Central European University

Department of Mathematics and its Applications

Master's Thesis

Factor models

Submitted by:

Kurdyukova Anna

Supervisor:

Dr. Marianna Bolla

Head of Program:

Dr. Gheorghe Morosanu

Budapest - 2010

Contents

List of Figures

List of Tables

1 Introduction

A central aim of factor analysis is the "orderly simplification" of a number of interrelated measures. It should be reassuring for the reader to discover that factor analysis seeks to do precisely what man has been engaged in throughout history – to make order out of the apparent chaos of his environment. This process of identifying and classifying the attributes of our surroundings in an attempt to make our world intelligible is a very familiar one.

One of the chief burdens of our research is to define a *factor*. However, the term crops up so frequently beforehand that a provisional definition would not be out of place. Intuitively, when a group of variables has, for some reason, a great deal in common a factor may be said to exist. These related variables are discovered using the technique of correlation. For example, if one took a group of people and correlated the lengths of their arms, legs, and bodies one would probably find a marked relationship between all three measures. This interconnection constitutes a factor.

Many authors (Child(1970), Kline (1994)) agree on that it was Galton, a brilliant scientist of the 19th and early 20th centuries, who laid the foundation of factorial study. Although he did not concern himself with the kind of mathematical analysis so familiar to the subject nowadays, he nevertheless inspired two lines of thought which have been essential to the development of factorial study. The first of these was the idea that general intellectual power was spread in a continuous fashion from the very dull to the very bright. This idea of a common casual thread running through all intelligent behavior was in marked contrast to the pluralistic theories expounded by faculty theorists. His idea lived on in a form of the general ability formulated, as we shall discover, using factor solutions. He denoted this ability as g. Galton took his argument a step further by proposing the existence of special powers, although he still believed the general intellectual power was the overriding influence in determining the quality of a man's responses in general. The second major contribution made by Galton was the concept of correlation. He developed quantitative methods to give some idea of the interdependence between two variables. This point is worth pursuing because we shall be making extensive use of the concept of correlation.

Karl Pearson, in a famous paper at the beginning of the 20th century, was the first to

make explicit a procedure for a factor analysis, and he derived his formulae by considering the geometry of multidimensional space. The earliest suggestion of an application for this new technique came in 1902 when Macdonnell wrote a paper on the study of "criminal anthropometry and the identification of criminals" (Child (1970)). And finally, it was Spearman's report in 1904 on 'general intelligence' which heralded the intensive study of human ability using mathematical models. In this paper he posited the well-known Two-Factor Theory. Of the immediate relevance to factor analysis, he states what he calls "our general theorem" which is

"Whatever branches of intellectual activity are at all dissimilar, then their correlations with one another appear wholly due to their being all variants wholly saturated with some common fundamental Function (or group of Functions)..."

He distinguished this central Function from "the specific function (which) seems in every instance new and wholly different from that in all the others". The simplest way to justify this analysis is to appeal to the theory of partial correlation. The essence of what Spearman needed is contained in the formula for the partial correlation between two variables, i and j say, given a third variable which following Spearman we call G. Thus

$$r_{i,j|G} = \frac{r_{ij} - r_{iG}r_{jG}}{\sqrt{(1 - r_{iG}^2)(1 - r_{jG}^2)}}. \tag{1}$$

If the correlation between i and j is wholly explained by their common dependence on G then $r_{ij|G}$ must be zero, or $r_{ij} = r_{iG}r_{jG}$, $(i, j = 1, 2, \ldots, p)$. So, if the correlation matrix $R = \{r_{ij}\}$ can be represented in this way then we have evidence for an underlying common factor. Thus, it was deducted (Bartholomew (2007)) that the first face of factor analysis thus starts from the correlation matrix. At this basic level, we see that the correlation matrix is the key to unearthing a common factor. In case of multivariate normal distribution - that is very common due to the multivariate Central Limit Theorem - the correlation, or covariance structure uniquely determines the interconnections between variables.

Elaborate and refined psychological and mathematical arguments blossomed from these early efforts of Galton, Pearson and Spearman. The range of subject areas in which factor analysis has played an important role is now very extensive. Many applications now exist, including politics, sociology, economics, man-machine systems, accident research, taxonomy, biology, medicine and geology. But such an extensive use of factor

analysis is closely connected with the second face of the history of its development. The second face of factor analysis started not with a correlation matrix, but with a model, and it makes possible to investigate statistical hypotheses on the model fit. The use of models in statistics, on a regular basis, seems to date from the 1950s. A statistical model is a statement about the distribution of a set of random variables. A simple linear regression model, for example, says that the dependent variable y is normally distributed with mean $a + bx$ and variance σ^2, where a and b are unknown constants and x is an observable variable. In the case of factor analysis the move to a model-based approach was gradual. The rudimentary idea was contained in the idea that an observed test score was composed of a common part and a specific part. However, it was in Lawley and Maxwell (1963) that a linear model was made the starting point for developing the theory in a systematic way. Actually, this formulation was incomplete but, in its essentials, it still holds sway today. In modern notation, Lawley and Maxwell (1963) supposed that

$$x_i = \lambda_{i1}y_1 + \lambda_{i2}y_2 + \ldots + \lambda_{iq}y_q + e_i, \qquad (i = 1, 2, \ldots, p), \tag{2}$$

where $q < p$ is the number of factors. In this equation the λs are constants and the xs, ys and es are random variables. Thus if one imagines that the y-values for item i are drawn at random from some distribution and the es are drawn similarly, then the model postulates that, if they are combined according to the equation above, the resulting random variable will be x_i. It is usually assumed that the ys are independent with normal distribution and (without loss of generality) unit variances; e_i is assumed to be normal with variance ψ_i; also e_is are usually independent of each other and of ys. With a few more assumptions we have the standard linear normal factor model in use today.

It is worth pausing to ask the justification for regarding this as another way of looking at Spearman's factor analysis. The answer (Bartholomew (2007)) is that the correlation structure to which it leads is the same. But the major advantage of the model-based approach is that it enables us to provide rigorous methods for answering the traditional questions addressed by factor analysis.

It is a curious fact of statistical history that there has been a strong focus on methods for continuous data. Regression and correlation analysis and then the analysis of variance have, for the most part, pre-supposed that the variables involved were continuous. Other multivariate methods, introduced along the way, such as discriminant analysis, principal components analysis and canonical correlation fall into the same mould. Of course, these

methods have been widely used on data which were not continuous. Coarsely grouped variables, ordered categorical variables, even binary variables, have been grist to the analysts' mill. Indeed, much ingenuity has been exercised to treat categorical data as if it were continuous by introducing, for example, pseudo-correlation coefficients of one sort or another. In practice, and especially in the social sciences, much of the data we encounter is not continuous but categorical. But in factor analysis we are asking whether the dependencies among a set of variables can be explained by their common dependence on one, or more, unobserved latent variables (or factors). There is nothing in this statement which refers to the level of measurement of the variables involved. If, therefore, we formulate the problem in sufficiently general terms we should have a general enough framework to include variables of all sorts. There exists a certain type of factor analysis of categorical data, which is called 'correspondence analysis'. The essential elements of the problem are the inter-dependence of a set of observable variables and the notion of conditional independence. Suppose we have p observable random variables $x' = (x_1, x_2, \ldots, x_p)$ with joint probability distribution $f(x)$. The question is: Do there exist factors $y_1, y_2, \ldots, y_q$, where $q < p$ such that the xs are conditionally independent? Lazarsfeld (1968) was the pioneer of this distinct kind of 'factor analysis' and he called it latent structure analysis. Essentially, he allowed one or both of the sets of variables x and y to be categorical.

In the 1950s, there seemed to be two schools of factor analysis: the psychometric school and the statistical school. The psychometric school regarded the battery of tests as a selection from a large domain of tests that could be developed for the same psychological phenomenon and focused on the factors in this domain. By contrast, the statistical school regarded the number of tests as fixed and focused on the inference from the individuals being tested to a hypothetical population of individuals. The distinction between the two perspectives is particularly contrasted with regard to the number of factors. In the psychometric perspective, it was assumed that there are a small number of major factors and possibly a large number of minor factors; whereas in the statistical perspective, the number of factors is assumed to be small relative to the number of tests.

Whereas the factor analysis literature in the first half of the 20th century was dominated by psychologists, the literature of the second half of the century was dominated by statisticians. In fact, there has been an enormous development of the statistical method-

ology for factor analysis in the last 50 years. This has been accompanied by an equally enormous development of computational methods for factor analysis. During this period, the applications of factor analysis spread from psychology to many other disciplines, for example, international relations, economics, sociology, communications, taxonomy, biology, physiology, medicine, geology and meteorology.

Whereas the initial approach to factor analysis was oriented to data originating from independent, identically distributed random variables and consisted in dimension reduction in the cross-sectional dimensions (i.e. the number of variables), the idea was further generalized to modelling of multivariate time series, thus compressing information in the cross-sectional and time dimensions. The idea has been pursued rather independently in a number of areas, such as signal processing or econometrics. In Dynamic factor models (DFMs; Geweke (1977)), the comovements of the observable time series are characterized by latent dynamic factors. Over the past decade, work on DFMs has focused on high-dimensional systems in which very many series depend on a handful of factors (Forni et al. (2000), Stock and Watson (2002), and many others).

In factor analysis, most authors (Child (1970), DeCoster (1998), Johnson and Wichern (2002), Kline (1994)) address the problem of reducing the dimension of a multivariate random variable, and we want to fix, from the start, the number of factors. Each factor will then be interpreted as a latent characteristic of the individuals revealed by the original variables. In a survey on household consumption, for example, the consumption levels, X, of p different goods during one month could be observed. The variations and covariations of the p components of X throughout the survey might in fact be explained by two or three main social behavior factors of the household. For instance, a basic desire of comfort or the willingness to achieve a certain social level or other social latent concepts might explain most of the consumption behavior. These unobserved factors are much more interesting to the social scientist than the observed quantitative measures (X) themselves, because they give a better understanding of the behavior of households.

From a statistical point of view, the essential purpose of factor analysis is to describe, if possible, the covariance relationships among many variables in terms of a few underlying, but unobservable, random quantities called factors. The ultimate goal is to find underlying reasons that explain the data variation. In achieving this goal we need to check the relation of the factors and original variables and give them an interpretation in the framework of

how the data were generated.

Dynamic factor models (Breitung and Eickmeier (2005), Deister and Zinner (2007), Doz and Lenglart (2001), Forni et al. (2000), Geweke (1977)) decompose the dynamics of observable variables $y_{i,t}$, $i = 1, \ldots, n$, $t = 1, \ldots, T$ into the sum of two unobservable components, one that affects all y_is, namely the factors F_t, and one that is idiosyncratic, e.g. specific to each i:

$$y_{i,t} = a_i + b_i F_t + \epsilon_{i,t}, \tag{3}$$

where a_i is a constant, and b_i is a loading of series i to the common actors. Both the factor and idiosyncratic components are usually assumed to follow autoregressive processes of order q and p_i respectively. The model just described is the standard dynamic factor model estimated for example in Stock and Watson (1989).

Among the very recent developments I would mention the Independent Component Analysis (ICA), which is also a statistical and computational technique for revealing hidden factors that underlie sets of random variables. ICA defines a generative model for the observed multivariate data, which is typically assumed to be nongaussian. As before, in the model, the data variables are assumed to be linear combinations of some unknown latent variables, and the coefficients of the system are also unknown. The latent variables are also assumed nongaussian and mutually independent, and they are called the independent components of the observed data. These independent components, also called sources or factors, can be found by ICA.

ICA is closely related to Principal component analysis and Factor analysis. However, ICA is a much more powerful technique.

Before we proceed with a detailed discussion of the subject, it is important to sound some notes of caution. There are some limitations (Child (1970)) which can be stated in general terms. One of the first and certainly one of the most important is to avoid reading too much into a correlation coefficient because causal relationships cannot be inferred from correlations alone. Also the size of the sample must enter into any speculations about errors and consequently the larger the sample the more notice we are likely to take of large correlations. The rule should be, in applying any tests of significance, to err on the side of rigour rather than leniency. Also in this connection, tests with low reliability, as we shall see later, should be avoided in factor analysis. Sample selection is important too; it often happens that a sample is homogeneous because of the special circumstances in

selection of a parent population. Samples collected from different populations should not be pooled when computing correlations. Factors which are specific to a population may become obscured when pooling is applied. Another important concern is the linearity of the correlation between two sets of data. Curved rather than straight relationships are suspect. Anything approaching a curvilinear shape should be treated with the utmost care. Besides, it is sometimes said that with factor analysis you only get what you put in so that it is difficult to see how the method can be useful.

The work is organized as follows. The first part of it is devoted to the fundamental models in the field of Factor analysis. In Section 2 we discuss the two main types of Factor analysis and describe the performance procedures. In Section 3 we provide the basic concepts we need for further research and mention the questions that must be answered in any factor analytic study. In Section 4 we discover Principal component analysis and give some connected results. Section 5 is devoted to a basic model of Factor analysis and the most popular methods of parameter estimation. All necessary steps of Factor analysis are also discussed in this section. In the second part of the work we study the applications of Factor models in economics. Section 6 provides us with basic concepts of Time series theory. The Dynamic Factor model is described in Section 7. Further, in Section 8 our particular model is given together with the extraction algorithm. Sections 9 discusses the problem of finding optima of inhomogeneous quadratic forms, which has to be solved while implementing the algorithm. We show the example of application of our model in Section 10. In Section 11 some conclusions are listed, as well as the topics of further possible research. Some figures and tables, illustrating the Example from Section 10, are shown at the end.

Part I

The Foundations of Factor Analysis

2 Classification of Factor Analyses

Typically it is accepted (DeCoster (1998), Kline (1994)), that there are two types of factor analysis: exploratory and confirmatory. Exploratory factor analysis attempts to

discover the nature of the constructs influencing a set of responses. Confirmatory factor analysis tests whether a specified set of constructs is influencing responses in a predicted way.

2.1 Exploratory factor analysis

The primary objectives of an Exploratory factor analysis are to determine the number of common factors influencing a set of measures and to evaluate the strength of the relationship between each factor and each observed measure. Some common uses of Exploratory factor analysis are to

1. Identify the nature of the constructs underlying responses in a specific content area.
2. Determine what sets of items "hang together".
3. Demonstrate the dimensionality of a measurement scale. Researchers often wish to develop scales that respond to a single characteristic.
4. Determine what features are most important when classifying a group of items.
5. Generate "factor scores" representing values of the underlying constructs for use in other analyses.

There are seven basic steps to performing an Exploratory factor analysis:

1. **Collect measurements.** You need to measure your variables on the same (or matched) experimental units. That is why correlations are preferred to covariances.
2. **Obtain the correlation matrix.** You need to obtain the correlations (or covariances) between each of your variables.
3. **Select the number of factors for inclusion.** Sometimes you have a specific hypothesis that will determine the number factors you will include, while other times you simply want your final model to account for as much of the covariance in your data with as few factors as possible. If you have k measures, then you can at most extract k factors. There are a number of methods to determine the "optimal" number of factors by examining your data, which will be discussed further in this work.

4. **Extract the initial set of factors.** You must submit your correlations or covariances into a computer program to extract your factors. This step is too complex to reasonably be done by hand. There are a number of different extraction methods, including maximum likelihood, principal component, and principal axis extraction.

5. **Rotate the factors to a final solution.** For any given set of correlations and number of factors there is actually an infinite number of ways that you can define your factors and still account for the same amount of covariance in your measures. Some of these definitions, however, are easier to interpret theoretically than others. By rotating your factors you attempt to find a factor solution that is equal to that obtained in the initial extraction but which has the simplest interpretation.

6. **Interpret the factor structure.** Each of your measures will be linearly related to each of your factors. The strength of this relationship is contained in the respective factor loading, produced by your rotation. This loading can be interpreted as a standardized regression coefficient, regressing the factor on the measures.

7. **Construct factor scores for further analysis.** If you wish to perform additional analyses using the factors as variables you will need to construct factor scores. The score for a given factor is a linear combination of all of the measures, weighted by the corresponding factor loading.

Exploratory factor analysis is often confused with Principal component analysis, a similar statistical procedure. However, there are significant differences between the two and they will provide somewhat different results when applied to the same data. The purpose of Principal component analysis is to derive a relatively small number of components that can account for the variability found in a relatively large number of measures. This procedure, called *data reduction*, is typically performed when a researcher does not want to include all of the original measures in analyses but still wants to work with the information that they contain. Thus, you should use Exploratory factor analysis when you are interested in making statements about the factors that are responsible for a set of observed responses, and you should use Principal component analysis when you are simply interested in performing data reduction.

2.2 Confirmatory factor analysis

The primary objective of a Confirmatory factor analysis is to determine the ability of a predefined factor model to fit an observed set of data. Some common uses of Confirmatory factor analysis are to

1. Establish the validity of a single factor model.
2. Compare the ability of two different models to account for the same set of data.
3. Test the significance of a specific factor loading.
4. Test the relationship between two or more factor loadings.
5. Test whether a set of factors are correlated or uncorrelated.
6. Assess the convergent and discriminant validity of a set of measures.

There are six basic steps to performing an Confirmatory factor analysis:

1. **Define the factor model.** The first thing you need to do is to precisely define the model you wish to test. This involves selecting the number of factors, and defining the nature of the loadings between the factors and the measures. These loadings can be fixed at zero, fixed at another constant value, allowed to vary freely, or be allowed to vary under specified constraints (such as being equal to another loading in the model).
2. **Collect measurements.** You need to measure your variables on the same (or matched) experimental units.
3. **Obtain the correlation matrix.** You need to obtain the correlations (or covariances) between each pair of your variables.
4. **Fit the model to the data.** You will need to choose a method to obtain the estimates of factor loadings that were free to vary. The most common model-fitting procedure is *Maximum likelihood estimation*, which should probably be used unless your measures seriously lack multivariate normality. In this case you might wish to try using *Asymptotically distribution free estimation*.

5. **Evaluate model adequacy.** When the factor model fits to the data, the factor loadings are chosen to minimize the discrepency between the correlation matrix implied by the model and the actual observed matrix. The amount of discrepency after the best parameters are chosen can be used as a measure of how consistent the model is with the data. The most commonly used test of model adequacy is the χ^2 *goodness-of-fit* test. The null hypothesis for this test is that the model adequately accounts for the data, while the alternative is that there is a significant amount of discepency. Unfortunately, this test is highly sensitive to the size of your sample, such that tests involving large samples will generally lead to a rejection of the null hypothesis, even when the factor model is appropriate.

6. **Compare with other models.** If you want to compare two models, one of which is a reduced form of the other, you can just examine the difference between their χ^2 statistics, which will also have an approximately χ^2 distribution. Almost all tests of individual factor loadings can be made as comparisons of full and reduced factor models.

Confirmatory factor analysis has strong links to structural equation modelling, a relatively nonstandard area of statistics. It is much more difficult to perform a Confirmatory factor analysis than it is to perform an Exploratory factor analysis. The first one requires a larger sample size than the second, basically because the Confirmatory factor analysis produces inferential statistics. The exact sample size necessary will vary heavily with the number of measures and factors in the model, but you can expect to require around 200 subjects for a standard model. As in Exploratory factor analysis, you should have at least three measures for each factor in your model. However, you should choose measures that are strongly associated with the factors in your model (rather than those that would be a "random sample" of potential measures).

In general, you want to use Exploratory factor analysis if you do not have strong theory about the constructs underlying responses to your measures and Confirmatory factor analysis if you do. It is reasonable to use an Exploratory factor analysis to generate a theory about the constructs underlying your measures and then follow this up with a Confirmatory factor analysis, but this must be done using separate data sets. You are merely fitting the data (and not testing theoretical constructs) if you directly put the

results of an Exploratory factor analysis directly into a Confirmatory factor analysis on the same data. An acceptable procedure is to perform an Exploratory factor analysis on one half of your data, and then test the generality of the extracted factors with a Confirmatory factor analysis on the second half of the data.

3 Basic concepts

3.1 Concept of variance

Many of the fundamental ideas in factor analysis derive from the concept of variance and the next step is to relate the above deductions to this concept. *Variance* is the square of the standard deviation. What is the total variance of a test? There are two important components of variance required to account for the total variance of a test. These are common and unique variance. When a factor contains two or more tests with significant loadings (or variance if the values are squared) it is referred to as a *common factor* and the variance of the tests in that factor is known as *common variance*. The primary aim of factor analysis is the discovery of those common factors. The techniques for extracting the factors generally endeavour to take out as much common variances as possible in the first factor. Subsequent factors are, in turn, intended to account for the maximum amount of the remaining common variance until, hopefully, no common variance remains. Common factors also come in two sizes: *general factors*, usually the first in a factor solution giving the maximum variance in the first factor, include significant loadings from most if not all the tests in the analysis, and *group factors*; group factors, as the term implies, arise when a few tests with significant loadings appear in the same factor. Often several group factors occur in the same analysis.

There remains that part of the total variance of a test resulting from the unique properties possessed by the test and as such would be entirely uncorrelated with the other tests in a particular analysis. This is referred to as unique variance. A factor containing only one significant loading for a particular test would be a unique factor. Unique variance can be broken down into two further elements of *specific* and *error variance*. Each test possesses some particular qualities which are not shared with any other test in the battery under consideration, and the variation in scores arising from these qualities will produce specific variance. Error variance, or *unreliability*, results from the imperfections of test

measurement. The difference between this and the total test variance does give a measure of the reliability of the test. In summary, we have the total variance of a test made up from common variance and unique variance, which in turn is divided into specific and error variance. As the variance is additive, the relationship can be expressed in its simplest form as

$$V_T = V_C + V_S + V_E \tag{4}$$

Strictly speaking, V_C consists of $V_{C_1} + V_{C_2} + V_{C_3} + \cdots + V_{C_q}$, where q is the total number of common factors. If the total variance of the test was made equal to one, the contributory variances on the right-hand side of the equation would become proportions of the total variance. Let these proportions be $C_1, C_2, C_3, \ldots, C_q$ for the common variance, S for the specific variances and E for the error variance. Our equation becomes

$$1 = C_1 + C_2 + C_3 + \ldots + C_q + S + E \tag{5}$$

This important fundamental statement is known as *the factor equation.* An equation of this kind applies to each individual test in a factor analysis. On taking the square roots of the variances we would obtain the loadings assigned to the factors.

The sum of all the common factor variance of a test is known as *the communality* (h^2), that is the variance shared in common with other tests. The communality will also be the sum of the squares of common factor holdings for a test.

3.2 Basic models

There are two basic models which can be adopted in factor solutions. They are generally (Child (1970), DeCoster (1998), Johnson and Wichern (2002), Kline (1994)) known as *the factor analysis* and *the component analysis* models. Without becoming technical, the distinction is that in factor analysis some account is taken of the presence of unique variance whereas in component analysis the intrusion of unique variance is ignored. The discrimination between these models rests entirely on the assumption one makes about the portions of the unit variances of each test which are to appear in the common factors. This is determined by the figure placed in the diagonal of the correlation matrix, because this diagonal value defines the total common variance (communality) of a test to be distributed amongst the common factors. If we assume the leading diagonal values to be unity, we are saying, in effect, that the test is completely reliable and without error.

This is the component analysis model. If we had some knowledge of the common variance of a test before commencing a factor solution and inserted this communality in the appropriate leading diagonal for the test we would automatically build into the model an allowance for unique variance. It is all the variance which remains after our predetermined communality has been accounted for $(1 - h^2)$. This is the factor analysis model. One of the dilemmas in this form of solution was finding an efficient and accurate procedure for determining the value of the communality before the analysis began. One of the earliest methods which is still in use is to take the largest correlation coefficient for a row of a correlation matrix and insert the value in the leading diagonal for that row.

There is a certain distinction between components and factors. Components are real factors as they are derived directly from the correlation matrix. Common factors of factor analysis are hypothetical because they are estimated from the data. With principal component analysis it is possible to take out as many components as variables, thus exhausting all the variance in the matrix. However, since one of the aims of exploratory factor analysis is to explain the matrix of correlations with as few factors as is possible, it is usual to take out less than this number.

3.3 Criteria for the number of factors to be extracted

How do we decide on how many factors to extract? Essentially, only the common factors are required and the method employed rest upon assumption as to when this has been achieved. The following two methods are most popular among factorists.

1. A technique in considerable use at present is *Kaiser's criterion* suggested by Guttman and adapted by Kaiser (Child (1970)). The rule is that only the factors having latent roots greater than 1 are considered as common factors. This method is particularly suitable for principal components designs. Kaiser's criterion is probably most reliable when the number of variables is between 20 and 50. When the number of variables is less than 20, there is a tendency, not too serious, for this method to extract a conservative number of factors. When more than 50 variables are involved, too many factors are taken out.

2. It should be mentioned that component analysis has the drawback of containing 'hybrid' factors, particularly in the later factors to be extracted, because unique

variance overlaps with common variance. Catelli has argued that some unique variance creeps into all factors to be extracted is so great as to swamp the common variance. We need to identify the optimum number of factors which can be taken out before the intrusion of non-common variance becomes serious. An intriguing method described by Catelli is the *scree test*. For this, a graph is plotted of latent roots against the factor number and the shape of the resulting curve employed to judge the cut-off point. Starting at the highest latent root, the plot is curved at first then develops into a linear relationship at some point. This point at which the curve straightens out is taken as the maximum number to be extracted.

3.4 Criteria for the significance of factor loadings

Ultimately we have to decide on which factor loadings are worth considering when it comes to interpreting the factors. Several methods have been suggested, of which three will be mentioned.

1. The first suggestion is not really based on any mathematical propositions, except that it represents roughly 10 per cent of the variance. Some idea of the pattern of significant items can be gained by underlining the loadings greater than 0.3 in absolute value provided the sample is not too small ($N = 50$ at least).

2. In deriving the factor loadings it became evident that they were, in effect, correlation coefficients. For the purpose of specifying an acceptable level of significance the loadings could be treated in a similar fashion to correlation coefficients. For a certain sample size one can compare the loadings with table values.

3. One distinct disadvantage in the last method is the absence of any adjustment for the number of variables, or the factors under consideration. Burt and Banks have shown that as one progresses from the first factor to higher factors the acceptable value for a loading to be judged significant should increase (it should get harder for coefficient to reach significance). These authors devised a formula (the Burt-Banks formula) which has the merit of allowing not only for the sample size but also for the number of tests correlated and number of factors up to and including the one

under examination. The formula is

$$LoadingSE = CorrelationSE \cdot \left(\sqrt{\frac{p}{p+1-q}} \right) \tag{6}$$

where SE stands for 'standard error', p is the number of variables in the analysis, and q is the factor number, that is the position of the factor during extraction. The correlation SE can be obtained from the tabes likewise in the previous case.

The formula, in addition to providing some assessment of the SE of a loading, can serve to evaluate the common factors. First the SE is calculated, then doubled. Only those factors possessing at least half the total number of variables with values in excess of a doubled SE should be considered. The method works out to be exceedingly stringent, especially for small samples.

3.5 Rotation of factors

The main objection to factor analysis is that there is an infinity of mathematically correct and equivalent solutions. While this is true, it is also the case that factorists have developed powerful methods or choosing the right solution.

The methods of analysis described so far are sometimes referred to as *direct methods* because the factor matrix obtained arises directly from the correlation matrix by the application of mathematical models. Are the reference axes from direct solutions always in a position to give the most illuminating evaluation of the variables? Mathematically equivalent designs can, and do, give rise to a variety of alternative and equally acceptable solutions. Most factor analysts are now agreed that some direct solutions are not sufficient. The process of manipulating the reference axes is known as rotation. The results of rotation methods are sometimes referred to as *derived solutions* because they are obtained as a second stage from the results of direct solutions. The term 'rotation' applied to the reference axes means exactly what it says, namely, the axes are turned about the origin until some alternative position has been reached. The simplest case arises when the axes are maintained at 90 degrees thus giving an orthogonal rotation. Further, it is quite possible, and more popular, to rotate the axes through different angles to arrive at an oblique rotation. The orthogonal procedure can be easily seen using a graphical approach, an approach which is still used by a few researches in the final stage of rotation. But in general, of course, hand rotation is a laborious technique and the broad use of computer

facilities has encouraged most researches to rely on 'analytical' rotation, that is, obtained computerized solutions using mathematical approximations.

The earliest attempts at rotation came in the 1930's when Thurstone expounded his theory of *simple structure*. His primary objective was to organize the factor axes (and hence the loadings) so that their meaning would make better sense in terms of the problem discovered. In his particular case, he was concerned with the structure of mental abilities. For him, direct solutions, which satisfied the principle of parsimony in reducing a large number of related variables to a small number of independent factors, were not adequate. In addition to parsimony, solutions should be invariant, unique and in accord with non-factorial research findings. He believed that factor analysis was most appropriately used as the first stage in mapping out new domains and not as an end in itself.

By invariance, Thurstone was referring to the constancy of factor content from one analysis to the next. In an attempt to fulfil the requirements of unique and invariant factors, Thurstone established several criteria to assist in the decision as to when rotation should cease. The criteria were intuitive and are not rigidly adhered to nowadays, although they still lurk in the background of most subsequent formulations. They are based on the principle that the simplest explanation involving only a few variables is the best. The fullest statement appears in his book on *Multiple Factor Analysis* in which he proposes five conditions for the fulfilment of simple structure for an orthogonal or oblique analysis. If we take a factor matrix from a direct evaluation, the derived matrix after rotation should meet the following requirements.

1. Each row of the derived matrix, that is the loadings associated with one variable, should contain at least one zero loading. A zero loading would include numerical values which were not statistically significant.

2. If there are m common factors being used in the rotation (selected using one of the criteria for deciding on the number of worthwhile factors as indicated above) there should be at least m zero loadings in each factor.

3. For every pair of factors there should be several variables with zero loadings in one factor but having at the same time significant loadings in the other.

4. For every pair of factors a large proportion of the loadings should have zero values in both factors where there are four or more factors.

5. For every pair of factors there should only be a small proportion of loadings with significant values in both factors.

These criteria have the effect of maximizing the number of loadings having negligible values whilst leaving a few with large loadings. This makes the job of interpreting a factor very much easier than would a collection of moderately sized loadings.

3.5.1 Orthogonal rotation

As it was mentioned above, in orthogonal rotation the factors are rotated such that they are always at right angles to each other. This means that the factors are uncorrelated. But as Catelli has argued, in searching for factors which are fundamental dimensions for understanding the (psychological) phenomena it is unlikely, *a priori*, that factors would be uncorrelated.

3.5.2 Oblique rotation

In oblique rotation the factor axes can take any position in factor space, hence the name. The cosine of the angle between the factor axes indicates the correlation between them. The oblique case is rather more complex than the orthogonal. No entirely satisfactory analytical rotations have been devised for oblique solutions and they are still the subject of considerable experimentation and controversy. Cattelll (1952) is a keen protagonist of the method and he has made a significant contribution in this direction.

3.6 How good is the solution?

One of the tests of the quality of a factor analysis is to see how accurately the correlations can be reproduced from the factors. Indeed it is this ability to reproduce the correlations which further demonstrates that factors account for variance. Even more importantly it is this ability which makes factors useful in research. If all the principal components of a matrix are extracted, then the correlations between the variables can be perfectly reproduced. However, this is not a simplification or a reduction of the dimension of a problem since there are as many components as variables. In practice, of course, only a few components are extracted, the largest in terms of variance accounted for, and these should be able to account for the correlations partially. A good test of the adequacy of

an analysis is to reproduce the correlations and then subtract them from their originals. What is left is referred to as residual matrix. If the elements of this are small then the analysis is satisfactory.

The following equation shows how correlations are reproduced from factor loadings, where two factors have been extracted

$$r_{xy} = r_{x_1y_1} + r_{x_2y_2} \tag{7}$$

where r_{xy} is the correlation of variables x and y, $r_{x_1y_1}$ is the cross product of the factor loadings of variables x and y on factor1, and $r_{x_2y_2}$ is the cross product of the factor loadings of variables x and y on factor2.

4 Principal component analysis

4.1 Description of a model

A principal component analysis is concerned with explaining the variance-covariance structure of a set of variables through a few linear combinations of these variables. Its general objectives are

1. data reduction

2. interpretation

Although p components are required to reproduce the total system variability, often much part of it can be accounted by small number k of the principal components. The k principal components can then replace the initial p variables, and the original data set, consisting of n measurements on p variables, is reduces to a data set consisting of n measurements on k principal components. Analyses of principal components are more of a means to an end rather than an end in themselves, because they frequently serve as intermediate steps in much larger investigations. For example, principal components may be inputs to a multiple regression or cluster analysis. Algebraically, principal components are particular linear combinations of the p random variables $X_1, X_2, \ldots, X_p$. Geometrically, these linear combinations represent the selection of a new coordinate system obtained by rotating the original system with $X_1, X_2, \ldots, X_p$ as the coordinate axes. The new

axes represent the directions with maximum variability and provide a simpler and more parsimonious description of the covariance structure.

Let the random vector

$$\mathbf{X}^t = [X_1, X_2, \ldots, X_p]$$

where t stands for 'transformed' have the covariance matrix Σ with eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p \geqslant 0$.

Consider the linear combinations

$$\begin{aligned} Y_1 &= \mathbf{a}_1^t \mathbf{X} = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p \\ Y_2 &= \mathbf{a}_2^t \mathbf{X} = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p \\ &\ldots \\ Y_p &= \mathbf{a}_p^t \mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p \end{aligned} \tag{8}$$

Then we have

$$\mathrm{Var}(Y_i) = \mathbf{a}_i^t \mathbf{\Sigma} \mathbf{a}_i, \qquad i = 1, 2, \ldots, p \tag{9}$$

$$\mathrm{Cov}(Y_i, Y_k) = \mathbf{a}_i^t \mathbf{\Sigma} \mathbf{a}_k, \qquad i, k = 1, 2, \ldots, p \tag{10}$$

The principal components are those uncorrelated linear combinations $Y_1, Y_2, \ldots, Y_p$ whose variances are as large as possible. The first principal component is the linear combination with maximum variance. That is it maximizes $\mathrm{Var}(Y_1) = \mathbf{a}_1^t \mathbf{\Sigma} \mathbf{a}_1$. But it is clear that $\mathrm{Var}(Y_1) = \mathbf{a}_1^t \mathbf{\Sigma} \mathbf{a}_1$ can be increased by multiplying $\mathbf{a}_1$ by some constant. To eliminate this indeterminacy, it is convenient to restrict attention to coefficient vectors of unit length. We therefore define

1. First principal component = linear combination $a_1^t X$ that maximizes $\mathrm{Var}(\mathbf{a}_1^t \mathbf{X})$ subject to $\mathbf{a}_1^t \mathbf{a}_1 = 1$

2. Second principal component = linear combination $\mathbf{a}_2^t \mathbf{X}$ that maximizes $\mathrm{Var}(\mathbf{a}_2^t \mathbf{X})$ subject to $\mathbf{a}_2^t \mathbf{a}_2 = 1$ and $\mathrm{Cov}(\mathbf{a}_1^t \mathbf{X}, \mathbf{a}_2^t \mathbf{X}) = 0$

   ...

3. ith principal component = linear combination $\mathbf{a}_i^t \mathbf{X}$ that maximizes $\mathrm{Var}(\mathbf{a}_i^t \mathbf{X})$ subject to $\mathbf{a}_i^t \mathbf{a}_i = 1$ and $\mathrm{Cov}(\mathbf{a}_i^t \mathbf{X}, \mathbf{a}_k^t \mathbf{X}) = 0$ for $k < i$

We are ready now to formulate the following

Theorem 1. Let the symmetric positive definite matrix $\mathbf{\Sigma}$ be the covariance matrix associated with the random vector $\mathbf{X}^t = [X_1, X_2, \ldots, X_p]$. Let $\mathbf{\Sigma}$ have the eigenvalue-eigenvector pairs $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ where $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p \geqslant 0$ and the eigenvectors have unit norms. Then the ith principal component is given by

$$Y_i = \mathbf{e}_i^t \mathbf{X} = e_{i1}X_1 + e_{i2}X_2 + \ldots + e_{ip}X_p, \qquad i = 1, \ldots, p \tag{11}$$

With these choices

$$\mathrm{Var}(Y_i) = \mathbf{e}_i^t \mathbf{\Sigma} \mathbf{e}_i = \lambda_i, \qquad i = 1, 2, \ldots, p \tag{12}$$

$$\mathrm{Cov}(Y_i, Y_k) = \mathbf{e}_i^t \mathbf{\Sigma} \mathbf{e}_k, \qquad i \neq k \tag{13}$$

In case of multiple eigenvalues, the choices of the corresponding coefficient vectors $\mathbf{e}_i$, and hence Y_i, are not unique; only the subspaces are unique.

Proof. It is known from matrix algebra that

$$\max_{\mathbf{a} \neq \mathbf{0}} \frac{\mathbf{a}^t \mathbf{\Sigma} \mathbf{a}}{\mathbf{a}^t \mathbf{a}} = \lambda_1 \tag{14}$$

which is attained when $\mathbf{a} = \mathbf{e}_1$. But $\mathbf{e}_1^t \mathbf{e}_1 = 1$ since the eigenvectors are normalized. Thus,

$$\max_{\mathbf{a} \neq \mathbf{0}} \frac{\mathbf{a}^t \mathbf{\Sigma} \mathbf{a}}{\mathbf{a}^t \mathbf{a}} = \lambda_1 = \frac{\mathbf{e}_1^t \mathbf{\Sigma} \mathbf{e}_1}{\mathbf{e}_1^t \mathbf{e}_1} = \mathbf{e}_1^t \mathbf{\Sigma} \mathbf{e}_1 = \mathrm{Var}(Y_1) \tag{15}$$

Similarly we get

$$\max_{\mathbf{a} \perp \mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k} \frac{\mathbf{a}^t \mathbf{\Sigma} \mathbf{a}}{\mathbf{a}^t \mathbf{a}} = \lambda_{k+1}, \qquad k = 1, 2, \ldots, p-1 \tag{16}$$

For the choice $\mathbf{a} = \mathbf{e}_{k+1}$, with $\mathbf{e}_{k+1}^t \mathbf{e}_i = 0$, for $i = 1, 2, \ldots, k$ and $k = 1, 2, \ldots, p-1$,

$$\frac{\mathbf{e}_{k+1}^t \mathbf{\Sigma} \mathbf{e}_{k+1}}{\mathbf{e}_{k+1}^t \mathbf{e}_{k+1}} = \mathbf{e}_{k+1}^t \mathbf{\Sigma} \mathbf{e}_{k+1} = \mathrm{Var}(Y_{k+1}) \tag{17}$$

But $\mathbf{e}_{k+1}^t (\mathbf{\Sigma} \mathbf{e}_{k+1}) = \lambda_{k+1} \mathbf{e}_{k+1}^t \mathbf{e}_{k+1} = \lambda_{k+1}$ so $\mathrm{Var}(Y_{k+1}) = \lambda_{k+1}$. It remains to show that $\mathbf{e}_i^t \mathbf{e}_k = 0, \quad i \neq k$ gives $\mathrm{Cov}(Y_i, Y_k) = 0$. Now, the eigenvectors of $\mathbf{\Sigma}$ are orthogonal if all the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ are distinct. If they are not, the eigenvectors corresponding to multiple eigenvalues can be chosen to be orthogonal within the corresponding eigenspaces. Therefore, for any two eigenvectors $\mathbf{e}_i$ and $\mathbf{e}_k$, $\mathbf{e}_i^t \mathbf{e}_k = 0, \quad i \neq k$. Since $\mathbf{\Sigma} \mathbf{e}_k = \lambda_k \mathbf{e}_k$, premultiplication by $\mathbf{e}_i^t$ gives

$$\mathrm{Cov}(Y_i, Y_k) = \mathbf{e}_i^t \mathbf{\Sigma} \mathbf{e}_k = \mathbf{e}_i^t \lambda_k \mathbf{e}_k = \lambda_k \mathbf{e}_i^t \mathbf{e}_k = 0 \tag{18}$$

for any $i \neq k$, and the proof is complete.

Thus, according to **Theorem 1**, the principal components are uncorrelated and have variances equal to the eigenvalues of $\mathbf{\Sigma}$.

Theorem 2. Let $\mathbf{X}^t = [X_1, X_2, \ldots, X_p]$ have covariance matrix $\mathbf{\Sigma}$, with eigenvalue-eigenvector pairs $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ where $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p \geqslant 0$. Let $Y_1 = \mathbf{e}_1^t\mathbf{X}, Y_2 = \mathbf{e}_2^t\mathbf{X}, \ldots, Y_p = \mathbf{e}_p^t\mathbf{X}$ be the principal components. Then the total variance of the data is equal to the total variance of the components.

$$\sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp} = \sum_{i=1}^{p} \mathrm{Var}(Y_i) \tag{19}$$

Proof. On one hand

$$\sum_{i=1}^{p} \mathrm{Var}(X_i) = \sum_{i=1}^{p} \sigma_{ii} = \mathrm{tr}\mathbf{\Sigma} = \mathrm{tr}\mathbf{\Lambda} = \sum_{i=1}^{p} \lambda_i \tag{20}$$

On the other hand

$$\sum_{i=1}^{p} \mathrm{Var}(Y_i) = \sum_{i=1}^{p} \lambda_i \tag{21}$$

This completes the proof.

As a consequence, the proportion of total variance due to (explained by) the kth principal component is

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} \qquad k = 1, 2, \ldots, p \tag{22}$$

If most of the total variance, for large p, can be attributed to the first one, two, or three components, then these components can "replace" the original p variables without much loss of information.

Each component of the coefficient vector $\mathbf{e}_i^t = [e_{i1}, \ldots, e_{ik}, \ldots, e_{ip},]$ also merits inspection. The magnitude of e_{ik} is proportional to the correlation coefficient between Y_i and X_k.

Theorem 3. If $Y_1 = \mathbf{e}_1^t\mathbf{X}, Y_2 = \mathbf{e}_2^t\mathbf{X}, \ldots, Y_p = \mathbf{e}_p^t\mathbf{X}$ are the principal components obtained from the covariance matrix $\mathbf{\Sigma}$, then

$$\rho_{Y_i, X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \qquad i, k = 1, 2, \ldots, p \tag{23}$$

are the correlation coefficients between the components Y_i and the variables X_k.

Proof. Set $\mathbf{a}_k^t = [0, \ldots, 0, 1, 0, \ldots, 0]$ so that $X_k = \mathbf{a}_k^t\mathbf{X}$ and $\mathrm{Cov}(X_k, Y_i) = \mathrm{Cov}(\mathbf{a}_k^t\mathbf{X}, \mathbf{e}_i^t\mathbf{X}) = \mathbf{a}_k^t\mathbf{\Sigma}\mathbf{e}_i$. Since $\mathbf{\Sigma}\mathbf{e}_i = \lambda_i\mathbf{e}_i$, $\mathrm{Cov}(X_k, Y_i) = \mathbf{a}_k^t\lambda_i\mathbf{e}_i = \lambda_i e_{ik}$. Then $\mathrm{Var}(Y_i) = \lambda_i and \mathrm{Var}(X_k) =$

σ_{kk} yield

$$\rho_{Y_i,X_k} = \frac{\text{Cov}(Y_i, X_k)}{\sqrt{\text{Var}(Y_i)}\sqrt{\text{Var}(X_k)}} = \frac{\lambda_i e_{ik}}{\sqrt{\lambda_i}\sqrt{\sigma_{kk}}} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \qquad i,k = 1,2,\ldots,p \tag{24}$$

The proof is complete.

Although the correlations of the variables with the principal components often help to interpret the components, they measure only the univariate contribution of an individual $\mathbf{X}$ to a component $\mathbf{Y}$. For this reason some statisticians recommend that only the coefficients e_{ik}, and not the correlations, are used to interpret the components. But in practice, variables with relatively large coefficients (in absolute value) tend to have relatively large correlations, so the two measures of importance, the first multivariate and the second univariate, frequently give similar results. We recommend that both the coefficients and the correlations are examined to help interpret the principal components.

4.2 Number of components to extract

There is always the question of how many components to retain. There is no definitive answer to this question. Things to consider include the amount of total sample variance explained, the relative sizes of the eigenvalues, and the subject-matter interpretations of the components. A useful visual aid to determining an appropriate number of principal components is a scree plot. Recall that a scree plot is a plot of the ordered from largest to smallest eigenvalues versus their number. To determine the appropriate number of components, we look for an elbow (bend) in the scree plot. The number of components is taken to be the point at which the remaining eigenvalues are relatively small and all about the same size.

4.3 Interpreting Principal Components

Since the principal components are linear combinations of the original variables, it is often necessary to interpret or provide a meaning to the linear combination. One can use the loadings for interpreting the principal components. The higher the loadings of a variable in absolute value are, the more influence it has in the formation of the principal component score and vice versa. Therefore, one can use the loadings to determine which variables are influential in the formation of principal components, and one can then assign a meaning or label to the principal component.

4.4 Use of principal component scores

The principal components scores can be plotted for further interpreting the results. Based on a visual examination of the plot, one might argue that there are some certain groups or clusters of variables. The scores resulting from the principal components can also be used as input variables for further analyzing the data using other multivariate techniques such as cluster analysis, regression, and discriminant analysis.

5 Factor analysis

5.1 Description of a model

The essential purpose of factor analysis is to describe, if possible, the covariance relationships among many variables in terms of a few underlying, but unobservable, random quantities called factors. Basically, the factor model is motivated by the following argument. Suppose variables can be grouped by their correlations. That is, suppose all variables within a particular group are highly correlated among themselves, but have relatively small correlations with variables in a different group. Then it is conceivable that each group of variables represents a single underlying construct, or factor, that is responsible for the observed correlations.

Factor analysis can be considered as an extension of principal component analysis. Both can be viewed as attempts to approximate the covariance matrix Σ. However, the approximation based on the factor analysis model is more elaborate. The primary question in factor analysis is whether the data are consistent with a prescribed structure.

5.2 The orthogonal factor model

Suppose the observable random vector $\mathbf{X}$, with p components, has mean μ and correlation matrix $\mathbf{\Sigma}$. The factor model postulates that $\mathbf{X}$ is linearly dependent upon a few unobservable variables $F_1, F_2, \ldots, F_q$, called common factors, and p additional sources of variation $\epsilon_1, \epsilon_2, \ldots, \epsilon_p$, called errors or, sometimes, specific factors. In particular, the

factor analysis model is

$$X_1 - \mu_1 = l_{11}F_1 + l_{12}F_2 + \ldots + l_{1q}F_q + \epsilon_1$$
$$X_2 - \mu_2 = l_{21}F_1 + l_{22}F_2 + \ldots + l_{2q}F_q + \epsilon_2 \tag{25}$$
$$\ldots$$
$$X_p - \mu_p = l_{p1}F_1 + l_{p2}F_2 + \ldots + l_{pq}F_q + \epsilon_p \tag{26}$$

or, in matrix notation,

$$\mathbf{X} - \mu = \mathbf{LF} + \epsilon \tag{27}$$

The coefficient l_{ij} is called the loading of the ith variable on the jth factor, so the matrix $\mathbf{L}$ is the matrix of factor loadings. Note that the ith specific factor ϵ_i is associated only with the ith response X_i. The p deviations $X_1 - \mu_1$, $X_2 - \mu_2$, ..., $X_p - \mu_p$ are expressed in terms of $p + q$ random variables $F_1, F_2, \ldots, F_q, \epsilon_1, \epsilon_2, \ldots, \epsilon_p$ which are unobservable. Thus with so many unobservable quantities, a direct verification of the factor model from observations on $X_1, X_2, \ldots, X_p$ is hopeless. However, with some additional assumptions about the random vectors $\mathbf{F}$ and ϵ, the model implies certain covariance relationships, which can be checked.

We assume that

$$E(\mathbf{F}) = 0, \qquad \text{Cov}(\mathbf{F}) = E[\mathbf{F}\mathbf{F}'] = \mathbf{I}$$

$$E(\epsilon) = 0, \qquad \text{Cov}(\epsilon) = E[\epsilon\epsilon'] = \mathbf{\Psi} = \begin{bmatrix} \psi_1 & 0 & \ldots & 0 \\ 0 & \psi_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \psi_p \end{bmatrix} \tag{28}$$

and that $\mathbf{F}$ and ϵ are independent, so

$$\text{Cov}(\epsilon, \mathbf{F}) = E(\epsilon\mathbf{F}') = \mathbf{0}. \tag{29}$$

The orthogonal factor model implies a covariance structure for X. From the model we have

$$(\mathbf{X} - \mu)(\mathbf{X} - \mu)' = (\mathbf{LF} + \epsilon)(\mathbf{LF} + \epsilon)'$$
$$= \mathbf{LF}(\mathbf{LF})' + \epsilon(\mathbf{LF})' + \mathbf{LF}\epsilon' + \epsilon\epsilon'$$

so that

$$\begin{aligned}\boldsymbol{\Sigma} = \text{Cov}(\mathbf{X}) = E(\mathbf{X}-\mu)(\mathbf{X}-\mu)' \\ = \mathbf{L}E(\mathbf{F}\mathbf{F}')\mathbf{L}' + E(\epsilon\mathbf{F}') + \mathbf{L}E(\mathbf{F}\epsilon') + E(\epsilon\epsilon') \\ = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}\end{aligned} \tag{30}$$

Also, by independence, $\text{Cov}(\epsilon, \mathbf{F}) = E(\epsilon, \mathbf{F}') = 0$. Also,by the model $(\mathbf{X}-\mu)\mathbf{F}' = (\mathbf{L}\mathbf{F}+\epsilon)\mathbf{F}' = \mathbf{L}\mathbf{F}\mathbf{F}'+\epsilon\mathbf{F}'$, so $\text{Cov}(\mathbf{X}, \mathbf{F}) = E((\mathbf{X}-\mu)\mathbf{F}') = \mathbf{L}E(\mathbf{F}\mathbf{F}')+E(\epsilon\mathbf{F}') = \mathbf{L}$. Thus, we get the following **Covariance structure for the orthogonal factor model.**

1. $\text{Cov}(\mathbf{X}) = \mathbf{L}\mathbf{L}^t + \boldsymbol{\Psi}$ or

$$\begin{aligned}\text{Var}(X_i) = l_{i1}^2 + \ldots + l_{iq}^2 + \psi_i \\ \text{Cov}(X_i, X_k) = l_{i1}l_{k1} + \ldots + l_{iq}l_{kq}\end{aligned} \tag{31}$$

2.

$$\mathbf{Cov}(\mathbf{X}, \mathbf{F}) = \mathbf{L} \tag{32}$$

or

$$\mathbf{Cov}(X_i, F_j) = l_{ij}$$

The model $\mathbf{X}-\mu = \mathbf{L}\mathbf{F}+\epsilon$ is linear in the common factors. If the p responses of $\mathbf{X}$ are, in fact, related to underlying factors, but the relationship is nonlinear, then the covariance structure $\mathbf{L}\mathbf{L}^t + \boldsymbol{\Psi}$ may not be adequate. The very important assumption of linearity is inherent in the formulation of the traditional factor model.

Recall that the portion of the variance of the ith variable contributed by the m common factors is called the ith communality. That portion of $\text{Var}(X_i) = \sigma_{ii}$ due to the specific factor is often called the uniqueness, or specific variance. Denoting the ith communality by h_i^2, we see that

$$\text{Var}(X_i) = \sigma_{ii} = h_i^2 + \psi_i, \qquad i = 1, 2, \ldots, p \tag{33}$$

where

$$h_i^2 = l_{i1}^2 + l_{i2}^2 + \ldots + l_{iq}^2 \tag{34}$$

The factor model assumes that the $p+p(p-1)/2 = p(p+1)/2$ variances and covariances for $\mathbf{X}$ can be reproduced from the pq factor loadings l_{ij} and the p specific variances ψ_i. When $q = p$, any covariance matrix $\boldsymbol{\Sigma}$ can be reproduced exactly as $\mathbf{L}\mathbf{L}^t$, so $\boldsymbol{\Psi}$ can be

the zero matrix. However, it is when q is small relative to p that factor analysis is more useful. In this case, the factor model provides a "simple" explanation of the covariation in $\mathbf{X}$ with fewer parameters than $p(p+1)/2$ parameters in $\mathbf{\Sigma}$. Unfortunately for the factor analyst, most covariance matrices cannot be factored as $\mathbf{LL}^t + \mathbf{\Psi}$, where the number of factors q is much less than p. Thus, a question, for which minimal $q < p$, $q \in \mathbb{N}$, $p \in \mathbb{N}$, could the $p-$dimensional vector of observations be explained by the $q-$dimensional vector of factors, has to be answered. We could evaluate this number by counting the number of parameters and the number of equations of the model. All together, there are $pq + p$ unknown parameters, as it was mentioned above, and $(1/2)p(p+1)$ equations. However, while estimating the model parameters, we will be assuming that the matrix $\Delta = \mathbf{L}^t\mathbf{\Psi}^{-1}\mathbf{L}$ is diagonal to make $\mathbf{L}$ be unique (otherwise $\mathbf{L}$ is unique only up to an orthogonal rotation). This condition gives us $(1/2)q(q-1)$ additional equations. Thus, we can expect to get a solution of the system if the difference between those two numbers,

$$s = (1/2)p(p+1) + (1/2)q(q-1) - (pq-p) = (1/2)[(p-q)^2 - (p+q)] \tag{35}$$

is non-positive. Solving this quadratic equation w.r.t. q, we get the following lower bound for the number of factors

$$q \geq (2p + 1 - \sqrt{8p+1})/2. \tag{36}$$

5.3 Methods of estimation

Given observations $x_1, x_2, \ldots, x_n$ on p generally correlated variables, factor analysis seeks to answer the question, Does the factor model (linear in our case), with a small number of factors, adequately represent the data? In essence, we tackle this statistical model-building problem by trying to verify the covariance relationship (**??**), (**??**). The sample covariance matrix $\mathbf{S}$ is an estimator of the unknown population covariance matrix $\mathbf{\Sigma}$. If the off-diagonal elements of $\mathbf{S}$ are small or those of the sample correlation matrix are $\mathbf{R}$ essentially zero, the variables are not related, and a factor analysis will not prove useful. In these circumstances, the specific factors play the dominant role, whereas the major aim of factor analysis is to determine a few important common factors.

If $\mathbf{\Sigma}$ appears to deviate significantly from a diagonal matrix, then a factor model can be entertained, and the initial problem is one of estimating the factor loadings l_{ij} and specific variables ψ_i. We shall consider two of the most popular methods of parameter

estimation, the principal component method and the maximum likelihood method. It is always prudent to try more than one method of solution; if the factor model is appropriate for the problem at hand, the solutions should be consistent one with another.

Current estimation methods require iterative calculations that must be done on a computer. Several computer programs are now available for this purpose.

5.3.1 The principal component method

The name of the method follows from the fact that the factor loadings are the scaled coefficients of the first few sample principal components discussed in the previous chapter.

Principal component solution of the factor model

The principal component factor analysis of the sample covariance matrix $\mathbf{S}$ is specified in terms of its eigenvalue-eigenvector pairs $(\hat{\lambda}_1\hat{\mathbf{e}}_1), (\hat{\lambda}_2, \hat{\mathbf{e}}_2), \ldots, (\hat{\lambda}_p, \hat{\mathbf{e}}_p)$, where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p$. Let $q < p$ be the number of common factors. Then the matrix of estimated factor loadings $\{\tilde{l}_{ij}\}$ is given by

$$\tilde{\mathbf{L}} = \left[\sqrt{\hat{\lambda}_1}\hat{\mathbf{e}}_1 | \sqrt{\hat{\lambda}_2}\hat{\mathbf{e}}_2 | \cdots | \sqrt{\hat{\lambda}_q}\hat{\mathbf{e}}_q\right]. \tag{37}$$

The estimated specific variances are provided by the diagonal elements of the matrix $\mathbf{S} - \tilde{\mathbf{L}}\tilde{\mathbf{L}}^t$, so

$$\tilde{\mathbf{\Psi}} = \begin{bmatrix} \tilde{\psi}_1 & 0 & \ldots & 0 \\ 0 & \tilde{\psi}_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \tilde{\psi}_p \end{bmatrix} \qquad with \qquad \tilde{\psi}_i = s_{ii} - \sum_{j=1}^{q} \tilde{l}_{ij}^2 \tag{38}$$

Communalities are estimated as

$$\tilde{h}_i^2 = \tilde{l}_{i1}^2 + \tilde{l}_{i2}^2 + \ldots + \tilde{l}_{iq}^2 \tag{39}$$

The principal component factor analysis of the sample correlation matrix is obtained by starting with $\mathbf{R}$ in place of $\mathbf{S}$.

By the definition of $\tilde{\psi}_i$, the diagonal elements of $\mathbf{S}$ are equal to the diagonal elements of $\tilde{\mathbf{L}}\tilde{\mathbf{L}}^t + \tilde{\mathbf{\Psi}}$. However, the off-diagonal elements of $\mathbf{S}$ are not usually reproduced by $\tilde{\mathbf{L}}\tilde{\mathbf{L}}^t + \tilde{\mathbf{\Psi}}$. How, then, do we select the number of factors q?

If the number of common factors is not determined by a priori considerations, such as by theory or the work of other researchers, the choice of q can be based on the estimated eigenvalues in much the same manner as with principal components.

Consider the *residual* matrix

$$\mathbf{S} - (\tilde{\mathbf{L}}\tilde{\mathbf{L}}^t + \tilde{\mathbf{\Psi}}) \tag{40}$$

resulting from the approximation of $\mathbf{S}$ by the principal component solution. The diagonal elements are zero, and if the other elements are also small, we may subjectively take the q factor model to be appropriate. Analytically, we have

$$\text{Sum of squared entries of } (\mathbf{S} - (\tilde{\mathbf{L}}\tilde{\mathbf{L}}^t + \tilde{\mathbf{\Psi}})) \leq \hat{\lambda}_{q+1}^2 + \ldots + \hat{\lambda}_p^2 \tag{41}$$

Consequently, a small value for the sum of the squares of the neglected eigenvalues implies a small value for the sum of the squared errors of approximation.

Ideally the contribution of the first few factors to the sample variances of the variables should be large. In general, proportion of total sample variance due to jth factor is

$$\frac{\hat{\lambda}_j}{s_{11} + s_{22} + \ldots + s_{pp}} \qquad for\ a\ FA\ of\ \mathbf{S} \tag{42}$$

$$\frac{\hat{\lambda}_j}{p} \qquad for\ a\ FA\ of\ \mathbf{R} \tag{43}$$

Criterion (**??**) is frequently used as a heuristic device for determining the appropriate number of common factors. The number of common factors retained in the model is increased until a "suitable proportion" of the total sample variance has been explained.

5.3.2 The Principal Factor Solution

A modification of the principal components approach is sometimes considered. Suppose that initial estimates ψ_i^* of the specific variances are available. Then replacing the ith diagonal element of $\mathbf{R}$ by $h_i^{*2} = 1 - \psi_i^*$, we obtain a "reduced" sample correlation matrix

$$\mathbf{R}_r = \begin{bmatrix} h_1^{*2} & r_{12} & \ldots & r_{1p} \\ r_{12} & h_2^{*2} & \ldots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \ldots & h_p^{*2} \end{bmatrix}$$

All the elements of the reduced sample correlation matrix $\mathbf{R_r}$ should be accounted for by the q common factors. In particular, $\mathbf{R_r}$ is factored as

$$\mathbf{R_r} = \mathbf{L_r^*}\mathbf{L_r^{*t}} \tag{44}$$

where $\mathbf{L}^*_{\mathbf{r}} = \left(l^*_{ij}\right)$. The principal factor method of FA employs the estimates

$$\mathbf{L}^*_r = \left[\sqrt{\hat{\lambda}^*_1}\hat{\mathbf{e}}^*_1|\sqrt{\hat{\lambda}^*_2}\hat{\mathbf{e}}^*_2|\cdots|\sqrt{\hat{\lambda}^*_q}\hat{\mathbf{e}}^*_q\right]$$

$$\psi^*_i = 1 - \sum_{j=1}^{q} l^{*2}_{ij} \tag{45}$$

where $(\hat{\lambda}^*_1, \hat{e}^*_1)$, $i = 1, 2, \ldots, q$ are the (largest) eigenvalue-eigenvector pairs determined from $\mathbf{R}_{\mathbf{r}}$. In turn, the communalities would be then reestimated by

$$\tilde{h}^{*2}_i = \sum_{j=1}^{q} l^{*2}_{ij} \tag{46}$$

The principal factor solution can be obtained iteratively, with the communality estimates of (**??**) becoming the initial estimates for the next stage.

Although there are many choices for initial estimates of specific variances, the most popular one, working with a correlation matrix, is $\psi^*_i = 1/r^{ii}$, where r^{ii} is the ith diagonal element of $\mathbf{R}^{-1}$.

5.3.3 The Maximum Likelihood Method

If the common factors $\mathbf{F}$ and the specific factors ϵ can be assumed to be normally distributed, then maximum likelihood estimates of the factor loadings and specific variances may be obtained. When $\mathbf{F}_j$ and ϵ_j are jointly normal, the observations $\mathbf{X}_j - \mu = \mathbf{L}\mathbf{F}_j + \epsilon_j$ are then normal, and the likelihood is

$$\begin{aligned} L(\mu, \mathbf{\Sigma}) &= (2\Pi)^{-\frac{np}{2}}|\mathbf{\Sigma}|^{-\frac{n}{2}} e^{-\frac{1}{2}tr\left[\mathbf{\Sigma}^{-1}(\sum_{j=1}^{n}(x_j-\overline{x})(x_j-\overline{x})' + n(\overline{x}-\mu)(\overline{x}-\mu)')\right]} \\ &= (2\Pi)^{\frac{-(n-1)p}{2}}|\mathbf{\Sigma}|^{-\frac{n-1}{2}} e^{-\frac{1}{2}tr\left[\mathbf{\Sigma}^{-1}(\sum_{j=1}^{n}(x_j-\overline{x})(x_j-\overline{x})')\right]} \\ &\quad \times (2\Pi)^{-\frac{p}{2}}|\mathbf{\Sigma}|^{-\frac{1}{2}} e^{-\frac{n}{2}(\overline{x}-\mu)'\mathbf{\Sigma}^{-1}(\overline{x}-\mu)} \end{aligned} \tag{47}$$

which depends on $\mathbf{L}$ and $\mathbf{\Psi}$ through $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^t + \mathbf{\Psi}$. This model is not still well defined, because of the multiplicity of choices for $\mathbf{L}$ made possible by orthogonal transformations. It is desirable to make $\mathbf{L}$ well defined by imposing the computationally convenient uniqueness condition $\mathbf{L}^t\mathbf{\Psi}^{-1}\mathbf{L} = \Delta$, where Δ is a diagonal matrix.

The maximum likelihood estimates $\hat{\mathbf{L}}$ and $\hat{\mathbf{\Psi}}$ must be obtained by numerical maximization of (**??**). Fortunately, efficient computer programs now exist that enable one to get these estimates rather easily.

Here we summarize some facts about maximum likelihood estimators.

Definition. Let $\mathbf{X_1}, \mathbf{X_1}, \ldots, \mathbf{X_n}$ be a random sample from $N_p(\mu, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^t + \mathbf{\Psi}$ is the covariance matrix for the q common factor model of (**??**). The maximum likelihood estimators $\hat{\mathbf{L}}$, $\hat{\mathbf{\Psi}}$, and $\hat{\mu} = \overline{\mathbf{x}}$ maximize (**??**) subject to $\hat{\mathbf{L}}^t\hat{\mathbf{\Psi}}^{-1}\hat{\mathbf{L}}$ being diagonal.

Denote

$$\mathbf{S} := \sum_{j=1}^{n} (\mathbf{X}_j - \overline{\mathbf{X}})(\mathbf{X}_j - \overline{\mathbf{X}})^t, \tag{48}$$

$$\hat{\mathbf{C}}_n := \frac{1}{n}\mathbf{S} \tag{49}$$

ML-estimator of the true covariance matrix $\mathbf{C}$, and

$$\hat{\mathbf{C}}_n^* := \frac{1}{n-1}\mathbf{S} \tag{50}$$

unbiased estimator of $\mathbf{C}$.

Remark. The eigenvalue-eigenvector pairs are unique functions of $\mathbf{C}$. The same functions of $\hat{\mathbf{C}}_n$ give ML-estimates of the factors.

Although a simple analytical expression cannot be obtained for the ML estimators $\hat{\mathbf{L}}$ and $\hat{\mathbf{\Psi}}$, they can be shown to satisfy certain equation. The conditions are stated in terms of the ML estimator $\hat{\mathbf{C}}_n$ of an unconstrained covariance matrix. Some factor analysts employ the corrected sample covariance $\hat{\mathbf{C}}_n^*$, but still use the title *maximum likelihood* to refer to resulting estimates. The factor analysis of $\mathbf{R}$ is, of course, unaffected by the choice of $\hat{\mathbf{C}}_n$ or $\hat{\mathbf{C}}_n^*$, since they both produce the same correlation matrix.

Theorem 4. With all the conditions above the ML estimates $\hat{\mathbf{L}}$ and $\hat{\mathbf{\Psi}}$ satisfy

$$(\hat{\mathbf{\Psi}}^{-1/2}\hat{\mathbf{C}}_n\hat{\mathbf{\Psi}}^{-1/2})(\hat{\mathbf{\Psi}}^{-1/2}\hat{\mathbf{L}}) = (\hat{\mathbf{\Psi}}^{-1/2}\hat{\mathbf{L}})(\mathbf{I}+\hat{\Delta}) \tag{51}$$

so the jth column of $\hat{\mathbf{\Psi}}^{-1/2}\hat{\mathbf{L}}$ is the eigenvector of $\hat{\mathbf{\Psi}}^{-1/2}\hat{\mathbf{C}}_n\hat{\mathbf{\Psi}}^{-1/2}$ corresponding to eigenvalue $1 + \hat{\Delta}_j$. Here $\hat{\Delta}_1 \geq \hat{\Delta}_2 \geq \ldots \hat{\Delta}_q$. Also, at convergence,

$$\hat{\psi}_i = i\text{th diagonal element of } \hat{\mathbf{C}}_n - \hat{\mathbf{L}}\hat{\mathbf{L}}^t \tag{52}$$

and

$$tr(\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{C}}_n) = p. \tag{53}$$

Sketch of the Proof. It is evident that $\hat{\mu} = \overline{\mathbf{x}}$ and a consideration of the log-likelihood leads to the maximization of

$$-(n/2)[ln|\mathbf{\Sigma}| + tr(\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{C}}_n)]$$

over $\mathbf{L}$ and $\mathbf{\Psi}$,as $\hat{\mathbf{C}}_n$ and p are constant with respect to the maximization. Thus we have to minimize

$$h(\mathbf{L}, \mathbf{\Psi}) = ln|\mathbf{\Sigma}| + \text{tr}(\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{C}}_n) \tag{54}$$

subject to $\mathbf{L}^t\mathbf{\Psi}^{-1}\mathbf{L} = \mathbf{\Delta}$, where $\mathbf{\Delta}$ is a diagonal matrix.

After differentiating w.r.t. $\mathbf{L}$ and $\mathbf{\Psi}$, and taking the derivatives equal to zero, we have the following system of equations:

$$\frac{\partial h}{\partial \mathbf{L}} = \mathbf{C}^{-1}(\mathbf{C} - \hat{\mathbf{C}}_n)\mathbf{C}^{-1}\mathbf{L} = \mathbf{0} \tag{55}$$

$$\frac{\partial h}{\partial \mathbf{\Psi}} = \mathbf{0}, \tag{56}$$

which implies

$$\text{diag}[\mathbf{C}^{-1}(\mathbf{C} - \hat{\mathbf{C}}_n)\mathbf{C}^{-1}] = \mathbf{0}. \tag{57}$$

Various numerical methods can be applied for solving this system of equations.

5.4 Factor rotation

When $q > 1$, there is always some inherent ambiguity associated with the factor model. To see this, let $\mathbf{T}$ be any $q \times q$ orthogonal matrix, so that $\mathbf{T}\mathbf{T}^t = \mathbf{T}^t\mathbf{T} = \mathbf{I}$. Thus, we can write the model in the following form

$$\mathbf{X} - \mu = \mathbf{L}\mathbf{F} + \epsilon = \mathbf{L}\mathbf{T}\mathbf{T}'\mathbf{F} + \epsilon = \mathbf{L}^*\mathbf{F}^* + \epsilon \tag{58}$$

where

$$\mathbf{L}^* = \mathbf{L}\mathbf{T} \qquad \textit{and} \qquad \mathbf{F}^* = \mathbf{T}^t\mathbf{F}, \tag{59}$$

since

$$E(\mathbf{F}^*) = \mathbf{T}'E(\mathbf{F}) = 0. \tag{60}$$

and

$$\text{Cov}(\mathbf{F}^*) = \mathbf{T}'\text{Cov}(\mathbf{F}) = \mathbf{T}'\mathbf{T} = \mathbf{I} \tag{61}$$

It is impossible, on the basis of observations on $\mathbf{X}$, to distinguish the loadings $\mathbf{L}$ from the loadings $\mathbf{L}^*$.That is, the factors $\mathbf{F}$ and $\mathbf{F}^* = \mathbf{T}^t\mathbf{F}$ have the same statistical properties, and even though the loadings $\mathbf{L}^*$ are, in general, different from the loadings $\mathbf{L}$, they both generate the same covariance matrix $\mathbf{\Sigma}$. That is

$$\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^t + \mathbf{\Psi} = \mathbf{L}\mathbf{T}\mathbf{T}^t\mathbf{L}^t + \mathbf{\Psi} = (\mathbf{L}^*)(\mathbf{L}^*)^t + \mathbf{\Psi} \tag{62}$$

This ambiguity provides the rationale for "factor rotation" since orthogonal matrices correspond to rotations of the coordinate system for $\mathbf{X}$. So we get the following rule

Factor loadings $\mathbf{L}$ *are determined only up to an orthogonal matrix* $\mathbf{T}$. *Thus, the loadings*

$$\mathbf{L}^* = \mathbf{LT} \qquad and \qquad \mathbf{L} \tag{63}$$

both give the same representations. The communalities, given by the diagonal elements of $\mathbf{LL}^t = (\mathbf{L}^*)(\mathbf{L}^*)^t$ *are also unaffected by the choice of* $\mathbf{T}$.

The analysis of the factor model proceeds by imposing conditions that allow one to uniquely estimate $\mathbf{L}$ and $\mathbf{\Psi}$. The loading matrix is then rotated (multiplied to an orthogonal matrix), where the rotation is determined by some "case-of-interpretation" criterion. Once the loadings and specific variances are obtained, factors are identified, and estimated values for the factors themselves (factor scores) are frequently constructed.

Since the original loadings may not be easily interpretable, it is usual practice to rotate them until a "simpler structure" in a sense mentioned above is achieved. The rationale is very much similar to sharpening the focus of a microscope in order to see the detail more clearly.

While the loadings are changed under rotation, the communality estimates remain unchanged, since $\mathbf{LL}^t = \mathbf{LTT}^t\mathbf{L}^t = \mathbf{L}^*\mathbf{L}^{*t}$, and the communalities are the diagonal elements of these matrices.

Kaiser has suggested an analytical measure of simple structure known as the *varimax criterion*. Define $\tilde{l}^*_{ij} = \frac{\hat{l}^*_{ij}}{\hat{h}_i}$ to be rotated coefficients scaled by the square roots of the communalities. The the varimax procedure selects the orthogonal transformation $\mathbf{T}$ that makes

$$V = \frac{1}{p}\sum_{j=1}^{q}\left[\sum_{i=1}^{p}\tilde{l}^{*4}_{ij} - \left(\sum_{i=1}^{p}\tilde{l}^{*2}_{ij}\right)^2/p\right] \tag{64}$$

as large as possible. Computing algorithms exist for maximizing V, and most popular factor analysis computer programs (for example, the statistical software packages SAS, SPSS, and others) provide varimax rotations.

5.5 Factor scores

In factor analysis, interest is usually centered on the parameters in the factor model. However, the estimated values of the common factors, called factor scores, may also be

required. These quantities are often used for diagnostic purposes, as well as inputs to a subsequent analysis.

Factor scores are not estimates of unknown parameters in the usual sense. Rather, they are estimates of values for the unobserved random factor vectors F_j, $j = 1, 2, \ldots, n$. That is, factor scores

$$\hat{\mathbf{f}}_\mathbf{j} = \text{estimate of the values } \mathbf{f_j} \text{ attained by } \mathbf{F}_j \ (j\text{th case})$$

The estimation situation is complicated by the fact that the unobserved quantities $\mathbf{f}_j$ and ϵ_j outnumber the observed $\mathbf{x}_j$. To overcome this difficulty, some rather heuristic, but reasoned, approaches to the problem of estimating factor values have been advanced. *Multiple regression* is one of such techniques. For example, the factor score for variable i on a given factor j can be represented as

$$\hat{F}_{ij} = \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip} \tag{65}$$

The equation can be represented in matrix form as

$$\hat{\mathbf{F}} = \mathbf{X}\hat{\mathbf{B}} \tag{66}$$

where $\hat{\mathbf{F}}$ is an $n \times q$ matrix of q factor scores for the n individuals, $\mathbf{X}$ is an $n \times p$ matrix of observed variables, and $\hat{\mathbf{B}}$ is a $p \times q$ matrix of estimated factor score coefficients.

5.6 Performing factor analysis

There are many decisions that must be made in any factor analytic study. Probably the most important one is the choice of q, the number of common factors. Most often the final choice of q is based on some combination of

1. the proportion of the sample variance explained
2. subject-matter knowledge
3. the "reasonableness" of the results.

The choice of the solution method and type of rotation is a less crucial decision. In fact, the most satisfactory factor analyses are those in which rotations are tried with more than one method and all the results substantially confirm the same factor structure. One possible option to perform factor analysis is

1. **Perform a principal component FA**. This method is particularly appropriate for a first pass through the data.

   (a) Look for suspicious observations by plotting factor scores.

   (b) Try a varimax rotation.

2. **Perform a maximum likelihood FA, including a varimax rortation.**

3. **Compare the solution obtained from the two previous analyses.**

   (a) Do the loadings group in the same manner?

   (b) Plot the factor scores obtained for principal components against scores from the maximum likelihood analysis.

4. **Repeat the first three steps for other number of common factors**. Do more factors contribute to the understanding and interpretation of the data?

5. **For large data sets, split them in half and perform a FA on each part**. Compare the two results with each other and with what obtained from the complete data set to check the stability of the solution.

5.7 Factor Analysis versus Principal Component Analysis

Although factor analysis and principal component analysis are typically labelled as data-reduction techniques, there are significant differences between these two. The objective of principal component analysis is to reduce the number of variables to a few components such that each component forms a new variable and the retained components explain the maximum amount of variance in the data. The objective of factor analysis, on the other hand, is to search or identify the underlying factor(s) or latent constructs that can explain the intercorrelation among the variables. There are the two major differences:

1. Principal component analysis places emphasis on explaining the variance in the data; the objective of factor analysis is to explain the correlation among the indicators;

2. In principal component analysis the variables form components; in factor analysis, on the other hand, the variables reflect the presence of unobservable constructs or factors.

Part II

Factor models in economics

Most of economics is concerned with modelling dynamics. There has been an explosion of research in this area in the last twenty years (Box et al. (1994), Deistler and Hamann (2005), Deistler and Zinner (2007), Forni et al. (2002 and 2002)), as "time series econometrics" has practically come to be synonymous with "empirical macroeconomics".

Since Geweke (1977), Box and Tiano (1977) generalized the classical factor model to a dynamic one, a lot of various dynamic factor models (Breitung and Eickmeier (2005), Forni et al. (2000)) have been developed and studied from the point of view of parameter estimation. The problem of describing comovements in multivariate time series by means of some nearly independent factors becomes more and more important when facing economic crises and looking for predictions.

Dynamic factor models are well designed to describe data having strong comovements: in such models, each of the series under study is supposed to depend linearly on a small number of common latent variables which are the sources of that comovements (the common factors) and on a residual term (the idiosyncratic component).

In recent years, large-dimensional dynamic factor models have become popular in empirical macroeconomics (Deister and Zinner (2007), Stock and Watson (2002)). They are more advantageous than other methods in various respects. Factor models can cope with many variables without running into scarce degrees of freedom problems often faced in regression-based analyses. Researchers and policy makers nowadays have more data at a more disaggregated level at their disposal than ever before. Once collected, the data can be processed easily and rapidly owing to the now wide-spread use of high-capacity computers. Exploiting a lot of information can lead to more precise forecasts and macroeconomic analyses. A second advantage of factor models is that idiosyncratic movements which possibly include measurement error and local shocks can be eliminated. This yields a more reliable signal for policy makers and prevents them from reacting to idiosyncratic movements. In addition, the estimation of common factors or common shocks is of intrinsic interest in some applications. A third important advantage is that factor modelers can remain agnostic about the structure of the economy and do not need

to rely on overly tight assumptions as is sometimes the case in structural models. It also represents an advantage over structural VAR models where the researcher has to take a stance on the variables to include which, in turn, determine the outcome, and where the number of variables determine the number of shocks. Dynamic factor models were traditionally used to construct economic indicators and for forecasting (Bánkövi et al. (1982), Deister and Hamann (2005)). Let us briefly discuss existing applications of dynamic factor models in these fields.

1. **Construction of economic indicators.** The two most prominent examples of monthly coincident business cycle indicators, to which policy makers and other economic agents often refer, are the Chicago Fed National Activity Index (CFNAI) for the US and EuroCOIN for the euro area. The CFNAI estimate, which dates back to 1967, is simply the first static principal component of a large macro data set. It is the most direct successor to indicators which were first developed by Stock and Watson but retired by the end of 2003. EuroCOIN is estimated as the common component of euro-area GDP based on dynamic principal component analysis.

2. **Forecasting.** Factor models are widely used in central banks and research institutions as a forecasting tool. The forecasting equation typically has the form

$$y_{t+h}^h = \mu + a(L)y_t + b(L)\hat{F}_t + \epsilon_{t+h}^h \tag{67}$$

where y_t is the variable to be predicted at period $t+h$ and ϵ_{t+h} denotes the h-step ahead prediction error. Accordingly, information used to forecast y_t are the past of the variable and the common factor estimates $\hat{F}_t$ extracted from an additional data set. Factor models have been used to predict real and nominal variables in the US, in the euro area, for Germany, for the UK, and for the Netherlands. The factor model forecasts are generally compared to simple linear benchmark time series models, such as AR models, AR models with single measurable leading indicators and VAR models. Overall, results are quite encouraging, and factor models are often shown to be more successful in terms of forecasting performance than smaller benchmark models. Three remarks are, however, in order. First, the forecasting performance of factor models apparently depends on the types of variable one wishes to forecast, the countries/regions of interest, the underlying data sets, the benchmark models and horizons. Unfortunately, a systematic assessment of the determinants of the relative

forecast performance of factor models is still not available. Second, it may not be sufficient to include just the first or the first few factors. Instead, a factor which explains not much of the entire panel, say, the fifth or sixth principal component, may be important for the variable one wishes to forecast. Finally, the selection of the variables to be included in the data set is ad hoc in most applications. The same data set is often used to predict different variables. This may, however, not be adequate. Instead, one should only include variables which exhibit high explanatory power with respect to the variable that one aims to forecast.

In making choices between alternative courses of action, decision makers at all structural levels often make predictions of economic variables. If time series observations are available for a variable of interest and the data from the past contain information about the future development of a variable, it is plausible to use as forecast some function of the data collected in the past. Assuming that the tendency prevails in future periods, forecasts can be based on current and past data. Let us go through some basic concepts of time series theory in order to be able to discuss the prediction-making techniques.

6 Time series analysis

A *time series* is a sequence of observations taken sequentially in time. If the data set is continuous, the time series is said to be continuous. If the set is discrete, the time series is said to be discrete.

The time series to be analyzed may then be thought of as a particular realization of the system under study. A statistical phenomenon that evolves in time according to probabilistic laws is called a *stochastic process*. So, in other words, in analyzing a time series we regard it as a realization of a stochastic process.

Given a particular realization such as $\{y_t^{(1)}\}_{t=-\infty}^{\infty}$ on a time series process, consider constructing a vector $\mathbf{x}_t^{(1)}$ associated with date t. This vector consists of the $[j+1]$ most recent observations on y as of date t for that realization:

$$x_t^{(1)} = \begin{bmatrix} y_t^{(1)} \\ y_{t-1}^{(1)} \\ \vdots \\ y_{t-j}^{(1)} \end{bmatrix} \tag{68}$$

We think of each realization $\{y_t^{(1)}\}_{t=-\infty}^{\infty}$ as generating one particular value of the vector $\mathbf{x}_t$ and want to calculate the probability distribution of this vector $\mathbf{x}_t^{(i)}$ across realizations i. This distribution is the joint distribution of $(Y_t, Y_{t-1}, \ldots, Y_{t-j})$. From this distribution we can calculate the following quantity:

$$\begin{aligned}\gamma_{jt} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(y_t-\mu_t)(y_{t-j}-\mu_{t-j})\times \\ \times f_{Y_t,Y_{t-1},\ldots,Y_{t-j}}(y_t, y_{t-1}, \ldots, y_{t-j})dy_t dy_{y-1}\ldots dy_{t-j} = \\ = E(Y_t-\mu_t)(Y_{t-j}-\mu_{t-j}).\end{aligned} \tag{69}$$

Note that it has the form of a covariance between two variables, thus it could be described as the covariance of Y_t with its own lagged value; hence, the term autocovariance is used. Note further that the 0th autocovariance is just the variance of Y_t.

So we come to the following

Definition. The covariance between y_t and its value y_{t+j}, separated by j intervals of time, is called the autocovariance at lag j and is defined by

$$\gamma_{jt} = cov[y_t, y_{t+j}] = E[(y_t-\mu)(y_{t+j}-\mu)]. \tag{70}$$

If neither the mean μ_t, nor the autocovariance γ_{jt} depend on the date t, then the process for T_t is said to be covariance-stationary or weakly stationary.

$$E(Y_t) = \mu$$

$$E(Y_t-\mu)(Y_{t-j}-\mu) = \gamma_j$$

for all t and any j. In this case we have the following

Definition. The autocorrelation at lag j is

$$\rho_j = \frac{\gamma_j}{\sqrt{E[(y_t-\mu)]^2 E[(y_{t+j}-\mu)^2]}} = \frac{\gamma_j}{\sigma_y^2}. \tag{71}$$

as, obviously, the variance $\sigma_y^2 = \gamma_0$ is the same at time $t+j$ as at time t. Thus, the autocorrelation at lag j, that is, the correlation between y_t and y_{t+j}, is

$$\rho_j = \frac{\gamma_j}{\gamma_0} \tag{72}$$

which implies that $\rho_0 = 1$.

Many sets of data appear as time series: a monthly sequence of the quantity of goods shipped from the factory, a weekly series of the number of road accidents, hourly observations made on the yield of a chemical process, and so on. Examples of time series abound

in such fields as economics, business, engineering, the natural sciences, and the social sciences. An intrinsic feature of a time series is that, typically, adjacent observations are dependent. The nature of this dependence among observations of a time series is of considerable practical interest. Time series analysis is concerned with techniques for the analysis of this dependence. This required the development of stochastic and dynamic models for time series data (Box and Tiao (1977), Box et al. (1994), Fernandez-Macho (1997)) and the use of such models in important areas of applications. The forecasting of future values of a time series from current and past values is probably the most important one. The use at time t of available observations from a time series to forecast its value at some future tie $t+h$ can provide a basis for (1) economic and business planning, (2) production planning, (3) inventory and production control, and (4) control and optimization of industrial processes.

Formally, this approach to forecasting may be expressed as follows. Let y_t denote the value of the variable of interest in period t. Then a forecast for period $T+h$, made at the end of period T, may have the form

$$\hat{y}_{T+h} = f(y_T, y_{T-1}, \ldots), \tag{73}$$

where $f(\cdot)$ denotes some suitable function of the past observations $y_T, y_{T-1}, \ldots$. One major goal of univariate time series analysis is to specify sensible forms of functions $f(\cdot)$. In many applications, linear functions have been used so that, for example,

$$\hat{y}_{T+h} = \nu + \alpha_1 y_T + \alpha_2 y_{T-1} + \ldots. \tag{74}$$

Because linear functions are relatively easy to deal with, it makes sense to begin with forecasts that are linear functions of past observations. Let us consider a univariate time series y_t and a forecast $h = 1$ period into the future. If $f(\cdot)$ is a linear function, we have

$$\hat{y}_{T+1} = \nu + \alpha_1 y_T + \alpha_2 y_{T-1} + \ldots. \tag{75}$$

Assuming that only a finite number p, say, of past y values are used in the prediction formula, we get

$$\hat{y}_{T+1} = \nu + \alpha_1 y_T + \alpha_2 y_{T-1} + \ldots + \alpha_p y_{T-p+1}. \tag{76}$$

Of course, the true value y_{T+1} will usually not be exactly equal to the forecast $\hat{y}_{T+1}$. Let us denote the forecast error by $u_{T+1} := y_{T+1} - \hat{y}_{T+1}$ so that

$$y_{T+1} = \hat{y}_{T+1} + u_{T+1} = \nu + \alpha_1 y_T + \ldots + \alpha_p y_{T-p+1} + u_{T+1}. \tag{77}$$

Now, assuming that our numbers are realizations of random variables and that the same data generation law prevails in each period T, (**??**) has the form of an autoregressive process of order p,

$$y_t = \nu + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots + \alpha_p y_{t-p}, \tag{78}$$

where the quantities $y_t, y_{t-1}, \ldots, y_{t-p}$, and u_t are now random variables. To actually get an autoregressive (AR(p)) process we assume that the forecast errors ut for different periods are uncorrelated, that is, u_t and u_s are uncorrelated for $s \neq t$. In other words, we assume that all useful information in the past y_ts is used in the forecasts so that there are no systematic forecast errors. Obviously such a model represents a tremendous simplification compared with the general form. Because of its simple structure, it enjoys great popularity in applied work.

7 Dynamic Factor model

Whenever we have a multidimensional time series, e.g., financial or economic data observed at regular time intervals, we want to describe its components with a smaller number of uncorrelated factors. As we stated before, the usual factor model of multivariate analysis cannot be applied immediately as the factor process also varies in time. For this reason, we call it "static factor model". Hence, there is a dynamic part, added to the usual linear factor model, the autoregressive process of the factors. As we noted before, the main difference between what we call "dynamic factor model" and "static factor model" is the autocorrelation structure of the common factors and of the idiosyncratic components. We still assume in the dynamic case that the factors only have a contemporaneous correlation with the observable variables, so that the only difference between the static and dynamic models is that the hypothesis $\forall\ (t,s):\ t \neq s\ E(F_t F_s') = 0$ and $\forall\ (t,s):\ t \neq s\ E(u_t u_s') = 0$ are no longer maintained in the dynamic model. However, we consider here exact factor models, so that we still assume that the processes (u_{it}) are uncorrelated with each other at all leads and lags.

Recall that in a classical factor analytic setting, it is assumed that a $p-$dimensional random vector of observations, y_t, depends linearly on a $q-$dimensional vector of unobserved common factors, f_t, and on individual or idiosyncratic components u_t. So

$$y_t = \mathbf{L} f_t + u_t, \tag{79}$$

where $y_t \in \mathbb{R}^p$, $f_t \in \mathbb{R}^q$, $\mathbf{L}$ is a $(p \times q)$ matrix of factor loadings, and the components of u_t are assumed to be uncorrelated, that is, $\mathbf{\Sigma}_u$ is a diagonal matrix. Still the main point of the dynamic model is that the components of the underlying multivariate stochastic process are, apart from noise, linear functions of the same, relatively small number of dynamic factors that can be identified with some latent driving forces of the whole process. Based on factor loadings, factors, e.g., monetary or macroeconomic ones, can be identified by an expert.

However, in case of time series variables, it is more reasonable to assume that the factors are autocorrelated. Also the idiosyncratic components u_t may be autocorrelated. Assume that

$$f_t = A_1 f_{t-1} + \ldots + A_p f_{t-p} + \eta_t \tag{80}$$

and

$$u_t = C_1 u_{t-1} + \ldots + C_p u_{t-p} + \epsilon_t. \tag{81}$$

For a time-varying dynamic factor model we consider the possibility that the factor loadings, $\mathbf{L}$, can change over time, i.e. $\mathbf{L} = \mathbf{L}_t$.

Here each idiosyncratic component is orthogonal at any lead and lag both to the common factors and to the idiosyncratic components of the other variables. This feature represents a serious weakness of the model. Thus, a new model, the Generalized Dynamic Factor Model, was introduced and analyzed (Forni et al. (2000)). What differentiated it from the dynamic factor models mentioned above, was that they were not assuming mutual orthogonality of the idiosyncratic components. In this work we consider only the basic model, leaving the analysis of the generalized model for further research.

Methods for estimating the models parameters were also developed. Geweke and Singleton (1981) gave maximum likelihood estimates of the factors, while Bánkövi et al. (1981, 1983) introduced an iteration that uses regression methods and principal components to find the factors one by one; they applied their results for Hungarian macroeconomic data spanning 1953-1979. Here we consider the improved algorithm which enables us to extract dynamic factors not only sequentially, but simultaneously. As the input of the algorithm, we have observations for an n-dimensional random vector in equidistant dates between t_1 and t_2. Here n is not necessarily larger than $t_2 - t_1 + 1$, cf. Stock and Watson (2002). For a given positive integer $k < n$ (k is usually much less than n) we are looking for uncorrelated factors satisfying both a linear and an autoregressive model.

The time lag, that is the order of the autoregressive model is the same for the factors and is in the range [1, 4]. To estimate the model's parameters we minimize a quadratic cost function on conditions concerning the orthogonality of the factors, the variances of the factors, and the weights balancing between the dynamic and the static part.

We use a linear algebraic method developed for this purpose to find a so-called compromise system of distinct symmetric matrices of the same size. This makes it possible to find factors simultaneously and hence, minimize the nonnegative objective function step by step in an outer and inner cycle. The method first introduced in Bolla et al. (1998) for finding maxima is interesting for its own right and makes it possible to obtain the factors by an exact compromise decomposition of several matrices, and hence, extends the method of principal components, instead of using sophisticated numerical algorithms.

8 Our model

The input data are $n-$dimensional observations $y(t) = (y_1(t), \ldots, y_n(t))$, where t is the time and the process is observed at discrete instances between two limits $(t = t_1, \ldots, t_2)$. For given positive integer $M < n$ we are looking for uncorrelated factors $F_1(t), ..., F_M(t)$ such that they satisfy the following model equations:

1. As in the usual linear model,

$$F_m(t) = \sum_{i=1}^{n} b_{mi} y_i(t), \quad t = t_1, \ldots, t_2; \quad m = 1, \ldots, M.$$

2. The second is the dynamic equation of the factors:

$$\hat{F}_m(t) = c_{m0} + \sum_{k=1}^{L} c_{mk} F_m(t-k), \quad t = t_1 + L, \ldots, t_2; \quad m = 1, \ldots, M,$$

where the time-lag L is a given positive integer and $\hat{F}_m(t)$ is the auto-regressive prediction of the mth factor at date t (the white-noise term is omitted, therefore we use $\hat{F}_m(t)$ instead of F_m).

3. The third is the linear prediction of the variables by the factors as in the usual factor model:

$$\hat{y}_i(t) = d_{0i} + \sum_{m=1}^{M} d_{mi} F_m(t), \quad t = t_1, \ldots, t_2; \quad i = 1, \ldots, n.$$

We want to estimate the parameters of the model: $\mathbf{B} = (b_{mi})$, $\mathbf{C} = (c_{mk})$, $\mathbf{D} = (d_{mi})$ $(m = 1, \ldots, M;\ i = 1, \ldots, n;\ k = 1, \ldots, L)$ in matrix notation (estimates of the parameters $c_{m0},\ d_{0i}$ follow from these) such that the objective function

$$w_0 \sum_{m=1}^{M} \mathrm{var}(F_m - \hat{F}_m)_L + \sum_{i=1}^{n} w_i \mathrm{var}(y_i - \hat{y}_i)$$

is minimum on the conditions for the orthogonality and variance of the factors:

$$\mathrm{cov}(F_m, F_l) = 0, \quad m \neq l; \quad \mathrm{var}(F_m) = v_m, \quad m = 1, \ldots, M,$$

where $w_0, w_1, \ldots, w_n$ are given non-negative constants (balancing between the dynamic and static part), while the positive numbers v_ms indicate the relative importance of the individual factors. Actually we can use the same weights $v_m = t_2 - t_1 + 1, \quad m = 1, \ldots, M$.

First we introduce some notations

$$\overline{y}_i = \frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} y_i(t) \tag{82}$$

the sample mean (average with respect to the time) of the ith component,

$$\mathrm{cov}(y_i, y_j) = \frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} (y_i(t) - \overline{y}_i)(y_j(t) - \overline{y}_j) \tag{83}$$

the sample covariance between the ith and jth components, while

$$\mathrm{cov}^*(y_i, y_j) = \frac{1}{t_2 - t_1} \sum_{t=t_1}^{t_2} (y_i(t) - \overline{y}_i)(y_j(t) - \overline{y}_j) \tag{84}$$

the corrected empirical covariance between them. The pairwise covariances between the factor components are zeroes in this sense. In the $i = j$ special case the covariance becomes the variance of the ith component. Taking the variance of the factor components, the superscript L above indicates that the summation with respect to the time is restricted to dates $t_1 + L, \ldots, t_2$ only.

Observe that the parameters c_{m0}, d_{0i} can be written in terms of other parameters:

$$c_{m0} = \frac{1}{t_2 - t_1 - L + 1} \sum_{t=t_1+L}^{t_2} (F_m(t) - \sum_{k=1}^{L} c_{mk} F_m(t-k)), \qquad m = 1, \ldots, M \tag{85}$$

and

$$d_{0i} = \overline{y}_i - \sum_{m=1}^{M} d_{mi} \overline{F}_m, \qquad i = 1, \ldots, n. \tag{86}$$

Thus, the parameters to be estimated are collected in the $M \times n$ matrices $\mathbf{B}, \mathbf{D}$, and in the $M \times L$ matrix $\mathbf{C}$. Let us denote by $\mathbf{b}_m \in \mathbb{R}^n$ the mth row of matrix $\mathbf{B}$ (we are going to use it as a column vector), $m = 1, \ldots, M$. We also introduce the notation

$$Y_{ij} := \mathrm{cov}(y_i, y_j), \qquad i, j = 1, \ldots, n \tag{87}$$

where $\mathbf{Y} := (Y_{ij})$ is the $n \times n$ symmetric, positive semidefinite empirical covariance matrix of the sample (sometimes it is corrected). We also define the lagged time series

$$z_i^m(t) = y_i(t) - \sum_{k=1}^{L} c_{mk} y_i(t-k), \quad t = t_1 + L, \ldots, t_2; \quad i = 1, \ldots, n; \quad m = 1, \ldots, M \tag{88}$$

and its empirical covariance matrix of entries

$$Z_{ij}^m := cov(z_i^m, z_j^m) = \frac{1}{t_2 - t_1 - L + 1} \sum_{t=t_1+L}^{t_2} (z_i^m(t) - \overline{z}_i^m)(z_j^m(t) - \overline{z}_j^m), \qquad i, j = 1, \ldots, n, \tag{89}$$

where $z_i^m = \frac{1}{t_2 - t_1 - L + 1} \sum_{t=t_1+L}^{t_2} z_i^m(t)$, $i = 1, \ldots, n$; $m = 1, \ldots, M$. Further, let $\mathbf{Z}^m = (Z_{ij}^m)$ be the $\times n$ symmetric, positive, semidefinite covariance matrix of these variables.

To write the objective function in terms of these quantities, we make the following argument:

$$F_m(t) - \hat{F}_m(t) = \sum_{j=1}^{n} b_{mj} z_j^m(t) - c_{m0},$$
$$var(F_m - \hat{F}_m)_L = \mathbf{b}_m^T \mathbf{Z}^m \mathbf{b}_m. \tag{90}$$

Because of the constraints of the model we have

$$\mathrm{var}(F_m) = \mathbf{b}_m^T \mathbf{Y} \mathbf{b}_m, \qquad m = 1, \ldots, M \tag{91}$$

and

$$\mathrm{cov}(y_i, F_m) = \sum_{j=1}^{n} b_{mj} Y_{ij}, \qquad i = 1, \ldots, n; \qquad m = 1, \ldots, M. \tag{92}$$

Further, due to the orthogonality of the factors, and due to the linear prediction of the variables by the factors

$$\mathrm{var}(y_i - \hat{y}_i) = Y_{ii} - 2 \sum_{m=1}^{M} d_{mi} cov(Y_i, F_m) + \sum_{m=1}^{M} d_{mi}^2 v_m = Y_{ii} - 2 \sum_{m=1}^{M} d_{mi} \sum_{j=1}^{n} b_{mj} Y_{ij} + \sum_{m=1}^{M} d_{mi}^2 v_m. \tag{93}$$

With these, the objective function to be minimized

$$G(\mathbf{B}, \mathbf{C}, \mathbf{D}) = w_0 \sum_{m=1}^{M} b_m^T Z^m b_m + \sum_{i=1}^{n} w_i Y_{ii} - 2 \sum_{i=1}^{n} w_i \sum_{m=1}^{M} d_{mi} \sum_{j=1}^{n} b_{mj} Y_{ij} + \sum_{i=1}^{n} w_i \sum_{m=1}^{M} d_{mi}^2 v_m, \tag{94}$$

where the minimum is taken on the constraints

$$\mathbf{b}_m^T \mathbf{Y} \mathbf{b}_l = \delta_{ml} v_m, \quad m, l = 1, \ldots, M.$$

The procedure of finding the minimum is based on the following so-called outer cycle. Choosing an initial $\mathbf{B}$, satisfying all the constraints, the following two steps of an iteration are alternated:

1. Starting with $\mathbf{B}$ we calculate the F_ms, then we fit a linear model (solve the Gaussian normal equations) to estimate the parameters of the autoregressive model. Hence, the current value of $\mathbf{C}$ is obtained.

2. Based on this $\mathbf{C}$, we take the minimum of G($\mathbf{B}$,$\mathbf{C}$,$\mathbf{D}$) with respect to $\mathbf{B}$ and $\mathbf{D}$, while keeping $\mathbf{C}$ fixed. Here we build in a longer inner cycle to find $\mathbf{B}$. With this new $\mathbf{B}$, we return to Step 1 of the outer cycle and proceed until convergence.

In Step 2: fixing $\mathbf{C}$, the part of the objective function to be minimized in $\mathbf{B}$ and $\mathbf{D}$ is

$$F(\mathbf{B}, \mathbf{D}) = w_0 \sum_{m=1}^{M} \mathbf{b}_m^T \mathbf{Z}^m \mathbf{b}_m + \sum_{i=1}^{n} w_i \sum_{m=1}^{M} d_{mi}^2 v_m - 2 \sum_{i=1}^{n} w_i \sum_{m=1}^{M} d_{mi} \sum_{j=1}^{n} b_{mj} Y_{ij},$$

that is optimized first in $\mathbf{D}$, then in $\mathbf{B}$. For solving the first problem, we solve the equation

$$\frac{\partial f(\mathbf{B}, \mathbf{D})}{\partial d_{mi}} = 2 w_i v_m d_{mi} - 2 w_i \sum_{j=1}^{n} b_{mj} Y_{ij} = 0$$

for $\mathbf{D}$. The solution is

$$d_{mi}^{opt} = \frac{1}{v_m} \sum_{j=1}^{n} b_{mj} Y_{ij},$$

and it gives a local minimum. Hence

$$F(\mathbf{B}, \mathbf{D}^{opt}) = w_0 \sum_{m=1}^{M} \mathbf{b}_m^T \mathbf{Z}^m \mathbf{b}_m - \sum_{i=1}^{n} w_i \sum_{m=1}^{M} \frac{1}{v_m} (\sum_{j=1}^{n} b_{mj} Y_{ij})^2,$$

from which, with the $n \times n$ symmetric matrix $\mathbf{V} = (V_{jk})$ of entries $V_{jk} = \sum_{i=1}^{n} w_i Y_{ij} Y_{ik}$ and with the $n \times n$ symmetric matrix

$$\mathbf{S}_m = w_0 \mathbf{Z}^m - \frac{1}{v_m} \mathbf{V}, \quad 1, \ldots, M$$

we have

$$F(\mathbf{B}, \mathbf{D}^{opt}) = \sum_{m=1}^{M} \mathbf{b}_m^T \mathbf{S}_m \mathbf{b}_m,$$

to be minimized on the constraints for b_m's. In order to find the minimum with respect to $\mathbf{B}$, we have to transform the vectors $\mathbf{b}_1, \ldots, \mathbf{b}_m$ into an orthonormal set. Because of the constraints, the transformation

$$\mathbf{x}_m := \frac{1}{\sqrt{v_m}} \mathbf{Y}^{1/2} \mathbf{b}_m, \quad \mathbf{A}_m := v_m \mathbf{Y}^{-1/2} \mathbf{S}_m \mathbf{Y}^{-1/2}, \quad m = 1, \ldots, M$$

will result in an orthonormal set $\mathbf{x}_1, \ldots, \mathbf{x}_M \in \mathbb{R}^n$, further

$$\mathbf{b}_m^T \mathbf{S}_m \mathbf{b}_m = \mathbf{x}_m^T \mathbf{A}_m \mathbf{x}_m$$

and hence

$$F(\mathbf{B}, \mathbf{D}^{opt}) = \sum_{m=1}^{M} x_m^T A_m x_m.$$

The sum of inhomogeneous quadratic forms is minimized by the algorithm of the next section. Let $\mathbf{x}_1^{opt}, \ldots, \mathbf{x}_M^{opt}$ denote the orthonormal set giving the minimum. Inverting the first transformation of $\mathbf{x}$s, the vectors

$$\mathbf{b}_m^{opt} = \sqrt{v_m} \mathbf{Y}^{/1/2} \mathbf{x}_m^{opt}, \qquad m = 1, \ldots, M \tag{95}$$

will give the row vectors of $\mathbf{B}^{opt}$, minimizing $F(\mathbf{B}, \mathbf{D}^{opt})$.

9 Finding optima of inhomogeneuos quadratic forms

Given the $n \times n$ symmetric matrices $\mathbf{A}_1, \ldots, \mathbf{A}_k \quad (k < n)$ we are looking for an orthonormal set of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{R}^n$ for which

$$\sum_{i=1}^{k} \mathbf{x}_i^T \mathbf{A}_i \mathbf{x}_i$$

is maximum.

1. **Theoretical solution:** by Lagrange's multipliers the x_i's giving the optimum satisfy the system of linear equations

   $$A(\mathbf{X}) = \mathbf{X}\mathbf{S}$$

   with some $k \times k$ symmetric matrix $\mathbf{S}$ (its entries are the multipliers), where the $n \times k$ matrices $\mathbf{X}$ and $\mathbf{A}(\mathbf{X})$ consist of the following columns:

   $$\mathbf{X} = (x_1, \ldots, x_k), \qquad \mathbf{A}(\mathbf{X}) = (\mathbf{A}_1 x_1, \ldots, \mathbf{A}_k x_k). \tag{96}$$

Due to the constraints imposed on $x_1, \ldots, x_k$, the non-linear system of equations

$$\mathbf{X}^T\mathbf{X} = \mathbf{I}_k \tag{97}$$

must also hold. As $\mathbf{X}$ and the symmetric matrix $\mathbf{S}$ contain all together $nk + k(k+1)/2$ free parameters, while the equations above the same number of equations, the solution of the problem is expected. Transforming our system into a homogeneous system of linear equations, a non-trivial solution of it exists, if

$$|\mathbf{A} - \mathbf{I}_n \bigotimes \mathbf{S}| = 0, \tag{98}$$

where the $nk \times nk$ matrix $\mathbf{A}$ is a Kronecker-sum $\mathbf{A} = \mathbf{A}_1 \bigoplus \ldots \bigoplus \mathbf{A}_k$.

2. **Numerical solution:** starting with a matrix $\mathbf{X}^{(0)}$ of orthonormal columns, the mth step of the iteration based on the $(m-1)$th one is as follows $(m = 1, 2, \ldots)$). Take the polar decomposition of $A(\mathbf{X}^{(m-1)})$ into an $n \times k$ matrix of orthonormal columns and a $k \times k$ symmetric matrix. Let the first factor be $\mathbf{X}^{(m)}$, etc., until convergence. The polar decomposition is obtained by spectral vector decomposition.

The above iteration is easily adopted to negative semidefinite or indefinite matrices and to minima in instead of maxima in the following way. Find the minimum of

$$\sum_{i=1}^{k} \mathbf{x}_i^T \mathbf{A}_i \mathbf{x}_i$$

on constraints (link), where $\mathbf{A}_1, \ldots, \mathbf{A}_n$ are $n \times n$ symmetric matrices. Let λ_i^{max} denote the largest eigenvalues as $\mathbf{A}_1$ $(i = 1, \ldots, k)$, and set

$$\lambda := \max_{i \in \{1,\ldots,k\}} \lambda_i^{max} + \epsilon,$$

where ϵ is an arbitrary small positive constant. The matrices

$$\mathbf{A}_i' := -\mathbf{A}_i + \lambda \mathbf{I}_n, \qquad i = 1, \ldots, k$$

are positive definite and

$$\min \sum_{i=1}^{k} \mathbf{x}_i^T \mathbf{A}_i \mathbf{x}_i = -\max \sum_{i=1}^{k} \mathbf{x}_i^T (-\mathbf{A}_i) \mathbf{x}_i = \max \sum_{i=1}^{k} \mathbf{x}_i^T \mathbf{A}_i' \mathbf{x}_i + \lambda k, \tag{99}$$

further, the minimum is taken on the same x_i's as the maximum in terms of $\mathbf{A}_i'$s.

10 Example

For an application of a method described above we used aggregate data of the Hungarian Statistical Office. We consider 10 highly correlated macroeconomic time series of the Hungarian Republic, registered yearly, spanning 19932007. Names and mnemonics of the components are as follows:

- Gross Domestic Product (1000 million HUF) GDP
- Number of Students in Higher Education EDU
- Number of Hospital Beds HEALTH
- Industrial Production (1000 million HUF) IND
- Agricultural Area (1000 ha) AGR
- Energy Production (petajoule) ENERGY
- Energy Import (petajoule) IMP
- Energy Export (petajoule) EXP
- National Economic Investments (1000 million HUF) INV
- Number of Publications INNOV

We extracted 3 factors out of the data, using lag length 4. As the variables were measured in different units we normalized them such that we made adjustments, where necessary, so as to produce numbers of comparable magnitude in the different series; later we used the reciprocals of their standard deviations as weights $w_1, \ldots, w_n$ in the objective function (**??**). In ....., the authors use the same weights $v_m = t_2 - t_1 + 1$ $(m = 1, \ldots, k)$ for the factors. We also used these weights; furthermore, we used the suggested choice $w_0 = n/kv_m$ ensuring the equilibrium between the dynamic and static parts.

In Figure 1, the first factor demonstrates a decrease, then an increase, and reaches its peak in 1996 (when restrictions on goverment spendings and social benefits were introduced and investments started). Since 1997 this factor has made slight periodic movements. Based on Table 1, variables GDP, ENERGY, and HEALTH are mainly responsible for this factor (in the middle of the 1990s there were also reforms in the health care system).

In Figure 2, the second factor slowly increases, then decreases, with highest values around the turn of the century. The variables EDU, ENERGY, and AGR have the highest coefficients in it. Note that the number of students in higher education steadily increased in the 1990's, however, since the beginning of the century the interest in some areas of study has dropped as people with higher degrees had difficulties finding jobs.

As Figure 3 demonstrates, the third factor is somewhat antipodal to the first one, with highest absolute value coefficients in GDP, ENERGY, and HEALTH; further, it shows smaller fluctuations. Future analysis is required to obtain a reasonable explanation for this phenomenon. Possibly, only the first two factors are significant, while the next ones are dampened dummies of them. We remark that in our model k is, in fact, the maximum number of factors, which does not contradict to certain rank conditions, see e.g., Deistler and coauthors (2005, 2007). The actual number of factors can be less, depending on the least square errors and practical considerations; it is an expert's job to decide how many factors to retain.

The coefficients of matrices $\mathbf{B}$, $\mathbf{D}$, and $\mathbf{C}$ are shown in Tables 1, 2, and 3, respectively. The relatively high constant terms in the linear prediction of the components by the factors (see Table 2) refer to "small" communalities. However, the constant coefficients in the autoregressive model are small (see Table 3) and the coefficient belonging to lag 2 is the largest in all the three factors. Notice that since 1990, different governments have changed each other in every 4 years, and lag 2 corresponds to the mid-period, when the measures introduced by the new government probably had the higher impact on the economy.

We also made predictions for the factors for 2 years ahead by means of matrix $\mathbf{C}$. The predicted factor values for 2008 and 2009 are illustrated by dashed lines and they show decline in all the three factors, possibly indicating the evolving economic crisis. Based on matrix $\mathbf{D}$, we predicted the variables by the factors for the period 1993-2007 and calculated the static part of the objective function, which represents one possible source of error in the algorithm. We also forecasted the components for 2008 and 2009 based on the predicted values of the factors. Data for 2009 are not available yet, however, the 2008's estimates showed a good fit to the factual data in case of most variables. We found that the squared error 1.16 of this only year is comparable to the cumulated error 11.54 of 15 years.

11 Conclusion

In this work we have reviewed the historical development of Factor Analysis and theory of dimension reduction. The most important questions arising in a factorial study were highlighted. Also we collected the most effective methods and algorithms for factor extraction and interpretation. The second part of the work was devoted to the applications of Factor Analysis in economics, to Dynamic factor models particularly. One model was described precisely and a new method (Bolla et.al) was implemented for extracting of the factors for Hungarian macroeconomic data spanning 1991-2007. Interesting relations between the factors and economic indicators were found. We focused on the first three factors, explained them, and made a prediction for 2 years ahead based on the matrix of factor loadings. The future research could be done on the generalized dynamic factor models, which was mentioned in the work, as it might better reflect the real-life economic processes. Of course, more empirical work is necessary to assess the potentials and pitfalls of dynamic factor models in empirical macroeconomic.

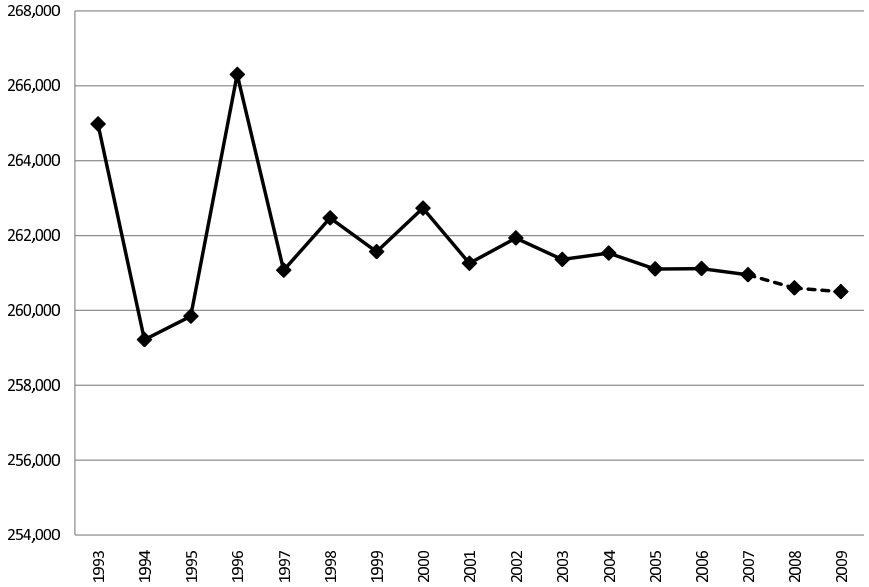

Figure 1: Dynamic Factor 1

	Factor 1	**Factor 2**	**Factor 3**
GDP	38.324	-2.541	-6.116
EDU	-1.775	5.725	0.015
HEALTH	10.166	0.837	-1.650
IND	-0.261	0.255	-0.107
AGR	6.146	2.919	-1.124
ENERGY	24.082	4.592	-4.054
IMP	1.560	-1.209	-0.213
EXP	-3.907	-0.233	0.615
INV	2.864	0.038	-0.510
INNOV	-0.608	0.197	0.089

Table 1: Factors Expressed in Terms of the Components (matrix **B**)

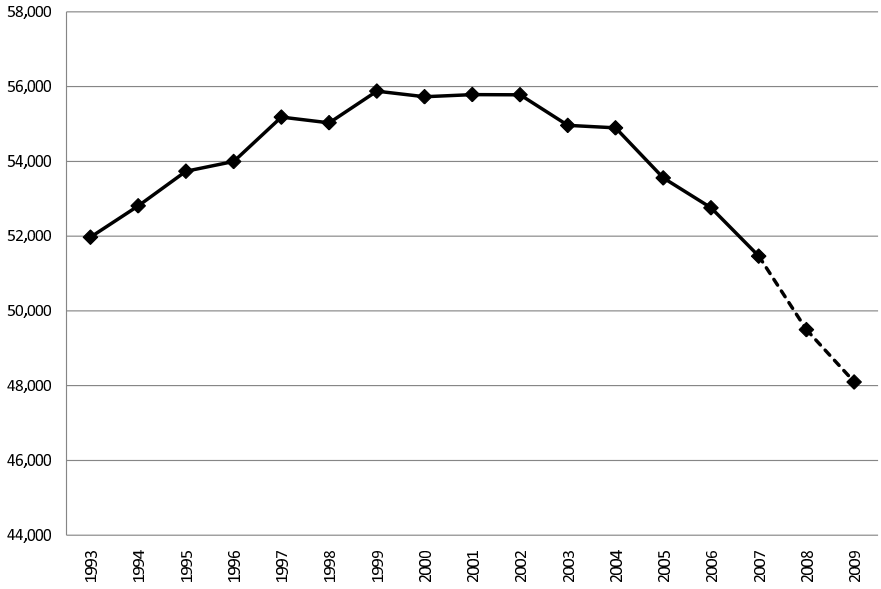

Figure 2: Dynamic Factor 2

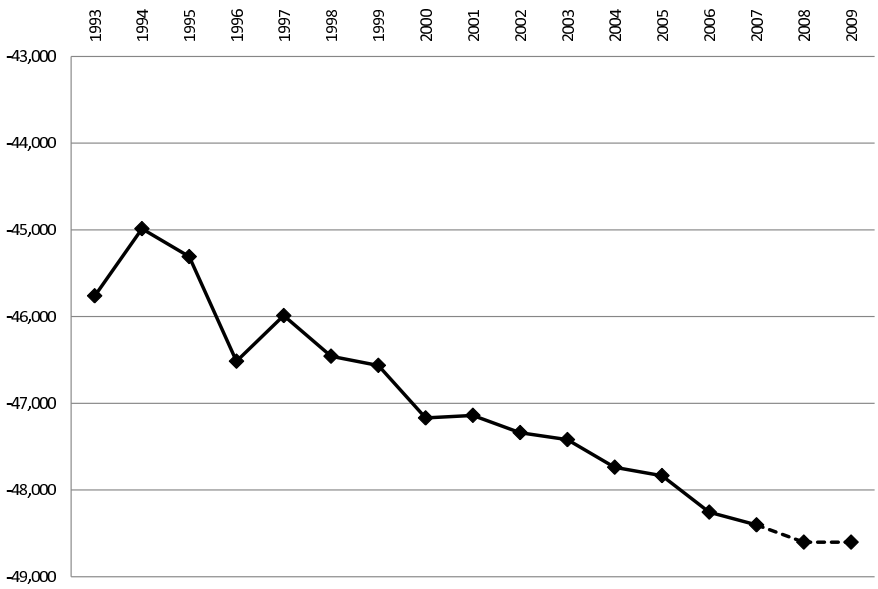

Figure 3: Dynamic Factor 3

	Factor 1	Factor 2	Factor 3	Constant
GDP	-0.108	-0.025	-0.677	-0.670
EDU	-0.142	0.145	-0.877	-8.637
HEALTH	0.115	-0.132	0.656	16.250
IND	-0.898	-0.187	-5.784	-14.690
AGR	0.021	0.005	0.137	6.809
ENERGY	0.085	-0.038	0.543	10.055
IMP	-0.098	-0.152	-0.868	0.311
EXP	-0.516	-0.931	-1.840	109.915
INV	-0.209	0.026	-1.341	-6.779
INNOV	-0.061	0.121	-0.484	-9.867

Table 2: Components Estimated by the Factors (matrix **D**)

Lag	Factor 1	Factor 2	Factor 3
0	-0.000	0.001	-0.000
1	0.069	0.283	0.117
2	0.473	1.644	0.495
3	0.205	0.229	0.141
4	0.251	-1.168	0.258

Table 3: Dynamic Equations of the Factors (matrix **C**)

References

[1] Bánkövi, G., Veliczky, J., Ziermann, M., (1982). Multivariate time series analysis and forecast. In: Grossmann, V., Pflug, G., Wertz, W. (Eds.), Probability and Statistical Inference. D. Reidel Publishing Company, Dordrecht, Holland, pp. 29–34.

[2] Bánkövi, G., Veliczky, J., Ziermann, M., (1983). Estimating and forecasting dynamic economic relations on the basis of multiple time series. Zeitschrift für Angewandte Mathematik und Mechanik 63, pp. 398–399.

[3] Bartholomew, D.J., (2007). Three Faces of Factor Analysis, in Cudeck, R., MacCallum R.C., Factor Analysis at 100, Historical Developments and Future Directions. Lawrence Erlbaum Associates, Publishers, Inc.

[4] Breitung, J., Eickmeier, S., (2005). Dymanic Factor Madels. Economic Studies, Series 1, No.38, Discussion paper.

[5] Bolla, M., Michaletzky, G., Tusnády, G., Ziermann, M., (1998). Extrema of sums of heterogeneous quadratic forms. Linear Algebra and its Applications 269, p. 331–365.

[6] Box, G. E. P., Tiao, G. C., (1977). A canonical analysis of multiple time series. Biometrika 64 (2), pp. 355–365.

[7] Box, G.E., Jenkins, G.M., Reinsel, G.C., (1994). Time Series Analysis: Forecasting and Control. Prentice-Hall International, Inc.

[8] Child, D., (1970). The Essentials of Factor Analysis. Holt, Rinehart and Winston Ltd.

[9] DeCoster, J., (1998). Overview of Factor Analysis, from http://www.stathelp.com/notes.html.

[10] Deistler, M., Hamann, E., (2005). Identification of factor models for forecasting returns. Journal of Financial Econometrics 3 (2), pp. 256–281.

[11] Deistler, M., Zinner, C., (2007). Modelling high-dimensional time series by generalized linear dynamic factor models: an introductory survey. Communications in Information and Systems 7 (2), pp. 153–166.

[12] Denton, F. T., (1978). Single-equation estimators and aggregation restrictions when equations have the same set of regressors. Journal of Econometrics 8, pp.173–179.

[13] Doz, C., Lenglart, F., (2001). Dynamic Factor Analysis: Estimation and Test with an Application to European Business Surveys. Paper presented at the CEPR/Banca d'Italia Conference, Rome.

[14] Forni, M., Hallin, M., Lippi, M., Reichlin, L., (2000). The generalized dynamic factor model: identification and estimation, from http://www.eabcn.org/research/documents/rev3.pdf

[15] Forni, M., Lippi, M., (2001). The Generalized Dynamic Factor Model: Representation Theory. Econometric Theory, Vol. 17, No. 6, pp. 1113–1141

[16] Fernandez-Macho, F.J., (1997). A Dynamic Factor Models for Economic Time Series. Kybernetika, vol.33, No.6, pp. 583–606.

[17] Geweke, J. F., (1977). The dynamic factor analysis of economic time series. In: Aigner, D., Goldberger, A. (Eds.), Latent Variables in Socio-economic Models. North-Holland, Amsterdam, pp. 365–382.

[18] Geweke, J. F., Singleton, K. J., (1981). Maximum likelihood "confirmatory" factor analysis of economic time series. International Economic Review 22, pp. 37–54.

[19] Hamilton, J.D., (1994). Time Series Analysis. Princeton University Press, Princeton, New Jersey.

[20] Lutkepohl,H.,(1993). Introduction to Multiple Time Series Analysis, 2nd Edition. Springer-Verlag Berlin.

[21] Johnson, R.A., Wichern, D.W., (2002). Applied Multivariate Statistical Analysis, 5th Edition. Prentice-Hall, Inc.

[22] Kline, P., (1994). An Easy Guide to factor Analysis. Rouledge, London.

[23] Lawley, D. N., Maxwell, A. E., (1963). Factor Analysis as a Statistical Method, 2nd edition (1971). Butterworth, London.

[24] Mardia, K.V., Kent, J.T., Bibby, J.M., (1979). Multivariate Analysis. Academic Press, Reprinted 2003.

[25] Stock, J. H., Watson, M. W., (2002). Forecasting using principal components from a large number of predictors. Journal of the American Statistical Association 97 (460), pp. 1167–1179.

[26] Tsay, R.S., (2005). Analysis of Financial Time Series. John Wiley and Sons, Inc., London.