Appendix D

More on PCA and factor analysis

D.1 Introduction

This chapter provides an account of some of the mathematics that underlies the derivation of principal components, and factors in factor analysis. It providea a source of reference for Chapter 7 and 8 without burdening those chapters with too much algebraic detail. There are fundamental differences between factor analysis and PCA that can be presented in starker relief than is possible using purely verbal descriptions, among them the dependence of factor analysis on a model for the data, and the role that rotation plays in applications of the methods. Rotation is central to factor analysis and an option in PCA, not exercised that much except, perhaps, in papers that confuse the two methodologies (Section 8.4).

A reservation often expressed about the use of factor analysis is what might be called the 'unavoidable indeterminacy' of factor analysis solutions. The model that forms the basis of factor analysis is expressed in terms of unknown parameters that must be estimated to obtain a 'solution'. A variety of choices need to be made to obtain a specific solution, among them distributional assumptions about the nature of random variation in the model, the method of factor extraction adopted, the number of factors selected for subsequent rotation, and the choice of method of rotation itself. What has been published in archaeological applications is often a matter of convention and convenience (e.g., what has gone before, often dictated by defaults in software packages). Earlier applications were dominated by PCA with varimax rotation, wrongly regarded as 'factor analysis' (Section 8.4). It is not an accident that these were the default options in widely-used commercial statistical software packages.

Any model-based method of statistical analysis involves similar kinds of choices (e.g., distributional assumptions, method of estimation) but the outcome of a factor analysis is perhaps more subject to the choices made than other methods. By contrast, PCA depends on the method of data pre-treatment chosen (Section 7.2) but, given this and as usually employed, the results from a PCA are obtained by mathematical means, leading to a unique 'solution' that does not involve assumptions about random variation, the need for estimation and so on. As emphasized in comparative reviews such as the statistical texts of Krzanowski (1985), Everitt and Dunn (2001) and Jolliffe (2002), the methods have different aims and should not be viewed as 'competitors'.

The issue of the indeterminacy of factor analysis solutions has generated a considerable literature. The choices made may sometimes not matter much, though this an empirical matter best resolved on a case-by-case basis. Some scholars are not troubled by such indeterminacy, but sceptics express concern that it allows unwarranted latitude in selecting outcomes that conform with 'theoretical' preconceptions about the phenomenon studied.

Whatever the view taken, it is a fact that any data set can be subjected to a large number of different specific factor analyses. The `fa` function from the `psych` package, used in the examples of Section 8.3.2, has the option of six different methods of factor extraction and 15 methods of rotation – eight orthogonal and seven oblique (Section D.3.2) – so 90 in all. Some of these can be expected to produce similar results, but scope for variation exists. The illustrative examples of Section 8.3.2 show some of the variation that can occur. Other (non-archaeological) illustrations are provided by Jolliffe (2002: 161–65). For convenience of reference Section D.4 summarizes some of the possible sources of indeterminacy in factor analysis applications; these are referred to in Section 8.3.2 without much additional discussion.

D.2 The singular value decomposition

For p variables, $(Y_1, Y_2, \ldots, Y_p)$, define p principal components, $(Z_1, Z_2, \ldots, Z_p)$, where component j is

$$Z_j = \mathbf{a}'\mathbf{y} = a_{j1}Y_1 + a_{j2}Y_2 + \ldots + a_{jp}Y_p$$

and $\mathbf{a} = (a_{j1}\ a_{j2}\ \ldots\ a_{jp})$ is a $(p \times p)$ column vector with transpose $\mathbf{a}'$, a $(p \times 1)$ column vector, and $\mathbf{y} = (Y_1\ Y_2\ \ldots\ Y_p)$ a $p \times 1$ column vector. Define $\mathbf{Y}$ and $\mathbf{Z}$ as $n \times p$ matrices of the data and component score with typical elements y_{ij} and z_{ij} with $\mathbf{A}$ the $p \times p$ matrix of coefficients with typical element a_{ij}; then

$$\mathbf{Z} = \mathbf{YA}'. \tag{D.1}$$

The data matrix $\mathbf{Y}$ can be factorized, using the *singular value decomposition* (SVD), as

$$\mathbf{Y} = \mathbf{UDV}' \tag{D.2}$$

where $\mathbf{U}$ is $n \times p$, $\mathbf{V}$ is $p \times p$ and $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}$ the $p \times p$ identity matrix. The matrix $\mathbf{D}$ is diagonal; the diagonal elements are the *singular values*, σ_i, and their squares, $\lambda_i = \sigma_i^2$ are *eigenvalues*. The matrix with diagonal elements given by the eigenvalues is $\mathbf{\Lambda}$.

Let $y_{ij} = (x_{ij} - \bar{x}_j)/(n-1)^{-1/2}$ (i.e. it is centered, but not standardized, and rescaled for convenience). Then

$$\mathbf{Y}'\mathbf{Y} = \mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}' \tag{D.3}$$

where $\mathbf{S}$ is the covariance matrix of the data (Appendix C)[1]. It follows from equation (D.3), on post-multiplication by $\mathbf{V}$, that

$$\mathbf{S}\mathbf{V} = \mathbf{V}\mathbf{\Lambda}.$$

By definition the columns of $\mathbf{V}$ are the *eigenvectors* of $\mathbf{S}$, with the diagonal elements of $\mathbf{\Lambda}$ the associated *eigenvalues.*

From equations (D.1) and (D.2) $\mathbf{Y} = \mathbf{Z}\mathbf{A}' = \mathbf{U}\mathbf{D}\mathbf{V}'$ define $\mathbf{Z} = \mathbf{U}\mathbf{D}$ and $\mathbf{V} = \mathbf{A}$. Thus column j of $\mathbf{V} = \mathbf{A}$ is the jth eigenvector of the estimated covariance matrix $\mathbf{S}$ and λ_j is the jth eigenvalue. Furthermore, if $\mathbf{S}_Z$ is the covariance matrix of $\mathbf{Z}$,

$$\mathbf{S}_Z = \mathbf{Z}'\mathbf{Z} = \mathbf{D}\mathbf{U}'\mathbf{U}\mathbf{D} = \mathbf{D}^2 = \mathbf{\Lambda}.$$

The sum of the diagonal elements of $\mathbf{\Lambda}$ (the eigenvalues) is the total variance of the variables defined by $\mathbf{Z}$. It can be shown that this is the same as the sum of the variances of $\mathbf{Y}$ – that is, the sum of the diagonal elements of $\mathbf{S}$ (Baxter, 2003: 68). This shows that the linear combinations, Z_j, are uncorrelated (because $\mathbf{\Lambda}$ is diagonal); that the variances of the Z_j are the eigenvalues of $\mathbf{S}$; and that (by arrangement) the Z_j are ordered in terms of importance as measured by the variances. The variances of the Z_j 'redistribute' the variances of the original data.

These are the properties required of principal components[2]. Numerical values for the a_{ij}, component variances and so on can be obtained by extracting eigenvectors and eigenvalues via the SVD. The necessary computations are applied in `prcomp` and other functions in `R`.

D.3 The factor analysis model

D.3.1 The model

The fundamental difference between factor analysis and PCA in that a statistical model needs to be formulated for the data in factor analysis whereas PCA as

[1] Note: $\mathbf{Y}'\mathbf{Y} = \mathbf{V}\mathbf{D}'\mathbf{U}'\mathbf{U}\mathbf{D}\mathbf{V}' = \mathbf{V}\mathbf{D}^2\mathbf{V}' = \mathbf{V}\mathbf{\Lambda}\mathbf{V}'$ from previous results/definitions.

[2] This development uses the covariance matrix (i.e. unstandardized data). For standardized data the covariance matrix is the correlation matrix $\mathbf{R}$, and defining $y_{ij} = (x_{ij} - \bar{x}_j)/s_j(n-1)^{-1/2}$ so that $\mathbf{S} = \mathbf{R}$ does not affect the development.

usually applied implies no such model (Jolliffe, 2002: 151)). From equation (D.1), $\mathbf{Z} = \mathbf{Y}\mathbf{A}'$; from the SVD of equation (D.2) $\mathbf{V}'\mathbf{V} = \mathbf{I}$; and following from these $\mathbf{V} = \mathbf{A}$ so $\mathbf{A}'\mathbf{A} = \mathbf{I}$. Thus, on post-multiplying both sides of the expression for $\mathbf{Z}$ by $\mathbf{A}$

$$\mathbf{Y} = \mathbf{Z}\mathbf{A} \tag{D.4}$$

or

$$Y_j = a_{1j}Z_1 + a_{2j}Z_2 + \ldots + a_{pj}Z_p.$$

This might be thought of as a 'model' for the data, but in fact is simply a mathematical consequence of the way that components are defined as linear function of the variables, showing that the latter can be expressed as linear functions of the components. This does not involve distributional assumptions of the kind typically associated with models.

In contrast, factor analysis involves a model

$$Y_j = b_{1j}F_1 + b_{2j}F_2 + \ldots + b_{qj}F_q + \varepsilon_j$$

where ε_j is an error term with variance ψ_j. The matrix formulation for this is

$$\mathbf{Y} = \mathbf{F}\mathbf{B} + \boldsymbol{\varepsilon}. \tag{D.5}$$

There are important differences between this and the PCA formulation of equation (D.4).

1. Unlike PCA, factor analysis involves a statistical model for the data, expressed as the sum of a systematic and random component.

2. In equation (D.4) $\mathbf{F}$ is an $n \times q$ matrix of unobserved factor scores where $q < p$. If p is large q will typically be a lot smaller. The F_j are *common factors*. In PCA the number of components, p, is known; in factor analysis q is not, and determining or confirming a suitable value is an aspect of the analysis.

3. In equation (D.4) $\mathbf{B}$ is a $q \times p$ matrix of parameters that must be estimated. Determining coefficients in the PCs is a mathematical exercise.

4. Assumptions are needed for estimation; minimally, that errors are uncorrelated and common factors are uncorrelated with the errors and each other. This last assumption can be relaxed, giving rise to *oblique* factors.

D.3.2 Rotation

There is another additional and important difference between the two methods. As seen from equation (D.3) the PCs and their variances can be determined by finding the eigenvectors and eigenvalues of the covariance matrix of the data, $\mathbf{Y}'\mathbf{Y} = \mathbf{S}$. Subject to the constraint imposed on the eigenvectors and the properties required of the PCs this leads to a unique solution.

Given the assumptions of factor analysis the covariance matrix can be written as

$$\mathbf{S} = \mathbf{B}'\mathbf{B} + \mathbf{\Psi} \tag{D.6}$$

where $\mathbf{\Psi}$ is a diagonal matrix with diagonal elements ψ_i. If $\mathbf{T}$ is a $p \times p$ orthogonal matrix (i.e. $\mathbf{T}'\mathbf{T} = \mathbf{I}$) then, defining $\tilde{\mathbf{B}} = \mathbf{TB}$ it follows that

$$\tilde{\mathbf{B}}'\mathbf{T}'\mathbf{T}\tilde{\mathbf{B}} = \mathbf{B}'\mathbf{B}$$

and $\mathbf{TB}$ is as valid a solution to equation (D.6) as $\mathbf{B}$. This solution is called an *orthogonal rotation* and it can be obtained in innumerable ways.

To obtain specific solutions for the factor loadings some choice of $\mathbf{T}$ is needed which requires the imposition of constraints on the factor loadings. Almost invariably 'simple structure' is aimed for, ideally resulting in factor loadings that are either 'high' or close to zero. Ideally a variable will have a high loading with respect to only one factor. Factor analysis, and ideas of rotation are driven by the desire for 'interpretability'; the assumption is that the covariances/correlations between observable variables reflect their relationship with a smaller number of common factors (or latent variables) or constructs that can be assigned some sort of (theoretical) 'meaning' within the domain of study involved.

Matters are complicated by the view that if factors do represent some aspect of an unobservable 'reality' there is no particular reason to expect them to be uncorrelated (Cattell, 1978, 128; cited in Jolliffe, 2002: 152). That is, orthogonal rotation does not lead to the identification 'of correct factor structure'. To allow for this, methods of *oblique rotation* have been developed where the constraint that $\mathbf{T}'\mathbf{T} = \mathbf{I}$ is dropped, leading to the identification of correlated factors[3].

An obvious question is 'how should the choice of rotation method be made?'. There is no prescriptive answer to this question

D.3.3 Factor extraction

No attempt is made here to provide a comprehensive mathematical account of methods of factor extraction, which necessarily precede rotation. The references

[3] The notion of 'correct' factor structure, explicit here, leads to the thought there is no obvious logical reason why 'correct' structure should also be 'simple' – that is, just because a factor is 'interpretable', and why simple structure is sought, doesn't make it 'real' or 'correct'.

given at the end of the chapter may be pursued for technical details. Binford and Binford's (1966) seminal paper that popularized the use of 'factor analysis' in archaeology was, in fact, PCA with rotation, as was the majority of 'factor analysis' applications to the mid-1980s when usage began to decline. It is a commonplace observation in the statistical literature that treating unrotated PCA as factor analysis is wrong. The same is true of PCA with rotation since it takes no account of the error structure which is one of the distinguishing features of the factor analysis model (Section 8.4).

It is, however, possible to use principal component ideas applied to the 'reduced covariance matrix' found from equation (D.6).

$$\mathbf{B}'\mathbf{B} = \mathbf{S} - \mathbf{\Psi}$$

as a starting point in iterative methods of estimation. This requires an estimate of $\mathbf{\Psi}$, leading to many specific varieties of factor analysis, none of which have any great claim to 'absolute validity' (Jolliffe, 2002: 159). This method of *principal axis* factor analysis has been one of the most commonly used methods of factor extraction when not confused with PCA.

Extraction methods do not necessarily require distributional assumptions to be made about the error terms, other than that they are independent with zero mean. The statistically more 'thoroughgoing' method of maximum-likelihood does, however, require the strong assumption that the errors have a multivariate normal distribution. The method has the advantage, unlike PCA and unlike other methods of factor analysis, that results do not depend on data standardization; it also facilitates the use of inferential methods to assess the quality of results obtained. The 'downside' is that the multivariate normality assumption might often be considered to be unrealistic, though the estimates have some reasonable properties even when normality does not hold[4].

Jolliffe (2002: 156–57) draws an analogy with least squares regression which can be applied regardless of distributional assumptions, but produces maximum-likelihood estimates and inherits their optimality properties if normality (of the errors) can be assumed. A variety of least squares methods exist for parameter

[4] One quite often reads the assertion that factor analysis and PCA require normality assumptions. Other than MLE this is not a requirement in most applications. There is sometimes confusion between normality of the error terms and 'normality' of the variables. As far as the former is concerned the usual PCA formulation does not even involve an error term. As with the commoner applications of regression analysis, where the same erroneous statement is sometimes found, there is no requirement of a normal distribution for the variables. What is the case is that if the observed distribution of variables is 'badly-behaved', containing clear outliers for example, this may unduly affect results, and the problem(s) need to be identified and remedial action taken. This is a matter of practical data analysis rather than the imposition of unnecessary distributional requirements. Variables can be perfectly 'well-behaved' for the purposes of data analysis without approaching a normal distribution.

estimation in factor analysis, several of which are available in the `fa` function in the `psych` package for `R`. The default method of factor extraction in `fa`, `minres` – 'minimum residual' using Ordinary Least Squares estimation – is noted in the documentation to produce 'solutions very similar to maximum-likelihood even for badly behaved matrices', and weighted and generalized least squares (`wls` and `gls`) methods are also available. The documentation also states that maximum-likelihood 'is probably preferred' provided the data are well-behaved.

The default method of rotation in `fa`, `oblimin`, is oblique, replacing the orthogonal method, `varimax`, that was the default in earlier versions. Varimax is the most widely used method in published applications, probably for historical reasons (Jolliffe, 2002: 153–54; Section 8.4).

D.4 Discussion

Here and in Chapter 8 it has been emphasized that PCA and factor analysis are different methods of analysis that should be clearly distinguished. The distinction has not always troubled practitioners and comment to the effect that the methods will often 'produce similar results' is not uncommon (see Section 8.4).

The choices leading to indeterminacy in factor analysis are mostly not relevant in PCA and may produce substantively different interpretations. What are the important choices?

1. *Factor extraction* can make a difference. It might be expected that those methods that require prior estimation of $\mathbf{\Psi}$ will, if the factor model is reasonable and a sensible method of estimation used, produce fairly similar results (Jolliffe, 2002: 159). If, as the `fa` documentation claims, least-squares methods approximate maximum-likelihood it may not be worth troubling too much about which is used unless the latter is wanted for inferential purposes. The example in Section 8.3.2 contrasts methods towards either end of the 'spectrum' in terms of the assumptions needed – principal factor analysis and maximum-likelihood – to see if the choice makes much difference.

2. *Rotation* of factors can be undertaken in many different ways. Jolliffe (2002: 271) suggests, within the class of orthogonal rotations, the choice 'often makes little difference to the results'. The choice between orthogonal and oblique rotation presumably matters; there would otherwise be no reason for developing the latter. Section 8.3.2 provides an example comparing varimax and oblimin rotations.

3. The *number of factors to rotate* is a choice that Jolliffe suggests may be more important than the factor extraction and rotation methods used. Kaiser's

rule and variants of it is a commonly used factor selection method. Where the choice of factors to rotate is not obvious experimenting with different choices, to see if interpretation is much affected, seems sensible.

Interrogating data using different methods and different variants of a method can lay one open to the accusation of being 'unscientific' and 'fishing' for palatable interpretations compatible with initial preconceptions. This *is* a danger. An implication – this is something of a caricature, but not entirely so – is that the proper approach is to make a principled choice of method and live with the consequences. This is a counsel of perfection open to the possible counter-accusation that such a 'principled choice' may itself conceal a variety of strongly held preconceptions and methodological/philosophical views.

The view adopted in these notes is that the exploration of different methods of data analysis is legitimate and sensible. It is in the selective reporting only of those results palatable to the investigator that the danger lies. The dictates of publication doubtless produce pressure to concentrate only on 'significant' results, but it would be helpful for potential consumers of a method to be made aware of circumstances when they don't 'work' and why.

Early developments of factor analysis occurred in the psychometric literature (e.g., Spearman, 1900). Statistician's came late the subject; the date and title of the first edition of Lawley and Maxwell's (1963) text, *Factor Analysis as a Statistical Method*, with the emphasis on *as a Statistical Method*, testifies to this. The development of efficient computational software during and either side of the 1960s promoted the wider use of factor analysis, archaeology not excluded. The popular SPSS package first surfaced in 1968; from some perspectives this popularity, and its longevity, has not necessarily been beneficial. The treatment of PCA as a particular case of factor analysis has engendered confusion and is to be regretted.

General statistical texts on multivariate analysis begin to appear in any profusion round about this period. Anderson's (1958) early text was a theoretical treatment and other texts such as Morrison (1967) appear in the 1960s. A 'flurry' of books appeared in the late-1970s to mid-1980s, among them Mardia *et al.* (1979), Chatfield and Collins (1980), Seber (1984) and Krzanowski (1988). The first edition of Jolliffe (2002), with its more specialized focus on PCA, appeared in 1986. Other texts noted elsewhere in these notes that include comparisons of PCA and factor analysis are Krzanowski and Marriott (1995) and Everitt and Dunn (2001). None of these can be described as 'recent' but the underlying mathematics and ideas don't change. What has changed is the ease and flexibility of implementation in packages such as R. Statistical developments, of more complex and computer intensive methods with essentially similar aims, have taken place but not yet influenced archaeological practice much. The framework that supports most archaeological applications of PCA, factor analysis and other multivariate methods was in place well before the turn of the century and remains valid.

Chatfield and Collins' (1980) take on factor analysis is, as already noted, a negative one (Section 8.4). Other statistical treatments have been more even-handed. Jolliffe's (2002) emphasis that factor analysis and PCA are different rather than competing methods, and that factor analysis, if appropriate to an analysis and properly executed, has its place as part of the analyst's 'toolkit', is echoed in similar statistical texts.

The questions for archaeologists are whether factor analysis is an appropriate tool for much of what they want to do and, if so, has its value in application been convincingly demonstrated? The first question is an interesting one, and for archaeologists rather than statisticians to address. My own view on the second question, nearly 50 years on from the publication of Binford and Binford (1966), it that is difficult to answer in the affirmative.