$$\mathbf{G} = \mathbf{G}_1 + \sum_{i=1}^{t-k} y_i\mathbf{N}_i. \tag{6}$$

The feasible subset specifies the eigenvalues $\mathbf{G}$ must have. As has been shown above, these eigenvalues are $\lambda_{r+1}, \lambda_{r+2}, \ldots, \lambda_k$, the last q eigenvalues of $\mathbf{R}$. Hence, the elements of the feasible subset must satisfy the q determinantal equations

$$|\mathbf{G} - \lambda_j \mathbf{I}_q| = |\mathbf{G}_1 + \sum_{i=1}^{t-k} y_i \mathbf{N}_i - \lambda_j \mathbf{I}_q| = 0, \tag{7}$$

$j = r + 1, \ldots, k$.

It is important to note that these q equations contain a redundancy: When (7) is satisfied for $q - 1$ different values of j, it is satisfied for all q values of j. This follows from the fact that every $\mathbf{G}$ satisfying (3) already has the correct sum of eigenvalues, this sum being equal to the sum of elements of $\mathbf{d}$. In general, when $t > k$, the $q - 1$ equations are not enough to determine $y_1, y_2, \ldots, y_{t-k}$, and hence $\mathbf{G}$, uniquely. However, there are some exceptions, that we shall now deal with.

The most obvious exception is the case where $t - k = 1$. In this case, $\mathbf{N}$ is a vector almost surely, and we can narrow down the feasible subset by collecting all values of y_1 that solve the $q - 1$ determinantal equations

$$|\mathbf{G}_1 + y_1 \mathbf{N}_1 - \lambda_j \mathbf{I}_q| = 0, \tag{8}$$

$j = r+1, \ldots, k-1$. Solving (8) for y_1 amounts to solving $q - 1$ polynomial equations of degree q, and yields $q - 1$ sets of q solutions for y_1. Only solutions that occur in each of the $q - 1$ sets need be considered: The matrices $\mathbf{G}$, associated with these solutions by virtue of (6), have the correct eigenvalues. We know that at least one $\mathbf{G}$ exists that does have the correct eigenvalues, which means that the feasible subset is not empty. If this is the only solution with the correct eigenvalues, we have retrieved $\mathbf{R}$.

The viability of the determinantal approach can easily be verified in the case $k = 5, r = 2$, where $t - k = 1$. The sets of three roots of the two determinantal equations (8) are readily obtained by standard techniques for solving a set of polynomial equations, and the correct solution invariably appears as the unique common solution for y_1 across the two sets.

Generally, cases where $t - k = 1$ can be shown to arise when $\phi(k + 1)$ is an integer and $r = \phi(k+1) - 1$. Hence, other cases besides $\{k = 5; r = 2\}$ are $\{k = 9; r = 5\}$, $\{k = 14, r = 9\}$, $\{k = 20, r = 14\}$, and so on. Retrieval of the correct solution can now again be verified by solving (8) for $q - 1$ different values λ_j. For this purpose, however, we have adopted a least-squares procedure, which can also be applied when $t - k > 1$. This will be explained in the next section.

4. An Alternating Least Squares Method to Solve (7) for $t - k \geq 1$

The inverse principal component problem is a constrained inverse eigenvalue problem, for example, see Friedland, Nocedal & Overton (1987) and Chu (1998). Neither of the two versions of our problem (the simple one with only the loadings on r principal components available, and the extended one with all eigenvalues available additionally) seems to have been considered in the literature on "Inverse Eigenvalue Problems." However, the specific (unconstrained) subproblem of solving (7) does explicitly appear in Chu (1998, p. 26) as the Least Squares Inverse Eigenvalue Problem LSIEPa1, also see Chen and Chu (1996, p. 2419). Chen and Chu (p. 2424) suggest solving (7) by minimizing the function

$$f(y_1, \ldots, y_{t-k}, \mathbf{U}) = \|\mathbf{G}_1 + \sum_{i=1}^{t-k} y_i \mathbf{N}_i - \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}'\|^2 \tag{9}$$

with $\mathbf{U}$ constrained to be orthonormal. They developed a hybrid LP-Newton algorithm for this purpose. However, we have used the following alternating least squares method instead:

Step 0: Initialize $\mathbf{y} = [y_1 | \ldots | y_{t-k}]'$ randomly, and $\mathbf{U}$ by Step 2 below.

Step 1: Determine the best conditional solution for $\mathbf{y}$ given $\mathbf{U}$ by minimizing $\| \mathrm{Vec}(\mathbf{U\Lambda U}' - \mathbf{G}_1) - \mathbf{Ny}\|^2$. Update $\mathbf{y}$ and go to Step 2.

Step 2: Determine the best conditional solution for $\mathbf{U}$ given $\mathbf{y}$ by taking the eigendecomposition $\mathbf{G}_1 + \sum_{i=1}^{t-k} y_i\mathbf{N}_i = \tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{U}}'$, with $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \ldots \geq \tilde{\lambda}_{t-k}$. Update $\mathbf{U}$ by $\tilde{\mathbf{U}}$ and go to Step 1, until convergence.

The conditional optimality of Step 1 is obvious. The conditional optimality of Step 2 follows from Friedland (1977, Theorem 3.1).

The alternating least squares procedure was used on simulated correlation matrices. To obtain random correlation matrices with realistic eigenvalues, we sampled no more than 30 cases with scores on k variables from a multivariate uncorrelated standard normal distribution. For each correlation matrix, the first r principal components are determined, and these are used in the program, together with the last eigenvalues (diagonal elements of $\mathbf{\Lambda}$), to determine the feasible subset. After convergence of the program, we take it that the global minimum has been attained when the function value of f is below 10^{-6}. When this condition is met, the procedure is considered to have retrieved $\mathbf{R}$ when the sum of squared differences between elements of the obtained and the original correlation matrix is less than 10^{-20}.

The procedure first has been tested in four cases with $t - k = 1$, namely, $\{5, 2\}$, $\{9, 5\}$, $\{14, 9\}$ and $\{20, 14\}$, with 10 replications of the correlation matrix, using 50 random starts. The global optimum was attained in 1569 (out of $50 \times 10 \times 4 = 2000$) runs. In all these runs, the correct solution for $\mathbf{G}$ appeared as the unique element of the feasible subset.

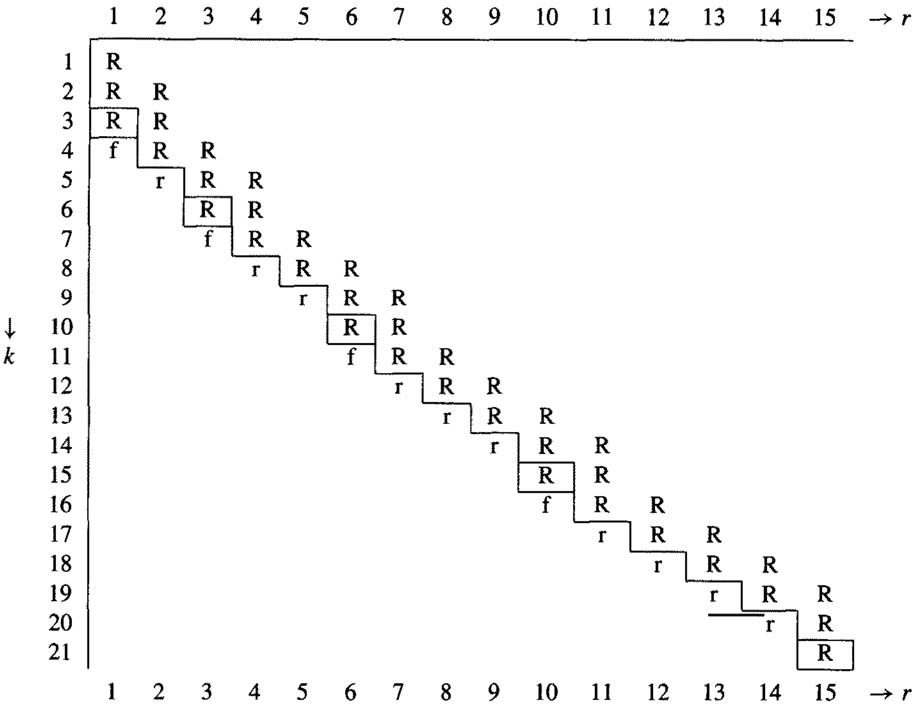
Next, various cases below the Ledermann bound, with $k \leq 21$, were examined. Surprisingly, uniqueness of $\mathbf{G}$ also appeared for values of $t - k$ higher than 1: In fact, no counter-examples to uniqueness below the Ledermann bound were found in those cases where the number of independent equations in (7) exceeds the number of parameters $y_1, \ldots, y_{t-k}$. This means that uniqueness below the Ledermann bound was found to hold if and only if $q - 1 > t - k$, or, equivalently, if and only if

$$r > \frac{1}{2}[2k - 1 - (8k - 7)^{1/2}]. \tag{10}$$

The $\{k, r\}$ combinations in point, including those with $t - k = 1$, for $k \leq 21$, are $\{5, 2\}$, $\{8, 4\}$, $\{9, 5\}$, $\{12, 7\}$, $\{13, 8\}$, $\{14, 9\}$, $\{17, 11\}$, $\{18, 12\}$, $\{19, 13\}$, and $\{20, 14\}$. Although we have no mathematical proof for the uniqueness encountered in these cases, the plausibility of our findings can be explained. We shall consider the case $\{8, 4\}$ as an example.

In the case $k = 8$, $r = 4$, we have $q = 4$, and $t - k = 2$, which means that three equations, see (7), have to be solved for y_1 and y_2. Each equation gives at most four different real values for y_2 given y_1 and vice versa. Therefore, each equation defines at most four curves in the real plane. The curves defined by the first two equations intersect only in isolated points. Usually, these points will not be on the curves defined by the third equation. In fact, when we deliberately replace the correct eigenvalues by incorrect eigenvalues with the same sum, we seem to find no solutions. Whenever the correct eigenvalues are used, however, we know that there is at least one solution for the three equations. Still, it would be unusual to have more than one solution.

The argument of the previous paragraph can be generalized to hypercurves in $t - k$ space, when $t - k > 2$. The argument implies that, in cases where the number of equations is *equal* to the number of parameters, only a small number of solutions are to be expected, among which the "correct one." Cases in point are $\{4, 1\}$, $\{7, 3\}$, $\{11, 6\}$, $\{16, 10\}, \ldots$ These cases arise when k is taken one higher than in the cases where Ledermann's bound is an integer. The numerical results for these cases confirmed the expectation, in the sense that for each case, the 50 runs of the algorithm led to a limited number of solutions, some of which were identical. For instance, when $k = 4$, $r = 1$, for each case we found at least 28 solutions (out of 50 runs), of which at most six were different. For the other cases, similar results were obtained. It should be noted that



↓ k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	→ r
1	R															
2	R	R														
3	$\boxed{\text{R}}$	R														
4	f	R	R													
5		r	R	R												
6			$\boxed{\text{R}}$	R												
7			f	R	R											
8				r	R	R										
9					r	R	R									
10						$\boxed{\text{R}}$	R									
11						f	R	R								
12							r	R	R							
13								r	R	R						
14									r	R	R					
15										$\boxed{\text{R}}$	R					
16										f	R	R				
17											r	R	R			
18												r	R	R		
19													r	R	R	
20														r	R	
21															$\boxed{\text{R}}$	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	→ r

FIGURE 1.
Cases "**R**" with retrieval from **A**, cases "*r*" with retrieval from **A** and **Λ**, and cases "*f*" with a finite number of solutions.

by using a limited number of runs, we cannot guarantee that among the solutions found, we will find the solution corresponding to the original correlation matrix, although in a large majority of cases we did.

As we have already mentioned above, no uniqueness has been found when r is below the value stipulated in (10). This conclusion is based on simulations in the cases {5, 1}, {6, 2}, {7, 2}, {8, 3}, {9, 4}, {10, 5}, {11, 5}, {12, 6}, {13, 7}, {14, 8}, {15, 9}, {16, 9}, {17, 10}, {18, 11}, {19, 12}, {20, 13} and {21, 14}. These cases were chosen because they are close to the cases at or above the Ledermann bound, but fall short of satisfying (10). The correct solution never emerged here as an element of the feasible subset. It can be concluded that retrieval of **R** from **A** and eigenvalues is not possible, in these cases. The same goes a fortiori for cases farther away from the Ledermann bound, like {6, 1}, {7, 1}, {8, 2}, and so on.

In Figure 1, the results are presented schematically. The boxes indicate cases where the Ledermann bound is an integer. For cases at or above this bound, some of which have been marked "R", retrieval of **R** from **A** is possible. For those cases below the Ledermann bound, marked "r", retrieval from **A** and **Λ** is possible, and for the cases marked "f" we obtain a finite number of different solutions, including the correct one, when **A** and **Λ** are given. The empty cells below the Ledermann bound are associated with an infinite number of solutions.

5. Discussion

Retrieving the correlation matrix from a truncated PCA solution has been shown possible at or above the Ledermann bound. This means that retrieving **R** from a published matrix of component loadings is possible in far more instances than is commonly believed. Although the issue of recovery of **R** from **A** initially arose from a practical question, the theoretical value of the result seems greater than its practical usefulness, which is rather limited because the number of components maintained in practice is typically below the Ledermann bound.

From a theoretical point of view, it is worth knowing that the Ledermann bound, so eminent in the context of factor analysis, also bears on the question how many components to maintain in PCA: Taking more than r components, when r is at or above the Ledermann bound, would be redundant, because a smaller number of components already carries all information there is. Of course, this does not mean that all variance is explained by these components (we do not have $\mathbf{AA}' = \mathbf{R}$, when $\mathbf{A}$ is $k \times r$, $r < k$). However, it does mean that the last $k - r$ components are fully determined by the first r components. Taking more than r components would create an embarrassing redundancy in that sense.

Although we have studied correlation matrices only, the unit diagonal property is not essential. To get retrieval of any Gramian matrix with known diagonal elements, it is sufficient to redefine δ_i, $i = 1, \ldots, k$, in (2). Hence, the result of this paper implies that, for any Gramian matrix, a complete retrieval is possible when only the diagonal elements and the first, or, in fact, any r eigenvectors and eigenvalues, with r at or above the Ledermann bound, are known. Also, any such a matrix can be decomposed as $\mathbf{AA}' + \mathbf{A}_2\mathbf{A}'_2$, with $\mathbf{A}$ and $\mathbf{A}_2$ of order $k \times r$ and $k \times (k - r)$, respectively, with $\mathbf{A}_2$ a function of $\mathbf{A}$, for any $r \geq \phi(k)$. These are unexpected and potentially useful results for Gramian matrices in general.

For certain cases just below the Ledermann bound, retrieval of $\mathbf{R}$ on the basis of $\mathbf{A}$ has also appeared possible when additional information in the form of the eigenvalues of $\mathbf{R}$ is available. It is conceivable that other forms of additional information may occur in certain contexts, for example, single elements of $\mathbf{R}$ are sometimes given. In such cases, a different strategy would be needed to retrieve $\mathbf{R}$. We have not pursued this possibility. It should be noted, however, that having just the determinant of $\mathbf{R}$ (which implies $|\mathbf{A}|$) as additional information is certainly not enough: In the search for the feasible subset, different solutions for $\mathbf{G}$, all implying the correct determinant of $\mathbf{G}$ and $\mathbf{R}$, appear to display different sets of (positive) eigenvalues more often than not. The example of Appendix A is a case in point.

Appendix A: Harman's Five Socio-Economic Variables

Consider the correlation matrix $\mathbf{R}$ and the 5×3 PCA loading matrix $\mathbf{A}$ for Harman's five socio-economic variables (Harman, 1967, p. 14, p. 137):

$$\overset{\mathbf{R}}{\begin{bmatrix} 1.0000 & 0.0100 & 0.9720 & 0.4390 & 0.0220 \\ 0.0100 & 1.0000 & 0.1540 & 0.6910 & 0.8630 \\ 0.9720 & 0.1540 & 1.0000 & 0.5150 & 0.1220 \\ 0.4390 & 0.6910 & 0.5150 & 1.0000 & 0.7780 \\ 0.0220 & 0.8630 & 0.1220 & 0.7780 & 1.0000 \end{bmatrix}}; \overset{\mathbf{A}}{\begin{bmatrix} 0.5808 & -0.8064 & 0.0281 \\ 0.7669 & 0.5447 & 0.3199 \\ 0.6723 & -0.7260 & 0.1144 \\ 0.9325 & 0.1040 & -0.3076 \\ 0.7912 & 0.5583 & -0.0654 \end{bmatrix}}.$$

The matrix $\mathbf{B}$, the vector $\mathbf{d}$, and the matrix $\mathbf{H}$ are

$$\overset{\mathbf{B}}{\begin{bmatrix} -0.6803 & -0.2327 \\ 0.0627 & -0.3880 \\ 0.5559 & 0.4235 \\ 0.2894 & -0.4095 \\ -0.3748 & 0.6695 \end{bmatrix}} \overset{\mathbf{d}}{\begin{bmatrix} 0.0116 \\ 0.0129 \\ 0.0078 \\ 0.0250 \\ 0.0580 \end{bmatrix}} \overset{\mathbf{H}}{\begin{bmatrix} 0.4628 & 0.3166 & 0.0541 \\ 0.0039 & -0.0486 & 0.1505 \\ 0.3090 & 0.4709 & 0.1794 \\ 0.0837 & -0.2370 & 0.1677 \\ 0.1405 & -0.5019 & 0.4483 \end{bmatrix}}.$$

The resulting $\mathbf{g} = (\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{d}$, rewritten in matrix form $\mathbf{G}$, is

$$\mathbf{g} = \begin{bmatrix} 0.0437 \\ -0.0393 \\ 0.0716 \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} 0.0437 & -0.0393 \\ -0.0393 & 0.0716 \end{bmatrix}.$$

The resulting $\mathbf{BGB}'$ is

$$\begin{bmatrix} 0.0116 & -0.0052 & -0.0071 & -0.0101 & 0.0145 \\ -0.0052 & 0.0129 & -0.0028 & 0.0176 & -0.0270 \\ -0.0071 & -0.0028 & 0.0078 & -0.0013 & 0.0028 \\ -0.0101 & 0.0176 & -0.0013 & 0.0250 & -0.0380 \\ 0.0145 & -0.0270 & 0.0028 & -0.0380 & 0.0580 \end{bmatrix},$$

and $\mathbf{AA}' + \mathbf{BGB}'$ now equals $\mathbf{R}$. This demonstrates perfect retrieval of $\mathbf{R}$ above the Ledermann bound. Now suppose that only the first two columns of $\mathbf{A}$ are used. Then r is below the Ledermann bound, and the feasible set contains more than one element. An alternative solution, $\mathbf{R}^*$ for instance, is

$$\begin{bmatrix} 1.0000 & & & & \\ 0.0293 & 1.0000 & & \text{(symmetric)} & \\ 0.9720 & 0.1330 & 1.0000 & & \\ 0.4391 & 0.6941 & 0.5149 & 1.0000 & \\ 0.0032 & 0.8630 & 0.1425 & 0.7750 & 1.0000 \end{bmatrix},$$

which differs slightly from $\mathbf{R}$, but does have the same loadings on the first two principal components. Its eigenvalues are 2.8732, 1.7965, .2079, .1073 and .0152. The eigenvalues of $\mathbf{R}$ are 2.8732, 1.7965, .2151, .0994 and .0159. This alternative solution $\mathbf{R}^*$ has been found by solving the equation $|\mathbf{G}_1 + y_1\mathbf{N}_1| = |\mathbf{\Lambda}|$. It can be verified that the last three eigenvalues of $\mathbf{R}^*$ have the same sum and the same product as those of $\mathbf{R}$.

References

Bekker, P.A., & ten Berge, J.M.F. (1997). Generic global identification in factor analysis. *Linear Algebra & Applications, 264*, 255–263.

Carroll, J.B. (1993). *Human cognitive abilities, A survey of factor-analytic studies*. New York: Cambridge University Press.

Chen, X., & Chu, M.T. (1996). On the least squares solution of inverse eigenvalue problems. *SIAM Journal of Numerical Analysis, 33*, 2417–2430.

Chu, M.T. (1998). Inverse eigenvalue problems. *SIAM Review, 40*, 1–39.

Friedland, S. (1977). Inverse eigenvalue problems. *Linear Algebra & Applications, 17*, 15–51.

Friedland, S., Nocedal, J., & Overton, M.L. (1987). The formulation and analysis of numerical methods for inverse eigenvalue problems. *SIAM Journal of Numerical Analysis, 24*, 634–667.

Guttman, L. (1958). To what extent can communalities reduce rank? *Psychometrika, 23*, 297–308.

Harman, H.H. (1967). *Modern factor analysis* (2nd. ed.). Chicago: The University of Chicago Press.

Ledermann, W. (1937). On the rank of the reduced correlation matrix in multiple-factor analysis. *Psychometrika, 2*, 85–93.

Shapiro, A. (1982). Rank-reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis. *Psychometrika, 47*, 187–199.

Shapiro, A. (1985). Identifiability of factor analysis: Some results and some open problems. *Linear Algebra & Applications, 70*, 1–7.

Wilson, E.B., & Worcester, J. (1939). The resolution of six tests into three general factors. *Proc. Nat. Acad. Sci. USA, 25*, 73–77.

Manuscript received 12 DEC 1996
Final version received 21 AUG 1998