

Underlying Dimensions of Knowledge Assessment: Factor Analysis of KAM Data

Derek Chen
The World Bank

Kishore Gawande
Texas A&M University

November 2006

Abstract: The Knowledge Assessment Methodology (KAM) database measures variables that may be used to provide an assessment of countries' readiness for the knowledge economy, and has many policy uses. Formal analysis employing KAM data is faced with the problem of which variables to choose and why. Rather than make these decisions in an ad hoc manner, we recommend factor-analytic methods to distill the information contained in the many KAM variables into a smaller set of "factors". The main objective of the paper is to quantify the factors for each country, and do so in a way that allow comparisons of the factor scores over time. We investigate both, principal components as well as true factor analytic methods, and emphasize simple structures which help to not only give a clear economic and institutional content to the factors but also allow comparisons over time.

1. Introduction

In order to facilitate the attempt of countries to make the transition to the knowledge economy, the Knowledge Assessment Methodology (KAM) was developed at the World Bank (Chen and Dahlman, 2004, 2005). It is designed to provide an assessment of countries' readiness for the knowledge economy, and identifies sectors or areas in which policymakers should focus their attention and make future investments. The KAM is currently being widely used both internally and externally to the World Bank, and frequently facilitates engagements and policy discussions with government officials from client countries.

The KAM database includes variables such as tariff and non-tariff barriers, regulatory quality, rule of law, adult literacy rate, secondary enrolment, tertiary enrolment, researchers in R&D, patent applications granted by the USPTO, scientific and technical journal articles, telephones, computers, and internet users.¹ They are constructed for over 150 countries, and are available at different points in time.

Any formal analysis employing KAM data must confront the problem of which variables to choose and why. Rather than make these decisions in an ad hoc manner, we recommend “reducing” the set of KAM variables to a smaller set of variables without losing information contained in the full set of variables. Factor-analytic methods are concerned with precisely this problem – reducing the data in a way that parsimoniously represent essentially the same information contained in the many variables. The parsimonious set of variables is the set of “factors” to which the data in the large number of variables is reduced.

Our main objective in undertaking the factor analysis is to quantify the factors for each country, that is, compute “factor scores” on each factor. Importantly, we wish to accomplish

¹Source: The Knowledge Assessment Methodology (KAM) website (www.worldbank.org/kam).

this in a way that allow comparisons of the factor scores over time. To this end, the paper details four issues in the factor analysis of the KAM data in detail. The first is whether the KAM data should be factor-analyzed and what factor-analytic method may be most appropriate; the second is determining the optimal dimensionality of the data, that is, the number of factors to which the data may be adequately reduced; the third, and perhaps most important, is giving clear meaning to the factors. Each of the above issues is treated exhaustively in the paper.

If subsets of variables are correlated, then depending on the extent of the correlation, factor analysis is worth doing. A formal test shows that the KAM data are not just amenable to factor analysis but they greatly benefit from it. There is enough inter-correlations among the variables that the real information in the data can be distilled down to a smaller number of dimensions.

What is the optimal dimensionality to which the information contained in the variables can be reduced? Depending on the factor analytic method that is chosen, the answer is different. For example, in principal components analysis this is determined by the number of principal components required to explain, say, 95% of the total variance in the data. In “true” factor analysis (which we estimate using maximum likelihood) a formal chi-squared test or information criteria that measures fit in terms of explained intercorrelations, not just variance, are used to determine optimal dimensionality.

The most important contribution of the paper is that it gives economic and institutional meaning to the dimensions, whether they go by the name of “factors” or “principal components”. Ultimately, we want the factor analysis to be a policy tool to indicate warning signals about the essential health of countries. The tool we will use to give economic and institutional meaning to the factors are the “factor loadings”. Intuitively, these are the coef-

ficients of the regression of each variable on the factors. Thus, if one variable has a very high coefficient on one factor but not on any of the others, we say that variable loads heavily on that factor. If the data, or the information in the variables can be reduced to a smaller set of factors then what we should find is that some variables load heavily on the same factor and other variables load heavily on other factors. That is, the structure of factor loadings should be simple”. One definition of a simple structure of loadings is as follows: a structure in which any single variable loads only on one factor and minimally on the others, and that more than one variable load on one factor. We will spend considerable effort in producing simple structures, because simple structures make the economic and institutional content of the factors unambiguous and clear. We will also test for the adequacy of our simple structures.

Obviously, the preceding discussion about factors loadings as regression coefficients is meant only to set ideas because, unlike in regression analysis, the factors themselves are unknown. In other words, the factor scores, or the value that the factors take, are not known and no regression in the usual sense can be estimated. Section 2 provides the theory behind how factor scores and factor loadings are computed simultaneously within a factor analytic framework.

The paper proceeds as follows. In section 2, we outline the generic factor model. In this section, two fundamentally different methods of factor analysis – principal components analysis and true factor analysis are explained in detail. A special case of the true factor analysis, the error components method, is also discussed here. Section 3 discusses the data and sources. The analysis is carried out on twelve variables measured across 120 countries. The data are from two time periods, 1995 and a more recent vintage around 2003. The same section discusses how we impute missing data in order to cover the sample of 120 countries, not all of whom have complete data on all twelve variables. We also point to a data pitfall that

should be avoided before doing the factor analysis. Section 4 contains the empirical results and the main contribution of the paper. We analyze principal components separately from the true factor analysis results. There are three main components to this section. The first is the use of factor loadings in order to name the factors. We show that with the KAM data we are able to achieve a fairly simple structure. The second is a set of formal tests for the dimensionality. The third is another methodological pitfall, whose resolution confronts us with a tradeoff. We indicate how and why we choose to resolve this in the manner we do. The choice is obvious from our overriding objective of computing factor scores as precisely as possible. Section 5 discusses the output from this factor analysis. We use graphs to show how countries have changed their rankings on the underlying dimensions over this ten year period. We separate the analysis by unweighted versus weighted KAM data. For both data sets we graphically analyze the scores produced by principal components versus true factor analysis. Section 6 concludes.

2. Factor Analysis Models

The notation and material in this section borrows from Reymont and Joreskog (1993, Sections 2, 4). The general factor analysis model is

$$\mathbf{X}_{(N \times p)} = \mathbf{F}_{(N \times k)} \mathbf{A}'_{(k \times p)} + \mathbf{E}_{(N \times p)}, \quad (1)$$

where \mathbf{X} is the data matrix of p variables, \mathbf{F} is the matrix of $k < p$ factors, and N is the sample size. The $k \times p$ “factor loadings” matrix \mathbf{A}' is used to linearly sum the factors to predict each column of \mathbf{X} . What cannot be predicted is collected in the error matrix \mathbf{E} .

In the context of the KAM data, each column of \mathbf{X} is a measure (i.e. variable) containing

“scores” for a set of N countries. There are p such measures on which country scores have been compiled.² The individual components of \mathbf{F} are the “scores” for *common factors* since they are common to several different measures. The KAM measures are thus predicted as linear combinations of the factors. The coefficients of the factors, called the *factor loadings*, are the elements of \mathbf{A}' . For example, consider the i th measure (variable) x_i . It can be written as a regression model

$$x_i = a_{i1}f_1 + a_{i2}f_2 + \dots + a_{ik}f_k + e_i. \quad (2)$$

where f_1, \dots, f_k are the “exogenous” factors, and the coefficients a_{i1}, \dots, a_{ik} are the “loadings” contained in the i th column of \mathbf{A}' . While e_i is given the interpretation of a regression residual, in fact it is made up of the measurement error in the measure x_i plus a “specific” factor that x_i does not share in common with other measures. Thus, each of the p variables $x_i, i = 1, \dots, p$ can be written as a regression model with the factors acting as the common “exogenous” variables weighted by the coefficients a_{i1}, \dots, a_{ik} , and where e_i is the regression residual.

Written in this form makes it clear that factor analysis is a method of data-reduction. The method seeks to parsimoniously represent in a small set of variables (f_1, \dots, f_k) essentially the same information contained in a much larger set of variables (x_1, \dots, x_p). We will reduce the KAM data variables to their essential factors using two different factor-analytic methods.

The difference between model (1) and ordinary regression models is that the factors and coefficients are both unknown. That is, neither \mathbf{F} nor \mathbf{A}' is known and must be estimated.

² p need not be fixed. Factor analysis of the KAM variables may be performed separately on subsets of the KAM variables. For example, each of the four pillars of the KAM data – (i) economic and institutional regime data, (ii) education and skills data, (iii) infrastructure data, and (iv) innovation potential data – may be distilled down to one or two factors.

There is a fundamental indeterminacy in the model. If we (linearly) transform \mathbf{F} and \mathbf{A}' , respectively, as $\mathbf{F}^* = \mathbf{F} \mathbf{C}^{-1}$ and $\mathbf{A}^{*'} = \mathbf{C} \mathbf{A}'$, then (1) is equivalently written as:

$$\mathbf{X}_{(N \times p)} = \mathbf{F}^*_{(N \times k)} \mathbf{A}^{*'}_{(k \times p)} + \mathbf{E}_{(N \times p)}, \quad (3)$$

Then, by observing \mathbf{X} we cannot distinguish between these two models. This should be familiar from econometric textbook discussions on identification (e.g. Greene, 2004). Devising “simple” structures in which as many factor loading as possible are zeros, facilitates identification and interpretations of the factors. We will explore simple structures in detail. We now formally discuss the two popular methods of factor analysis that we will use: the Principal Components (PC) method and the pure factor analysis model which we estimate by maximum likelihood (ML).

Fixed versus Random Factors

A distinction is made between models that presume the factor matrix \mathbf{F} in (1) to be fixed, and models that presume \mathbf{F} to be random. The random factors model is appropriate when we want to extend our inferences to different samples (say, of individuals), while the non-random factors model is appropriate when the specific observations (here countries), and not just the model structure, are of interest. The KAM data pertain to specific countries, and are exhaustive across countries, which makes a compelling case for the use of fixed-factor models. However, if inferences from the factor analysis were to be applied to countries not in the sample, or to the same countries but in a future period, then it is advisable to use random-factor models. The likelihood function for (identified) models with random \mathbf{F} is well defined (see e.g. Anderson, 1984 p 552). Estimation of models with non-random \mathbf{F} proceeds based on least squares criteria (for which, unlike the random factors case, no distributional assumptions need be made unless statistical testing is to be done).

2.1 Principal Components analysis of the Fixed Factors model³

In Stata, estimation of the Principal components model proceeds as in a fixed factor model. Let \mathbf{Y} be the mean-removed data matrix and scaled by $1/\sqrt{N}$ so that the matrix $\mathbf{S} = \mathbf{Y}'\mathbf{Y}$ is the data covariance matrix. Consider the (non-random factor) model for \mathbf{Y} :

$$\mathbf{Y}_{(N \times p)} = \mathbf{F}_{(N \times k)}\mathbf{A}'_{(k \times p)} + \mathbf{E}_{(N \times p)}, \quad (4)$$

where \mathbf{A} is the factor loadings matrix and \mathbf{F} is the matrix containing the factor scores. Using least squares to fit the (fixed data) model implies estimating \mathbf{F} and \mathbf{A} (for a given k , see Section 4 on determining k) in order to minimize the sum of the squares of the residual matrix:

$$\mathbf{E} = \mathbf{Y} - \mathbf{F}\mathbf{A}'. \quad (5)$$

The singular value decomposition (SVD) theorem indicates that a solution with the largest k singular values $\gamma_1, \gamma_2, \dots, \gamma_k$ is given as:

$$\hat{\mathbf{F}}\hat{\mathbf{A}}' = \gamma_1\mathbf{v}_1\mathbf{u}'_1 + \gamma_2\mathbf{v}_2\mathbf{u}'_2 + \dots + \gamma_k\mathbf{v}_k\mathbf{u}'_k, \quad (6)$$

where \mathbf{u}_j is a $(p \times 1)$ vector, and \mathbf{v}_j is a $(N \times 1)$ vector, and γ_j is the j th eigenvalue of the data covariance \mathbf{S} . Define the matrices: $\mathbf{V}_k = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k]$, $\mathbf{U}_k = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k]$ and $\mathbf{\Gamma}_k = \text{diag}[\gamma_1, \gamma_2, \dots, \gamma_k]$. Their dimensionalities are: $\mathbf{V}_k : (N \times k)$, $\mathbf{U}_k : (p \times k)$, $\mathbf{\Gamma}_k : (k \times k)$. Then the solution is:

$$\hat{\mathbf{F}}\hat{\mathbf{A}}' = \mathbf{V}_k\mathbf{\Gamma}_k\mathbf{U}_k. \quad (7)$$

Note that there is not a unique solution for $\hat{\mathbf{F}}$ and $\hat{\mathbf{A}}$ individually. Our solution will be in the direction of “simple” structures for $\hat{\mathbf{A}}$.

³The principal components (PC) method is applicable to both, fixed and random factors models (Reyment and Joreskog, 1993). We focus on PC as applied to fixed factors since Stata estimates PC for the fixed factors model. We indicate how to estimate the PC model for random factors in fn. 4

Consider the following solution:

$$\hat{\mathbf{F}} = \mathbf{V}_k, \quad \hat{\mathbf{A}}' = \mathbf{\Gamma}_k \mathbf{U}_k. \quad (8)$$

Then the factor scores for the k factors, $\hat{\mathbf{F}}$, are also in standardized form with covariance equal to the identity matrix. That is, they are pairwise uncorrelated.

If \mathbf{E} is small so that \mathbf{Y} is approximated by $\hat{\mathbf{F}}\hat{\mathbf{A}}'$, then the data covariance is approximately:

$$\mathbf{S} = \mathbf{Y}'\mathbf{Y} \approx \hat{\mathbf{A}}\hat{\mathbf{F}}'\hat{\mathbf{F}}\hat{\mathbf{A}}' = \hat{\mathbf{A}}\hat{\mathbf{A}}' \quad (9)$$

The ‘‘PCA’’ routine in Stata calculates Principal Components in the following steps:

1. Compute the covariance matrix \mathbf{S} .⁴
2. Compute the k eigenvectors corresponding to the largest eigenvalues of \mathbf{S} . Arrange the eigenvectors in the $p \times k$ matrix \mathbf{U}_k .
3. Estimate factor loadings as $\hat{\mathbf{A}} = \mathbf{U}_k$
4. Estimate Factor Scores as $\hat{\mathbf{F}} = \mathbf{Z}\hat{\mathbf{A}}$, where \mathbf{Z} is the data matrix with the p variables standardized to have zero mean and unit variance.

Thus, the factor loading matrix is the set of eigenvectors corresponding to the largest k eigenvalues. This is also the factor scoring matrix. Note that Stata computes factor scores using the standardized variables \mathbf{Z} , not \mathbf{Y} . In this solution, the factors have different variances and they are not comparable (their ‘‘units’’ are different). They must be scaled by $\mathbf{\Lambda}_k^{-1/2}$ to be comparable (and have unit variance).

2.2 True Factor Analysis of Intercorrelations (using Maximum Likelihood)

⁴In our analysis we use the Stata default to analyze the data correlation matrix, which produces quantitatively somewhat different loadings and scores from the analysis of the variance matrix – known as the ‘‘scaling’’ problem of PC analysis – but qualitatively the results are close.

True factor analysis is based on the **random factors** model. While the model in the random factors case is the same as (1), the population covariance matrix is:

$$\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}' + \boldsymbol{\Psi} \quad (10)$$

if the factors are uncorrelated, and

$$\boldsymbol{\Sigma} = \mathbf{A}\mathbf{F}'\mathbf{F}\mathbf{A}' + \boldsymbol{\Psi} \quad (11)$$

if the factors are correlated. In (10) and (11) $\boldsymbol{\Psi}$ is the true error covariance matrix.⁵

In order to estimate the parameters of the model, we proceed by analyzing data that are mean-removed, so that the data covariance $\mathbf{S} = \mathbf{X}'\mathbf{X}$. We make the following assumptions about the true covariances:

$$\frac{1}{N}\mathbf{X}'\mathbf{X} \rightarrow \boldsymbol{\Sigma}, \quad \frac{1}{N}\mathbf{F}'\mathbf{F} \rightarrow \boldsymbol{\Phi}, \quad \frac{1}{N}\mathbf{F}'\mathbf{E} \rightarrow \mathbf{0}, \quad \frac{1}{N}\mathbf{E}'\mathbf{E} \rightarrow \boldsymbol{\Psi}, \quad (12)$$

that is, finite second moments and orthogonality of the error and factor score matrices. We will assume that the error covariance $\boldsymbol{\Psi}$ is diagonal, that is, measurement (and other) errors are uncorrelated across different variables. This diagonal error covariance is constant across observations (or “homoskedastic” covariance). The factors may be correlated, that is, $\boldsymbol{\Phi}$ is permitted to be non-diagonal (if the factors are uncorrelated – more on this below – then

⁵PC analysis of the random factors model is also possible, but requires the assumption that $\boldsymbol{\Psi}$ is small (that is, \mathbf{E} in (1) is small). The Unweighted Least-Squares (ULS) criteria fits the factor model so that the sum of squares of the elements of $\mathbf{S} - \mathbf{A}\mathbf{A}'$ (presuming factors are uncorrelated) is minimized. The PC solution to this problem may be computed using the following steps:

1. Compute the covariance matrix \mathbf{S} .
2. Compute the k largest eigenvalues and arrange them in a diagonal matrix $\boldsymbol{\Lambda}_k$.
3. Compute the corresponding k eigenvectors \mathbf{U}_k of \mathbf{S} . Compute $\hat{\mathbf{A}} = \mathbf{U}_k\boldsymbol{\Lambda}_k^{1/2}$. Each eigenvector is now scaled so that its length equals the corresponding eigenvalue.
4. Compute the factor scores as $\hat{\mathbf{F}} = \mathbf{Y}\hat{\mathbf{A}}\boldsymbol{\Lambda}_k^{-1}$.

While this solution is different from the fixed factor solution, it is applicable to the fixed factor case with the ULS criterion applied to the error matrix \mathbf{E} in (4) and (5).

$\Phi = \mathbf{I}$). Therefore, the population variance Σ is a function of the model parameters \mathbf{A} , Φ and Ψ .

$$\Sigma = \mathbf{A}'\Phi\mathbf{A} + \Psi. \tag{13}$$

In PC analysis of the random factors model (see fn 5), factors are determined so that they account for maximum variance of all the observed variables. Thus, the emphasis in PC analysis is on eigenvalues, because the sum of all eigenvalues *is* the total variance of all the variables. In true factor analysis, the factors are determined so that they best account for the intercorrelations of the variables. In true factor analysis, the errors are presumed to be uncorrelated with each other so that Ψ is diagonal (in PC Ψ is simply assumed to be small in the sense that $\Sigma \approx \mathbf{A}'\Phi\mathbf{A}$). The rank of $\mathbf{A}'\Phi\mathbf{A}$, and therefore of Σ is approximately k . In true factor analysis, in (13) Ψ has diagonal elements only, so that the off-diagonals of Σ are exactly equal to the off-diagonals elements of $\mathbf{A}'\Phi\mathbf{A}$, and the parameters are estimated to make the off-diagonal elements of the data correlation matrix as close as possible to the off-diagonals of $\mathbf{A}'\Phi\mathbf{A}$. The diagonal elements of Σ are equal to the sum of the diagonal elements of $\mathbf{A}'\Phi\mathbf{A}$ (the “communalities” of the variables) and those of Ψ (the “uniqueness” of the variables). The off-diagonal elements assume greater importance in true factor analysis than in PC analysis (where they are assumed away).

The ML estimates of \mathbf{A} and Ψ is based on the assumption that the error vector for observation i , \mathbf{E}_i , is multivariate normal with mean 0 and variance Ψ . The likelihood function for the multivariate data is

$$\ln|\Sigma| + \text{tr}(\mathbf{S}\Sigma^{-1}) - \ln|\mathbf{S}| - p, \tag{14}$$

which is maximized over the parameters \mathbf{A} and Ψ . Asymptotic properties of the MLE have well-defined limiting distributions which are used for testing.^{6f}

⁶In Stata, the “factor” command is used together with the “ml” option in order to estimate the parameters of the factor model.

Computing Factor Scores and Standard Errors

In order to estimate factor scores from the ML method, consider a single observation on the factor model:

$$\mathbf{x}'_{(p \times 1)} = \mathbf{A}'_{(p \times k)} \mathbf{f}'_{(k \times 1)} + \mathbf{e}'_{(p \times 1)}, \quad (15)$$

where the lower case letters denote the vector elements of their matrix counterparts in (1). We proceed as described in Anderson (1971, p. 575). The transposed vectors are column vectors. The data vector \mathbf{x}' and the factor score vector \mathbf{a}' have a joint normal distribution with mean $(\mathbf{0}', \mathbf{0}')$ and covariance:

$$\text{cov} \begin{pmatrix} \mathbf{x}' \\ \mathbf{f}' \end{pmatrix} = \begin{pmatrix} \mathbf{\Psi} + \mathbf{A}\mathbf{\Phi}\mathbf{A}' & \mathbf{A}\mathbf{\Phi} \\ \mathbf{\Phi}\mathbf{A}' & \mathbf{\Phi} \end{pmatrix}$$

. The factor scores are computed by the regression of \mathbf{f}' on \mathbf{x}' . In terms of the population parameters, this is:

$$E(\mathbf{f}'|\mathbf{x}') = \mathbf{\Phi}\mathbf{A}'(\mathbf{\Psi} + \mathbf{A}\mathbf{\Phi}\mathbf{A}')^{-1} \mathbf{x}'. \quad (16)$$

Using the conditional variance formula, the covariance of the regression is

$$\text{cov}(\mathbf{f}'|\mathbf{x}') = \mathbf{\Phi} - \mathbf{\Phi}\mathbf{A}'(\mathbf{\Psi} + \mathbf{A}\mathbf{\Phi}\mathbf{A}')^{-1} \mathbf{A}\mathbf{\Phi}. \quad (17)$$

Replacing the parameters by their ML estimates yields an estimate of the (conditional) covariance. They may be used to test hypotheses about scores (for an observation) on different factors. The square roots of the diagonal elements of the (estimated) covariance are the standard errors of the estimated k -vector of factor scores on that observation. These standard errors are constant across observations. Dividing the factor score with the corresponding

standard error produces a t -statistic for testing statistical significance of individual factor scores. To take a simple, example, suppose the data are aggregated into a single factor, $k = 1$. Then the matrix Φ collapses to unity. Then the estimator for the (scalar) factor score is

$$E(f'|\mathbf{x}') = \mathbf{A}' (\Psi + \mathbf{A}\mathbf{A}')^{-1} \mathbf{x}', \quad (18)$$

and it's (scalar) variance is

$$\text{cov}(f'|\mathbf{x}') = 1 - \mathbf{A}' (\Psi + \mathbf{A}\mathbf{A}')^{-1} \mathbf{A}. \quad (19)$$

For this single factor case, denoting the ML parameter estimates with “hats”, the factor score (for the single observation) is computed as the conditional mean

$$E(\hat{f}'|\mathbf{x}') = \widehat{\mathbf{A}}' (\widehat{\Psi} + \widehat{\mathbf{A}}\widehat{\mathbf{A}}')^{-1} \mathbf{x}', \quad (20)$$

and its standard error is

$$\text{se}(\hat{f}'|\mathbf{x}') = \left[1 - \widehat{\mathbf{A}}' (\widehat{\Psi} + \widehat{\mathbf{A}}\widehat{\mathbf{A}}')^{-1} \widehat{\mathbf{A}} \right]^{0.5}. \quad (21)$$

2.3 Error Components Method

The error-components approach (EC) used by Kaufmann et al. (2005) to measure governance across several countries is a random-factors approach based on econometric methods developed for latent data models (see e.g. Goldberger (1972) and MIMIC models of Joreskog (1967) and Joreskog and Sorbom (1979)). The Kaufmann et al. approach is to fix the number of variables that map into a factor and then estimate score for the factor as conditional

means, conditional on parameters estimated by maximum likelihood. Thus, one major difference from the PC and ML methods of factor analysis described above is that the number of variables that map into a factor is prespecified.⁷ Thus, the number of variables p (and which ones they are) is treated as prior information. The computation of EC factor scores proceeds in two steps. First, the model parameters are estimated using maximum likelihood. Next, they are used to compute the scores as conditional means. The method also produces conditional variances which may be used to construct confidence intervals for the factor scores or for testing.

They (implicitly) consider the following factor model:

$$\mathbf{X}_{(N \times p)} = \mathbf{F}_{(N \times 1)} \beta'_{(1 \times p)} + \mathbf{E}_{(N \times p)}. \quad (22)$$

This corresponds to (1) except that the factor loading vector β takes the place of the factor loading matrix \mathbf{A} in (1). Whereas in (1) p variables mapped into k factors, here p variables map into a single factor. Note that while we have chosen to use the same notation to indicate matrix dimensions, the number of variables p may be chosen to be a specific set of variables, and not the entire data matrix at hand (as we did in the case of the ML and PC methods in which the number of factors k is determined by the data). Since in the EC method $k = 1$, the p variables may be chosen to be a “homogeneous” subset of the variables designed for mapping into that factor.

The EC likelihood function is as follows. Let α, β, σ be $(p \times 1)$ parameter vectors. As before, Ψ is defined to be the diagonal $(p \times p)$ error covariance matrix with elements $\sigma_1, \dots, \sigma_p$. Let

⁷For example, Chen and Dahlman (2005) partition the KAM variables into four “pillars”: Economic Incentive and Institutional Regime, Education and Human Resources, Innovation System, and Information Infrastructure. In the Chen-Dahlman scheme, since tariff and non-tariff barriers, regulatory quality, rule of law represent the Economic Incentive and Institutional Regime pillar, $p = 3$ for this factor.

the $(p \times p)$ matrix $\Sigma = \beta\beta' + \Psi$. Then, the likelihood function for the data is:

$$L = -0.5 \times N \times \ln|\Sigma| + \sum_{j=1}^N (x'_j - \alpha)' \Sigma^{-1} (x'_j - \alpha). \quad (23)$$

In (23) the parameter α is simply the vector of the means of the p variables in \mathbf{X} . For observation j the $1 \times p$ data vector is denoted x_j .

Denoting the ML parameter estimates with “hats”, the factor score for observation j is computed as the conditional mean (conditional on x_j)

$$\widehat{F}_j = \widehat{\beta}' \widehat{\Sigma}^{-1} (x'_j - \widehat{\alpha}), \quad (24)$$

and the standard error of this estimates is computed as

$$\widehat{se}_j = \left(1 - \widehat{\beta}' \widehat{\Sigma}^{-1} \widehat{\beta}\right)^{0.5}. \quad (25)$$

This is exactly the same as (21), where $\widehat{\mathbf{A}} = \widehat{\beta}$ (so that $(\widehat{\Psi} + \widehat{\mathbf{A}}\widehat{\mathbf{A}}' = \beta\beta' + \Psi = \Sigma)$).

Where the EC method differs from the (random) factor method estimated by ML is in the specification of the likelihood functions. Whereas in the EC method the data likelihood is maximized over the parameters (\mathbf{A}, Σ) , in the ML factor method the likelihood of the inter-correlations in the data is maximized over the parameters. In this sense, the EC method is still a variance method (driven by a squared error loss objective) while the ML factor method pays attention to the intercorrelations among the variables.

3. Data

The Knowledge Assessment Methodology (KAM) database consists of more than 80 structural and qualitative variables to measure countries' performance on how they perform as “knowledge economies”. We will use the subset of twelve variables that are used by the KAM method to compute each country's “basic scorecard”. They are: tariff and non-tariff barriers,

regulatory quality, rule of law, adult literacy rate (% age 15 and above), secondary enrolment, tertiary enrolment, researchers in R&D, patent applications granted by the USPTO, scientific and technical journal articles, telephones (mainlines + mobile phones), computers, and internet users. The KAM website (see fn 1) indicates the variety of sources from which the data are drawn. In addition to this unscaled data, we will also perform factor analysis on these variables, but now a subset of the variables will be scaled so that country size does not influence the analysis. The scaled set of variables are: tariff and non-tariff barriers, regulatory quality, rule of law, adult literacy rate (% age 15 and above), secondary enrolment, tertiary enrolment, researchers in R&D (per million population), patent applications granted by the USPTO (per million population), scientific and technical journal articles (per million population) telephones per 1000 persons (telephone mainlines + mobile phones), computers per 1000 persons, and internet users per 10,000 persons.

Data on the twelve variables are available at two points in time, one measured in 1995 and another during a more recent period, between 2002-04. We will use the term “2002” to indicate the recent data. Table 1 describes the variables and reports descriptive statistics for the twelve variable-pairs.

3.1 Missing Data Imputation

The factor analysis restricts the sample to one which has complete data on all included variables. Hence, a crucial pre-estimation step is to impute missing data in order to have as broad a coverage of countries as possible. The imputations are carried out using a simple regression of the variable with missing data, using as the independent variable a conceptually closely related variable. For example, (unscaled) `research95` has data for only 86 countries. However, the closely related `research03` has data for 95 countries. Therefore, nine observations can be additionally imputed by regressing `research95` on `research03`. The first column

of Table 2.1 shows the results of this regression for the unweighted data. The R -squared of 0.91 indicates a good fit for the imputation. Having filled these nine data points, we now have data for 95 countries for research95. That is still not enough. The next closely related variable is technical journal output in 1995 (techjour95). The second column of Table 2.1 indicates that this regression has an averagely good fit, with an R -squared of 0.70. The variable, techjour95 is statistically significant at 1%. Therefore, the two-step regression process makes available data on research95 for 120 countries.

A similar two-step regression process is used to impute missing research03 data via the regressions shown in columns 3 and 4 of Table 2.1. The last three columns in Table 2.1 impute data for computer95, computer04 and tariffs and NTBs for '95 (tntb95) using, respectively, GDP per capita for the two computer variables and tntb05 as regressors.⁸ After completing the imputations we have available data for 120 out of the 128 countries. Unavailability of data on the regressors prevents imputing missing data for the remaining 8 countries. The factor analysis is based on the sample of these 120 countries. Chen and Gawande (2006) provides details on the countries for which variables are imputed and the imputed values.

3.2 Is it Worth Doing Factor Analysis on the KAM data?

The main objective of the factor analysis is to understand whether countries have advanced their positions over the 10-year period in terms of (i) the absolute measures of the factors, and (ii) their factor score ranks vis-a-vis other countries. Before proceeding with factor analysis and the computation of factor scores, it is important to understand whether and how much we can gain from undertaking a factor analysis. The cross-country data on the twelve variables have considerable correlations among them. However, if the correlations are

⁸For imputing missing country values we use not only available data across countries but also for the ten regions including the world. There is additional information in these aggregated regions which can be brought to bear on the imputations.

driven by common underlying factors, then the factors become the main objects of interest for us.

If two variables share a common factor with other variables, their partial correlation, controlling for all remaining variables, will be small. The Kaiser-Meyer-Olkin (KMO) statistic, based on this idea, computes the ratio of (i) the sum of squared correlations of each variable in the analysis with every other variable to (ii) the same sum plus the sum of squared partial correlations of each variable with every other variable, controlling for all remaining variables. Large values for this “overall” KMO measure indicate that the partial correlations are small, that is, common underlying factors are responsible for the correlations among variables. A large value for the KMO measure indicates considerable gains from undertaking a factor analysis. Table 3 indicates that the overall KMO measure is 0.875, which provides solid support for proceeding with factor analysis of the KAM data. Further, the KMO statistic for each variable individually indicates that their high correlations are driven by underlying factors.⁹

3.3 Avoiding a Pitfall

Performing factor analyses separately on the 1995 variables and the 2002 variables is a pitfall one should avoid if the purpose of the factor analysis is to compare factor scores across the two periods. Separate analyses produce factor scores (that is, the quantity of a factor contained in each country) that are not strictly comparable. For example, in PC analysis separate analyses produces factors with different variances so that their magnitudes are not comparable (that is, their “units” are different). In order to solve this problem, we proceed as follows. First, we combine the 1995 and 2002 variables into one set of twelve variables. To be consistent, the same factor analytic method is used to combine each pair

⁹The variables in Table 3 are an amalgam of each variable-pair over the two years, see below.

of variables as is used in the factor analysis for the full set of twelve variables. For example, Computer95 and Computer04 are factor-analytically combined into one computer variable using either maximum likelihood (ML) or principal components (PC). Second, we proceed with the factor analysis of this set of 12 amalgamated variables. Third, we use the common estimate of the “scoring coefficients” matrix (see below) produced by the factor analysis, but apply that matrix separately to the 1995 and 2002 variables in order to compute separate sets of factor scores, one 1995 and one for 2002. These scores are used to analyze changes in the factors over the two periods. In this paper the scores are used ordinally to rank countries. However, making simple adjustments to the mean and standard deviation allows cardinal comparisons as well, for example in regression analyses.

In the following section we report and analyze the results from the Principal Components (PC) method and the Maximum Likelihood (ML) method. The discussion is from the ground-up and provides details about (i) why the specific number of factors are chosen in each method, (ii) the reason why we choose “simple” factor loading structures for our analyses, (iii) the reason we choose the specific method for obtaining the simple structure, and (iv) an approximation that makes the structure especially simple and is essential to achieve our objective of comparing factor scores across years (plus a chi-squared test that tests whether the approximation is statistically accurate).

4. Empirical results

4.1 Principal Components Analysis: Unweighted data

The first step in factor analysis is choosing k , or the number of factors that will fit the data “adequately”. In PC analysis, an oft-used criterion is to set k to be no less than the number required to explain 95% or more of the total variance in the data.¹⁰ Table 4.1 shows

¹⁰While this criterion is the most popular, other criteria have also been used. They are:

that six factors are required to explain at least 95% of the variance in the twelve KAM variables. Thus, we choose $k = 6$. Even though all of six factors are required to account for 95% of the variance in the data, the first factor accounts for the giant's share of the data variance. Table 4.1 indicates that the first principal component accounts for 60.2% of the total variance. We might expect that this component will also have the maximum number of large loadings among all principal components. Table 4.2 reports the loadings with $k=6$. As expected, a majority of the variables load heavily on this factor. Not only does this make it a catch-all factor, but since variables do not load on the remaining factors they have little economic content. For this reason, factor analysis has sought to design factors with "simple" structures so that all factors have meaningful content.

Simple Structure: Orthogonal and Oblique Rotations

The criteria advanced by Thurstone (1947) have been influential in producing computationally feasible methods that deliver simple structures:

- There should be at least one zero in each row of the factor loadings matrix.
- There should be many (at least k) zeros in each column of the factor matrix.
- For every pair of factors

(i) The size of individual factor loadings: the factor loadings squared (for orthogonal factors) indicate the variance of a variable accounted for by a particular factor. Factors not contributing much may be dropped. if parsimony is a driving concern, the thumb-rule proposed by Reymont and Joreskog (1993) – that there should be at least three significant loadings in each factor– may be used.

(ii) The variance explained by a factor: The sum of squared loadings for a given factor represents the information content in the factor. The ratio of the sum of squares to the trace of the correlation matrix is the proportion of total information residing in the factor. A cutoff value can be used to then determine how many factors to retain.

(iii) Significant residuals: A residual correlation matrix may be calculated after each factor has been extracted. k is determined at the point when the residual matrix consists of correlations solely due to random error determines k . The standard error of the residual correlations (estimated roughly as $1/\sqrt{N-1}$) can be used to determine whether the correlations are significantly greater than zero.

- only a few variables should have near-zero loadings on both.
- some variables should load heavily on one and not at all on the other.
- several variables should have near-zero loadings on both factors.

Two classes of methods have evolved that produce simple structures. The first class of methods – orthogonal rotation – maintains the uncorrelatedness of the factors, while the second class of methods – oblique rotations - seeks to find simple structures with correlated factors. Since the latter relax the constraint on orthogonality of factors, they are capable of producing even simpler structures than orthogonal rotations.

Technically, rotations work as follows. In the Stata code, a $k \times k$ rotation matrix \mathbf{T} rotates the factor loadings in Step 4 (see Section 2.1) so that the rotated factor loadings matrix, denoted $\hat{\mathbf{A}}_R$, is given by

$$\hat{\mathbf{A}}_R = \hat{\mathbf{A}}\mathbf{T} = \mathbf{U}_k\mathbf{T}. \quad (26)$$

If \mathbf{T} is an orthogonal transformation matrix, the rotation preserves the orthogonality of the factor score matrix \mathbf{F} . Otherwise, the rotation is oblique, that is, the factors are correlated. Table 4.3 displays the oblique rotation matrix \mathbf{T} that produces the simple structure that we use to proceed with our analysis and computations. In PC analysis, even after rotation, the total variance explained by the factors is still the same (e.g. 95.30% in the case of the unweighted KAM data). But the portion accounted by each factor is now different. As Table 4.4 shows, rotation distributes the portion of the variance explained more evenly than the unrotated factors. This is the point of the simple structure: to identify a factor associated with only few variables.

At this point it is useful to digress into a graphical analysis that makes the connection between rotation and simple structure clear. Figure A1 (see appendix) plots the unrotated factor loadings for $k = 6$ factors. A row of the factor loadings matrix, indicating how

the corresponding variable loads on each of the 6 factors, is depicted as a point in the 6-dimensional space of factors. The projection of these points on the (fifteen) possible 2-dimensional subspaces is displayed, respectively, in each panel of Figure A1. Consider the top row of five graphs in which the y-axis measures the loadings of the 12 variables on the first principal component (C1). While the C1 vs. C2 graph shows some evidence of clustering, the structure of the loadings is not very simple as we move across the row of graphs. Now compare the same row in Figure A1 with the first row in Figure A2, in which the axes have been rotated orthogonally (that is, rotating the axes while keeping the origin at the same point and maintaining the angle between the axes at 90 degrees) in order to achieve a simpler structure. This type of rotation is known as the “varimax” rotation. There is a clear separation of the loadings into two y -clusters: one set of variables (patapp, techjour, tel) projects into high y -values, that is, high C1 loadings, and the other into low y -values.

This type of simple structure is in evidence not only for loadings on C1 but on C4 (tertiary and secondary enrollment), C5(adult literacy), and C6 (tariffs & ntbs). While the C2 and C3 rows indicate the presence of clusters with high loadings (regulation quality and law load on C2 and computers and net users load on C3), the structure of loadings on these two components is not as simple as Thurstone’s ideal. Regardless, the varimax rotation has made the structure of loadings much simpler.

Can an oblique rotation that relaxes the constraint on uncorrelatedness of the principal components (i.e. the 90 degree angle between the axes) achieve an even simpler structure? Figure A3 depicts the result of an oblique rotation (known as the “Oblimin” rotation).¹¹ There is no visible difference between the orthogonal rotation results in Figure A2 and the Oblique rotation loadings in Figure A3. As Table A2 indicates the difference in the loadings is

¹¹Although in the figure the axes appear perpendicular to each other, they are not. That is done merely for convenience. Correlation between two components implies that the angle is less than 90 degrees. However what is important to us is the projection of the points on the axes.

almost negligible. In other words, in Principal Components Analysis the orthogonal rotation considerably simplifies the structure of loadings, and the oblique rotation reproduces it but does not simplify it further. However, in the True Factor analysis below (using ML) an oblique rotation produces significant improvement over the orthogonal rotation. We therefore adopt the oblique rotation results for computing factor scores.

Outside of economics, in psychometrics for example, researchers conducting exploratory factor analysis have generally assumed orthogonal factors.¹² In economics, however, there is every reason for believing that factors should be correlated. Multiple regression is prevalent in economics and political science precisely because non-experimental economic data are correlated. In order to satisfy the *ceteris paribus* assumption, considerable care is taken to include appropriate control variables. We should embrace the idea that economic data are correlated when such data are determined in general equilibrium. It is therefore almost impossible for the data to be orthogonal. Within a data class there may be strong interdependencies while across data classes these interdependencies may be weak. In that case, the assumption of partial equilibrium for each data class may be justified. In Chen and Dahlman (2004) this assumption leads the authors to think of their data classes as “pillars”. Here, we let the data decide how to form into groups. There are two related messages here. The first is that the main objective of the factor analysis is to be able to identify the underlying dimensions that the observed data purport to measure. Second, and related to this objective, there is absolutely no theoretical reason why the underlying factors should be uncorrelated. The underlying dimensions are determined by the same general equilibrium mechanism that generates measures of these factors (i.e. the variables). Theoretically, factors should be

¹²Traditionally, a clear distinction has been made between confirmatory factor analysis (CFA) and exploratory factor analysis (EFA). In CFA, if theory suggests two factors are correlated, then an oblique rotation is justified. In EFA, there is neither is there a theoretical basis for knowing how many factors there are nor whether they are correlated. In economics and political science, we argue, CFA will generally indicate correlated factors due to interdependencies of general equilibrium under which the data are generated.

correlated.¹³

Economic and Institutional Dimensions of the Data: Naming the Factors

What names are appropriate for the principal components? Table 4.5 shows that the first principal component (RC1) has the variables researchers, technical journals, and patent applications load heavily on it. Therefore, this component is named the Innovation Potential factor, since the ability of an economy to innovate is appropriately measured by these important inputs. Since RC2 has law and the quality of regulations load heavily on it, we call it the Law and Regulation factor. RC3 is named the ICP factor since computers, net users and telephones lines load heavily on it. RC4 is the Education factor since secondary and tertiary enrollments load heavily on it. RC5 is named the Literacy factor after the single variable, adult literacy. The final factor RC6 is the Openness factor because tariffs and NTBs almost entirely load on it. Of note is the fact that the unexplained variance (last column), after accounting for the six principle components, is quite small for every variable. This indicates that the factor model with six components fit the data well at the individual variable level (therefore satisfying criteria (ii) in fn 5, as well).

Computing Factor Scores

The scores on any factor indicate how much of the factor is “contained” in a particular country. We use the oblique-rotated factor loadings as the basis for our factor score computations. As indicated in (15), the unrotated principal components (eigenvectors) $\hat{\mathbf{A}}$ are transformed into the rotated components $\hat{\mathbf{A}}_R$ as $\hat{\mathbf{A}}_R = \hat{\mathbf{A}}\mathbf{T} = \mathbf{U}_k\mathbf{T}$ (Table 4.3 displays the matrix \mathbf{T}). In order to estimate the factor scores, we use the direct method (as different from the regression method, see Reymont and Joreskog, pp 223-225).¹⁴ In this method, the

¹³It is reasonable that correlations among factors should be weaker than correlations among variables measuring any single factor (else the two factor should be combined into one).

¹⁴This is different from the method used to compute the scoring matrix in the ML method below.

factor scores $\hat{\mathbf{F}}$ are computed as

$$\hat{\mathbf{F}} = \mathbf{Z}[(\hat{\mathbf{A}}'_R \hat{\mathbf{A}}_R)^{-1} \hat{\mathbf{A}}'_R]', \quad (27)$$

where $\hat{\mathbf{F}}$ is the $(n \times k)$ matrix containing factor scores on each factor for the 120 countries, and $\hat{\mathbf{Z}}$ is the $(n \times p)$ matrix containing the standardized data variables. The $(p \times k)$ scoring coefficient matrix in Table 4.6 (produced by Stata) is the transpose of the coefficients $(\hat{\mathbf{A}}'_R \hat{\mathbf{A}}_R)^{-1} \hat{\mathbf{A}}'_R$. However, before computing factor scores using (16), an important step is required to avoid another pitfall.

Avoiding a Pitfall (and trade-offs involved)

The scoring coefficients in Table 4.6 show that a few coefficients, indicated in bold, should dominate the measurement of the factor scores. In practice, however, other elements of the matrix can and do influence the computation of the factor scores, with unexpected consequences. Consider the first factor, the ICT factor, in Table 4.6. It consists of three large positive scoring coefficients (computers, internet users and telephones) and nine small coefficients, some of which are negative. These negative coefficients can actually produce contrarian factor scores. Take Angola, for example. Applying the direct method in (16) produces an ICT factor score for Angola that ranks it 70th among the 120 countries. However, when the countries in the sample are ranked individually according to the three variables that measure ICT, Angola ranks near the bottom of the list, below 115th, in all three rankings. The reason why its rank on the ICT factor score is much higher than its rank on any of the three variables is because some negative coefficients multiply into (large) negative values of the corresponding standardized variables to create positive numbers. The point is that, although the small coefficients appear innocuous, using them in a formulaic manner can lead to mismeasuring factor scores, sometimes quite poorly.

Because accurately measuring the factors scores is critically important, we take care to

produce the simplest structure possible.¹⁵ The example above indicates that despite those efforts, the structure is still not as simple as Thurstone’s ideal. Had that ideal been achieved by the oblique rotation, it would also have produced accurate factor scores. In order to overcome the pitfall, exemplified by the Angola case, we propose to keep only the leading scoring coefficients in each column while computing factor scores, and to set the remaining coefficients to zero. This scoring matrix with the embedded zeros is presented in Table 4.7. For example, in the first column we retain the first three elements of the scoring coefficient matrix in Table 4.6 that correspond to the main loadings on this factor. The remaining elements are set to zero.

This approximation may be formally tested using Anderson’s eigenvector test (Reyment and Joreskog, 1993, p. 101). In order to test whether a specific vector \mathbf{b} is equal to the eigenvector \mathbf{a}_i associated with the eigenvalue λ_i of a matrix \mathbf{S} (\mathbf{a}_i is the i th principal component of the data correlation matrix), the Anderson test statistic is:

$$\chi_{\text{eig}}^2 = (N - 1) \left[\lambda_i \mathbf{b}' \mathbf{S}^{-1} \mathbf{b} + \frac{1}{\lambda_i} \mathbf{b}' \mathbf{S} \mathbf{b} - 2 \right] \quad (28)$$

The statistic is distributed as χ^2 with $p - 1$ degrees of freedom, where p is the number of elements of the eigenvector (here $p = 12$). Inserting the eigenvector \mathbf{a} in place of \mathbf{b} in (17) results in a value of $\chi_{\text{eig}}^2 = 0$. We will use this property to adapt (17) to test for the rotated factor loadings \mathbf{A}_R which was created by transforming \mathbf{A} using the rotation matrix T , $\mathbf{A}_R = \mathbf{A}T$. Even though the columns of \mathbf{A}_R are not eigenvectors, since the columns of $\mathbf{A}_R T^{-1}$ are the original eigenvectors (17) can equivalently be written as a test statistic for the equality of the rotated vector corresponding to eigenvector \mathbf{a}_i , \mathbf{a}_{Ri} , with a specific vector \mathbf{b}_R as:

$$\chi_{\text{Reig}}^2 = (N - 1) \left[\lambda_i \mathbf{T}_i^{-1'} \mathbf{b}'_R \mathbf{S}^{-1} \mathbf{b}_R \mathbf{T}_i^{-1} + \frac{1}{\lambda_i} \mathbf{T}_i^{-1'} \mathbf{b}'_R \mathbf{S} \mathbf{b}_R \mathbf{T}_i^{-1} - 2 \right], \quad (29)$$

¹⁵The ordinal measure is a unit-free standardized factor score, relative to the median country whose score is zero.

where \mathbf{T}_i^{-1} is the column of the \mathbf{T}^{-1} that corresponds to eigenvector being tested for equality with the specific vector \mathbf{b}_R . We use (18) to test for the equality of each component (column) of the rotated factor loading matrix $\hat{\mathbf{A}}_R$ – the simplest possible principal component structure given the data – with the corresponding column of the rotated factor loading matrix with embedded zeros $\hat{\mathbf{A}}_{R0}$ – our “ideal” Thurstonian structure given the data.¹⁶ The statistic re-rotates the components back into the original unrotated eigenvector space and compares whether the simpler structure can map back closely to the unrotated loadings (which is the basis for the test statistic). Thus, the computed statistics correspond to the order of the unrotated eigenvectors. For the six columns we get the calculated chi-squared statistics, with 11 degrees of freedom, to be: 304.8, 54.6, 2.17, 35.0, 28.8, and 7.60. The critical value of 24.7 rejects equality of four of the six principal components (i.e. columns of $\hat{\mathbf{A}}_R$ and $\hat{\mathbf{A}}_{R0}$).

The trade-off that this result forces is between (a) accepting the results of the test and using the full scoring matrix to compute (sometimes unreliably) the factor scores, and (b) to proceed using a scoring matrix with zeros replacing the elements for which the corresponding loadings are small. The latter option is the one we choose for two reasons. First, we believe that while imperfect, as shown by the formal test procedure, it is a good approximation because the simple structure has delivered a clear picture of the variables that are strong measures of each factor. They come close to approximating the Thurstone ideal, and replacing the small loadings with zeros accomplishes that ideal. Second, and more important, is the overwhelming need to have consistent estimates of factor scores. As the Angola example drives home, the scores on a factor must be consistent with the underlying rankings of those variables that overwhelmingly determine the characteristics of the factors.

For these reasons, we proceed with the use of the zero-embedded scoring coefficient matrix

¹⁶This method may not be used with the ML method below because the ML method’s loadings matrix is not the matrix of eigenvectors, so there is no correspondence between the loadings matrix and the eigenvalues.

in Table 4.7 to determine the scores on the six factors. The same matrix is used to compute the 1995 scores and the 2002 scores on the six factors. so that the scores across the two periods may be compared. These factor scores will be used to depict two important features of the sample. First, the scores allow us to rank each country according to the values of each factor, for any specific period. Second, the scores from the two periods indicate how a country's relative position on a factor has changed over that time. Before performing these comparisons, we undertake a different kind of factor analysis, estimated by maximum likelihood. We will then be able to draw on a richer and robust set of results when we inspect country rankings and how they have changed.

4.2 True Factor Analysis with Maximum Likelihood: Unweighted data

The salience of many of the issues discussed while analyzing the PC results – rotation, simple structures, correlation of factors – are relevant for true factor analysis as well. Here too, our objective is to achieve the simplest structure, which for the ML estimates requires an oblique factor rotation. As was the case with PC analysis, in order to avoid inconsistent estimates and rankings in the true factor analysis, we set the factor scoring coefficients corresponding to small factor loadings to zero. The important difference from the PC analysis is that ML favors fewer factors. The communalities are reasonably high as indicated by the fairly low (below 0.35) uniqueness in the variables.

With the focus now on intercorrelations rather than variances (see (14)), the appropriate measure of fit used to assess how many factors best fit the data, is no longer the amount of total variance explained by the factors as in PC analysis. Three measures are appropriate here, a chi-squared measure of fit, denotes χ^2_{fit} , and two information-based criteria - the Bayes information criterion (BIC), and Akaike's information criterion (AIC). The first column Table 5.1 indicates the number of factors. The next five columns are related to the chi-squared fit

statistic corresponding to the number of factors in the first column. χ_{fit}^2 is distributed with $0.5[(p - k)^2 - (p + k)]$ degrees of freedom. It is used to test the hypothesis that k or less factors are required to rationalize the data. At the 1% level of significance the calculated statistic rejects $k = 1$ and $k = 2$ but fails to reject $k = 3$. The smallest k is therefore three according to this measure of fit. Another use of this statistic is to see if the *difference* in the statistic with every increase in k is “statistically significant”. Thus, going from $k = 5$ to $k = 6$ is the first increase (starting from $k = 1$) for which the change in the statistic is not significant. According to this variant, $k = 5$.

The two information criteria reward parsimony and penalize over-parameterization, with the BIC penalizing over-parameterization more strictly. The smaller the BIC and AIC, the more preferred the model. The BIC chooses $k = 4$ while the AIC chooses $k = 5$. Thus, the statistical tests conclude that we should focus our attention on no more than five and no less than four factors. We estimated the model with both four and five factors. Upon examining the simplest loading structure we found the four factor model to have cleaner economic and institutional content since the fifth factor is not distinct in the sense of clearly generating even one of the variables. That is, it consists of many small undistinguished loadings that are collectively significant but not individually so. Thus, we proceed with $k = 4$.

Table 5.2 indicates the oblique-rotated factor loading matrix with four factors (the rotation matrix \mathbf{T} is reported in Table 5.3). The oblique rotation improves upon the orthogonal (varimax) rotation and produces a simple structure. The first factor is named the ICT factor because the variables computers, internet users and telephones load heavily on this factor. Further, these three variables do not load heavily on any of the remaining factors, thus satisfying an important simple structure requirement. The second factor is named the Law, Regulation and Openness Factor because the three variables law, quality of regulation, and tariff and non-tariff barriers load heavily on this factor. These variables also do not load

heavily on the other factors. Factor 3 is named the Literacy and Education Factor because the three variables adult literacy, secondary enrollment and tertiary enrollment load heavily on this factor (and not on any other). Finally, factor four is named the Innovations Factor because the number of patent applications, research and number of articles in technical journals load heavily on this factor. Thus, Table 3 indicates a clear and simple structure of factors. These four factors define the underlying dimensions in the data, which are measured by the observed variables. That is, computers, internet users and telephones are essentially different measures of the ICT dimension, and adult literacy, secondary enrollment and tertiary enrollment are different measures of the Literacy and Education dimension.

An attractive feature of the four factors is that they account for the communalities in the variables quite well. The residual variances are small, as indicated by the last column of Table 5.2. None of the variables have a large measure of “uniqueness”. If one of the variables did, then it would mean that the error variance from a regression of that variable on the factors would be large. As a thumb rule, a uniqueness measure for a variable greater than 0.50 would indicate the presence of a unique factor, uncorrelated with the four common factors. Fortunately, the four factors rationalize our data well. Finally, just as for the PC analysis, in order to compute factor scores we use the Thurstonian scoring coefficient matrix in Table 5.5, achieved by replacing the undistinguished elements in the full scoring coefficient matrix in Table 5.4 by zero and retaining the significant loadings in each column.

4.3 Weighted data

The weighted data are different from the unweighted data only with regard to three variables: patent applications, researchers in R&D, and scientific and technical journal articles. In the weighted data these three variables are scaled by the population (in millions, see Table 1). The other nine variables are exactly the same as in the unweighted data. The scaling of

these three variables can, and does, influence the optimal number of principal components required to rationalize the weighted data. The composition as well as the meaning of the factors is somewhat different than in the unweighted case. For brevity, we refer the reader to Chen and Gawande (2006) for details such as the factor loading matrices for the weighted data runs. The methods for estimating those matrices and then using them to estimate the factor scores are exactly as the methods used to analyze the unweighted data.

The table that is different and of interest is determining the number of factors for PCA and ML, respectively, with the weighted data. With the ML method the optimal number of factors is three, one less than for the unweighted data, while in PC analysis the optimal number of components is seven, one more than for the unweighted data. In the ML estimation, although the chi-squared test and the AIC choose four factors, the fourth factor has little economic or institutional content. Therefore, we choose $k = 3$. This is also in accord with the Bayes information criterion, just as was the case with the unweighted data.

5. Analysis of the Factor Output

The main objective of the factor score computations is to use them to describe how countries rank on the basis of these factors, and how those rankings have changed over the two periods. Chen and Gawande (2006) contains a more complete analysis for 20 underdeveloped, developing, emerging, oil-rich and industrialized economies. Here we discuss these results for five countries.

Figure 1 for Albania has four panels in it. The panel on the top left depicts Albania's rank vis-a-vis the other 120 countries in the sample on each of the six principal components computed using the unweighted data. The spider chart on the top right depicts Albania's rank on the four ML factors using the unweighted data. The bottom row panel contains the weighted data counterparts to the top row. There are seven principal components and three

ML factors in this data. The green lines inside the spider chart shows how Albania ranked on each principal component or ML factor in 1995. The red line in the spider graph shows Albania's ranking in the most recent period, around 2002. If, along any factor axis, the red line graph is closer to the center than the green line, then it indicates that Albania's position relative to other countries in the sample on that factor has worsened over the decade. This unpleasant and surprising finding applies to the Literacy factor, the ICP factor, and the Education factor. Albania's ranking on the Literacy factor dropped from being near the top 25th percentile to the bottom 35th percentile over this decade. Similar deteriorations are in evidence for the ICP factor and the Education factor. Whether this decline in rankings imply that Albania degraded in absolute terms on the factor score or whether it improved, but at a far slower pace than other countries, is not obvious from the graphs. However, since we have used a common factor scoring matrix for computing factor scores for the two periods, the scores can be put to use in cardinal comparisons as well. The ML factors from the unweighted data, although fewer in number, convey the same difficult message about the change in Albania's ranking on the ICP factor and the Literacy & Education factor.

In terms of the unweighted data, Angola ranks towards the bottom of the list of 120 countries in almost all dimensions whether measured by principal components are maximum likelihood. It ranks abysmally in literacy, law, education, and innovation potential. The unweighted data may stack the odds against small countries like Angola since the variable patent applications, number of researchers and technical journal output is unscaled by population. The weighted data do indicate hope for Angola. Its rank in terms of its (scaled) patent applications is closer to the median. Its ranking on net users and (scaled) number of researchers has also increased over the 10 year period indicating the country is taking steps to keep up with the technological changes in the world.

One reason for separately analyzing the weighted and unweighted data sets is the belief

that there is a scale effect in the sheer numbers. That is, there may be threshold effects in innovation potential based on the stock of intellectual R&D and capital as measured by technical journal output, number of patent applications, number of researchers. This is the sense in which the unscaled (unweighted) data are different from the scaled (weighted) data. In addition to the obvious examples of the US, Japan and Western European countries, India and China have also demonstrated such threshold effects. On the other hand, scaling these variables by population indicates the extent to which the full technical potential of the population is being tapped. High levels of these scaled measures are also indicators of innovation potential as countries like Finland and Iceland have demonstrated in the last decade. So while there is no compelling reason that sheer numbers should be more or less important than the proportion of the population that is involved in technical pursuits, it is clear that both have led to the potential to innovate.

Argentina has, as one might expect after a major currency and banking crisis, degraded along many dimensions. In the unweighted data, it has fallen to the bottom quartile on the law dimension, as well as in openness. Rising inequality due to the recession are probably responsible for the degradations in the law dimension. The devaluation was probably not enough to make their exports competitive and therefore, while the rest of the world has cut back on trade barriers, Argentina has maintained or increased them. The four dimensional ML factors show a stark picture on the law and openness dimension. Surprisingly, Argentina has not lost its ranking in the other three dimensions. Its literacy ranking has actually increased, on innovation potential it has kept pace, and on the ICT dimension it has maintained its position. The weighted data reiterate the same messages from the unweighted data.

Brazil has made gains and presents a contrasting picture to Argentina on at least the law dimension. While its high income inequality is probably responsible for placing Brazil in

the bottom 50% percentile on the law and regulation dimension, the country has improved on this dimension during the last decade. In the four dimensional ML graph, the green line contains the red line, indicating that over this ten year period Brazil has improved its ranking on each dimension. The principal components show that its rank on the openness dimension has lowered, which probably has to do with Mercosur (Argentina and Brazil shared similar rankings on openness in 1995 – or is a result of keeping trade barriers at fixed levels while the rest of the world has liberalized). The ML graph indicates impressive gains in literacy and education in Brazil. It is probably a good bet that this trend will also lead to an increase in Brazil's rankings on the law dimension in future years (recall that the factors are correlated). The weighted data paints a similar picture.

China, being a populous country, will obviously show different rankings for weighted versus unweighted dimensions. We should be cautious about interpreting the meaning of the innovation potential factors in the unweighted versus the weighted data. In the unscaled data China ranks high on the innovation potential list because of the sheer strength of its size. The weighed data are quite a contrast along the dimension measured by researcher and technical journals. In other words, while China has a critical mass in innovation potential (which may be the reason it attracts foreign direct investment), China still has a long way to go in achieving its full potential on innovation as measured by the scaled data. If it produced patents, researchers and technical journal at the same per capita rate as the more advanced countries, China would probably be an OECD country. Such trends are already in evidence. Along each of these dimensions in the weighted data, China is already at the median of the sample and has made strides to move ahead, especially in patent applications. On other dimensions, literacy has not improved greatly. However, the ICT factor leaped from the bottom quartile to close to the median among the sample.

6. Conclusion

We factor-analyze the Knowledge Assessment Methodology (KAM) data. The KAM data was developed at the World Bank to assess countries' readiness for the knowledge economy. The data potentially draw the attention of policymakers to specific areas deserving of more attention and future investments. We factor-analyze KAM data in order to reduce those variables to their essential dimensions or factors. Our main objective in undertaking the factor analysis is to quantify the factors for each country, that is, compute factor scores on each factor. To this end, the paper details these issues in the factor analysis of the KAM data in detail – whether the KAM data should be factor-analyzed, the optimal dimensionality of the data, and giving economic and institutional meaning to the factors. We find that the KAM data are not just amenable to factor analysis but they greatly benefit from it. There is enough inter-correlations among the variables that the real information in the data can be distilled down to a smaller number of dimensions.

We use two factor analytic methods – Principal Components (PC) analysis and “true” factor analysis which we estimate using maximum likelihood (ML). While PC analysis focuses on explaining the variance in the data, the ML method seeks to explain the intercorrelations in the data. We should therefore expect the two methods to produce different results. While the results are different (PC analysis requires many more dimensions to rationalize the data than ML analysis), there are common themes.

A contribution of the paper is identifying the economic and institutional dimensions in the KAM data and measuring them for (ordinal) comparisons over time. We embrace the idea of a simple structure of the dimensions and allow these dimensions to be correlated with each other. The output from the factor analysis is used to graphically analyze how countries have changed their rankings on the underlying dimensions over the 1995-2002 period.

References

- Anderson, T. W. 1984. *An Introduction to Multivariate Statistical Analysis*. New York, Wiley.
- Bohara, A. K., A. I. Camargo, T. Grijalva, and K. Gawande. 2005. "Fundamental Dimensions Underlying the Regulation of U.S. Trade." *Journal of International Economics* 65(1): 93-125.
- Bollen, K.A., 1989. *Structural Equations with Latent Variables*. New York, NY: Wiley.
- Chen, H. C. Derek, and K. Gawande, 2006. "Underlying Dimensions of Knowledge Assessment: Factor Analysis of KAM Data". World Bank Working Paper.
- Chen, H. C. Derek, and C. J. Dahlman, 2005. "The Knowledge Economy, the KAM Methodology, and World Bank Operations." Manuscript.
- Chen, H. C. Derek, and C. J. Dahlman, 2004. "Knowledge and Development: A Cross-Section Approach." World Bank Policy Research Working Paper #3366.
- Goldberger, A., 1972. "Maximum Likelihood Estimation of Regressions Containing Unobservable Independent Variables." *International Economic Review* 13: 1-15.
- Heckman, J.J. and Snyder, J.M., 1997. "Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators." *Rand Journal of Economics* 28, S142-S189.
- Joreskog, K.G. and Sorbom, D., 1996. *LISREL 8: User's Reference Guide*. Chicago, IL: Scientific Software International Inc.

Joreskog, K.G. and Sorbom, D., 1979. *Advances in Factor Analysis and Structural Equations Models*. Cambridge, MA: Abt Books.

Joreskog, K. G., 1967, "A general approach to confirmatory maximum likelihood factor analysis", *Psychometrika* 34, 183-202.

Kaufmann, D., A. Kraay, and M. Mastruzzi, 1999. "Aggregating Governance Indicators." World Bank Policy Research Working Paper #2195.

Kaufmann, D., A. Kraay, and M. Mastruzzi, 2004. "Governance Matters III: Governance Indicators for 1996, 1998, 2000, and 2002." *World Bank Economic Review* 18: 253-287.

Lawley, D. N. and A. E. Maxwell, 1971. *Factor analysis as a statistical method*. New York, NY: American Elsevier.

Reyment, R. and Joreskog, K.G., 1993. *Applied Factor Analysis in the Natural Sciences*. Cambridge, UK: Cambridge University Press.

Rubin, D. B. and D. T. Thayer, 1982. "EM Algorithms for ML Factor Analysis". *Psychometrika*, Vol 47, No. 1, March, 1982.

Theil, H., 1971. *Principles of Econometrics*. New York, NY: John Wiley.

Table 1: Description of Variables

Unweighted data

Description of Variables	KAM#	Variables	Obs	Mean	Std. Dev.	Min	Max
Adult literacy rate (% age 15 and above) 2002 (SIMA)	h1	adultlit02	120	84.724	18.984	24.8	100
Adult literacy rate (% age 15 and above) 2002 (SIMA)	h1	adultlit95	120	81.606	20.971	19.8	100
Computers per 1,000 people 2004 (ITU)	t4	comp04	120	192	240	2	950
Computers per 1,000, 1995	t4	comp95	120	55	84	0	371
Rule of Law	g2	law04	120	0.114	1.021	-1.59	2.01
Rule of Law 2004 (WBI)	g2	law95	120	0.207	0.952	-1.38	2.05
Internet users per 10,000 people, 2004 (ITU)	t9	netusers04	120	2,137	2,258	8	8195
Internet users per 10,000 people	t9	netusers95	120	89	206	0	1390
Patent applications granted by the USPTO / mil pop. 2004 (USPTO)	i18	patapp04	120	1,450	9,253	0	94,129
Patent applications granted by the USPTO / mil pop. 2004 (USPTO)	i18	patapp95	120	932	6,256	0	64,510
Regulatory Quality 2004 (WBI)	g1	regqual04	120	0.189	0.968	-2.15	2.02
Regulatory Quality 1995 (WBI)	g1	regqual95	120	0.199	0.765	-1.7	1.95
Researchers in R&D / million (UNESCO 2002)	i7	resear03	120	50,569	205,162	9	1,943,000
Researchers in R&D / mil pop.	i7	resear95	120	37,821	127,861	8	962,700
Secondary Enrollment 2002 (SIMA)	h3	secenroll02	120		32.650	5.81	178.15
Secondary Enrollment	h3	secenroll95	120	68.594	33.666	5.40	146.32
Scientific and technical journal articles / mil pop. 2001 (SIMA)	i14	techjour01	120	5,349	20,179	0	200,870
Scientific and technical journal articles / mil pop.	i14	techjour95	120	4,799	19,961	1	202,887
Telephones per 1,000 people 2004 (mainlines+mobile phones) (ITU)	t1	tel04	120	669	550	9	1,991
Telephones per 1,000 (mainlines + mobile)	t1	tel95	120	214	236	2	909
Tertiary Enrollment 2002 (SIMA)	h4	terenrol02	120	31.333	23.200	0.31	85.66
Tertiary Enrollment	h4	terenrol95	120	23.218	18.825	0.40	87.80
Tariff & nontariff barriers 2005 (Heritage Fdn)	e4	tntb05	120	5.883	2.338	2	10
Tariff & nontariff barriers	e4	tntb95imp	120	4.917	2.617	2	10

Weighted data

Patent applications granted by the USPTO / mil pop. 2004 (USPTO)	i18	patapp04	120	23.5	55.9	0	321
Patent applications granted by the USPTO / mil pop. 2004 (USPTO)	i18	patapp95	120	13.8	37.0	0	244
Researchers in R&D / million (UNESCO 2002)	i7	resear03	120	1125	1556	1.04	7431
Researchers in R&D / mil pop.	i7	resear95	120	922	1189	0.82	4922
Scientific and technical journal articles / mil pop. 2001 (SIMA)	i14	techjour01	120	169	283	0	1160
Scientific and technical journal articles / mil pop.	i14	techjour95	120	149	264	0.20	1068

Note:

1. Data for some variables have been imputed to complete the sample of 120 countries (see Tables 2.1 and 2.2)

Table 2.1: Imputation Regressions for missing data [UWEIGHTED Data]

	resear95		resear03		comp95	comp04	tntb95
	1st imputation	2nd imputation	1st imputation	2nd imputation			
resear03	0.605 (29.24)**						
techjour95		5.405 (14.73)**					
tntb05							0.805 (26.84)**
hdi05							
gdppc					0.005 (27.36)**	0.016 (24.47)**	
techjour01				9.35 (25.48)**			
resear95			1.503 (29.24)**				
<i>N</i>	86	95	86	95	130	130	135
<i>R</i> -sq	0.91	0.7	0.91	0.87	0.85	0.82	0.84

Notes:

1. Absolute value of t statistics in parentheses -- * significant at 5%; ** significant at 1%
2. Imputations use fully available data (countries and regions).

Table 2.2: Imputation Regressions for missing data [WEIGHTED Data]

	resear95		resear03		comp95	comp04	tntb95
	1st imputation	2nd imputation	1st imputation	2nd imputation			
resear03	0.754 (30.83)**						
techjour95		4.122 (13.83)**					
tntb05							0.805 (26.84)**
hdi05							
gdppc					0.005 (27.36)**	0.016 (24.47)**	
techjour01				5.189 (18.06)**			
resear95			1.217 (30.83)**				
<i>N</i>	86	95	86	95	130	130	135
<i>R</i> -sq	0.92	0.67	0.92	0.78	0.85	0.82	0.84

Notes:

1. Absolute value of t statistics in parentheses -- * significant at 5%; ** significant at 1%
2. Imputations use fully available data (countries and regions).

Table 3: Kaiser-Meyer-Olkin measures

Variable	KMO
p_comp	0.920
p_netusers	0.927
p_tel	0.946
p_adultlit	0.851
p_secenroll	0.837
p_terenrol	0.917
p_patapp	0.715
p_resear	0.897
p_techjour	0.713
p_law	0.857
p_regqual	0.870
p_tntb	0.955
Overall	0.875

Table 4.1: Variance explained by unrotated Principal Components (PC)

	Unrotated		
	eigenvalues	Proportion	Cumulative
PC1	7.222	0.602	0.602
PC2	2.467	0.206	0.808
PC3	0.861	0.072	0.879
PC4	0.435	0.036	0.916
PC5	0.238	0.020	0.935
PC6	0.216	0.018	0.953
PC7	0.147	0.012	0.966
PC8	0.141	0.012	0.977
PC9	0.112	0.009	0.987
PC10	0.076	0.006	0.993
PC11	0.055	0.005	0.998
PC12	0.030	0.003	1.000

Table 4.2: Unrotated Principal Components (A)
Unweighted data

Variable	PC1	PC2	PC3	PC4	PC5	PC6	Unexplained
p_comp	0.334	-0.0217	-0.278	-0.2743	-0.0477	0.292	0.075
p_netusers	0.327	-0.0368	-0.2407	-0.441	-0.1385	0.359	0.059
p_tel	0.350	-0.1088	-0.1044	-0.124	0.039	0.020	0.069
p_adultlit	0.263	-0.123	0.639	0.183	0.285	0.589	0.004
p_secenroll	0.313	-0.1464	0.396	-0.1168	0.170	-0.4421	0.049
p_terenrol	0.321	-0.0175	0.329	-0.3437	-0.3939	-0.3741	0.044
p_patapp	0.155	0.559	-0.0354	0.107	-0.0482	0.122	0.046
p_resear	0.149	0.544	0.114	0.053	0.152	-0.0985	0.088
p_techjour	0.189	0.536	-0.0107	0.052	-0.0083	-0.0622	0.032
p_law	0.334	-0.1169	-0.3037	0.107	0.364	-0.2374	0.034
p_regqual	0.320	-0.144	-0.2701	0.370	0.348	-0.1327	0.052
p_tntb	0.305	-0.1404	-0.036	0.620	-0.6579	0.037	0.008

Table 4.3: PC Rotation matrix (T)

Unweighted data

	RC1	RC2	RC3	RC4	RC5	RC6
PC1	0.281	0.567	0.541	0.476	0.277	0.307
PC2	0.950	-0.2149	-0.0676	-0.1083	-0.1269	-0.1459
PC3	0.032	-0.4003	-0.418	0.502	0.661	-0.0593
PC4	0.126	0.306	-0.5366	-0.3754	0.172	0.643
PC5	0.049	0.564	-0.1582	-0.1963	0.294	-0.6825
PC6	-0.0186	-0.273	0.477	-0.5752	0.598	0.049

Table 4.4: Variance explained by PCs after Rotating

Component	Variance	Proportion
RC1	2.803	0.234
RC2	2.705	0.225
RC3	2.457	0.205
RC4	2.023	0.169
RC5	1.080	0.090
RC6	1.029	0.086

Notes:

1. Rotated principal components, termed RC_n , are now correlated.

Table 4.5: Simple PC Structure: Oblique–Rotated Principal Components (\mathbf{A}_R)
Unweighted data

	Innovation	Law&Reg	ICT	Education	Literacy	Openness	
Variable	RC1	RC2	RC3	RC4	RC5	RC6	Unexplained
p_comp	0.022	0.115	0.593	−0.0342	0.025	−0.0074	0.075
p_netusers	−0.0199	−0.0213	0.710	0.025	0.034	−0.0517	0.059
p_tel	−0.0225	0.242	0.311	0.153	0.044	0.024	0.069
p_adultlit	0.004	−0.0245	0.021	−0.0039	0.977	0.013	0.004
p_secenroll	−0.0365	0.231	−0.1609	0.628	0.132	−0.1181	0.049
p_terenrol	0.029	−0.1711	0.105	0.742	−0.0895	0.111	0.044
p_patapp	0.582	−0.0455	0.069	−0.1053	0.026	0.076	0.046
p_resear	0.578	0.051	−0.1031	0.076	0.043	−0.1146	0.088
p_techjour	0.569	0.024	0.014	0.044	−0.0535	0.016	0.032
p_law	0.009	0.639	0.087	0.044	−0.1099	−0.0535	0.034
p_regqual	0.011	0.666	−0.0211	−0.0987	0.015	0.130	0.052
p_tntb	−0.004	0.026	−0.0213	0.018	0.014	0.966	0.008

Notes:

1. Principal components are now correlated.
2. RC_n is rotated principal component n . The unrotated PCs map into RCs via the transformation or rotation matrix \mathbf{T} .

Table 4.6: PC Scoring Coefficient Matrix: $((\mathbf{A}_R' \mathbf{A}_R)^{-1} \mathbf{A}_R')$

Unweighted data

Variable	RC1	RC2	RC3	RC4	RC5	RC6
p_comp	0.027	0.054	0.580	-0.0431	0.010	-0.0013
p_netusers	-0.012	-0.0935	0.709	0.018	0.011	-0.0398
p_tel	-0.0225	0.207	0.282	0.144	0.039	0.022
p_adultlit	0.000	0.001	-0.0073	-0.0152	0.976	0.010
p_secenroll	-0.0436	0.242	-0.198	0.623	0.137	-0.1263
p_terenrol	0.033	-0.2051	0.121	0.746	-0.1073	0.121
p_patapp	0.584	-0.0604	0.082	-0.1054	0.021	0.083
p_resear	0.575	0.056	-0.1058	0.075	0.044	-0.1129
p_techjour	0.569	0.012	0.020	0.044	-0.0563	0.021
p_law	0.000	0.627	0.022	0.030	-0.0945	-0.0703
p_regqual	0.002	0.667	-0.0883	-0.1132	0.035	0.109
p_tntb	0.003	-0.0004	-0.009	0.021	0.010	0.965

Table 4.7: PC Scoring Coefficient Matrix with Embedded Zeros

Unweighted data

Variable	RC1	RC2	RC3	RC4	RC5	RC6
p_comp	0	0	0.580	0	0	0
p_netusers	0	0	0.709	0	0	0
p_tel	0	0	0.282	0	0	0
p_adultlit	0	0	0	0	0.976	0
p_secenroll	0	0	0	0.623	0	0
p_terenrol	0	0	0	0.746	0	0
p_patapp	0.584	0	0	0	0	0
p_resear	0.575	0	0	0	0	0
p_techjour	0.569	0	0	0	0	0
p_law	0	0.627	0	0	0	0
p_regqual	0	0.667	0	0	0	0
p_tntb	0	0	0	0	0	0.964

Table 5.1: ML Testing for Factors

Unweighted data

# Factors	χ^2_{fit}	df	$P(\chi^2_{\text{fit}})$	$\Delta\chi^2_{\text{fit}}$	Δdf	BIC	AIC
2	253.99	43	0.000			377.9	313.7
3	114.29	33	0.000	139.7	10	279.2	187.2
4	42.94	24	0.010	71.35	9	246.9	129.8
5	15.15	16	0.514	27.79	8	255.6	116.3
6	7.92	9	0.542	7.23	7	282.1	123.2
7	2.68	3	0.443	5.24	6	304.5	128.9

Notes:

1. df=degrees of freedom.
2. BIC=Bayes' information criterion; AIC=Akaike's information criterion.

Table 5.2: Oblique-Rotated Factor Loading Matrix (A_R)

Unweighted data

Variable	ICT	Law, Reg & Openness	Literacy & Education	Innovation Potential	Uniqueness
	F1	F2	F3	F4	
Computers	0.851	0.098	-0.003	0.060	0.097
Internet users	0.835	0.051	0.052	0.021	0.156
p_tel	0.581	0.238	0.251	-0.023	0.074
p_adultlit	-0.147	0.115	0.859	0.008	0.289
p_secenroll	0.113	0.067	0.838	-0.032	0.088
p_terenrol	0.344	-0.101	0.680	0.124	0.174
p_patapp	0.007	0.027	-0.082	0.975	0.066
p_resear	-0.055	-0.057	0.090	0.906	0.192
p_techjour	0.032	0.030	0.004	0.976	0.009
p_law	0.349	0.632	0.028	0.010	0.087
p_regqual	0.005	0.968	0.005	0.026	0.040
p_tntb	0.054	0.592	0.264	0.014	0.304

Table 5.3: Factor Rotation Matrix
Unweighted data

	F1	F2	F3	F4
F1	-0.016	-0.269	0.572	0.006
F2	0.383	-0.177	-0.196	-0.014
F3	0.704	0.829	0.671	-0.304
F4	0.598	0.457	0.430	0.953

Table 5.4: Scoring coefficients

Unweighted data

Variable	F1	F2	F3	F4
p_comp	0.413	-0.010	-0.094	-0.006
p_netusers	0.249	-0.015	-0.034	-0.005
p_tel	0.325	0.057	0.130	-0.013
p_adultlit	-0.056	0.009	0.195	0.002
p_secenroll	-0.040	-0.001	0.593	-0.001
p_terenrol	0.056	-0.035	0.231	0.006
p_patapp	-0.001	0.001	-0.075	0.114
p_resear	-0.022	-0.013	0.036	0.037
p_techjour	0.025	-0.016	0.037	0.867
p_law	0.153	0.206	-0.034	-0.010
p_regqual	-0.107	0.740	-0.042	-0.018
p_tntb	-0.009	0.058	0.050	-0.001

Table 5.5: Scoring Coefficient Matrix with Embedded Zeros

Unweighted data

	F1	F2	F3	F4
p_comp	0.413	0	0	0
p_netusers	0.249	0	0	0
p_tel	0.325	0	0	0
p_adultlit	0	0	0.195	0
p_secenroll	0	0	0.593	0
p_terenrol	0	0	0.231	0
p_patapp	0	0	0	0.114
p_resear	0	0	0	0.037
p_techjour	0	0	0	0.867
p_law	0	0.206	0	0
p_regqual	0	0.740	0	0
p_tntb	0	0.058	0	0

Figure 1: Albania

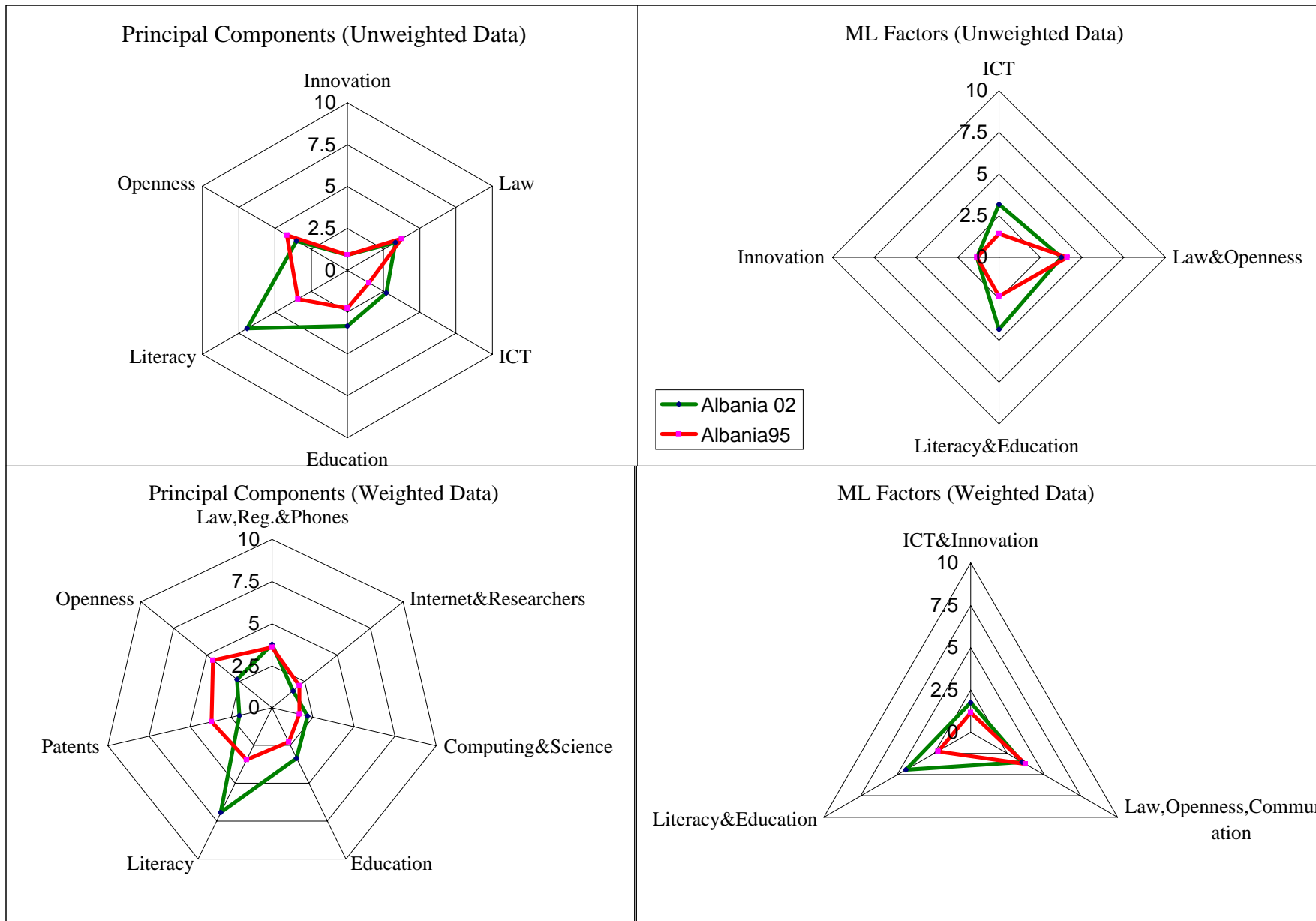


Figure 2: Angola

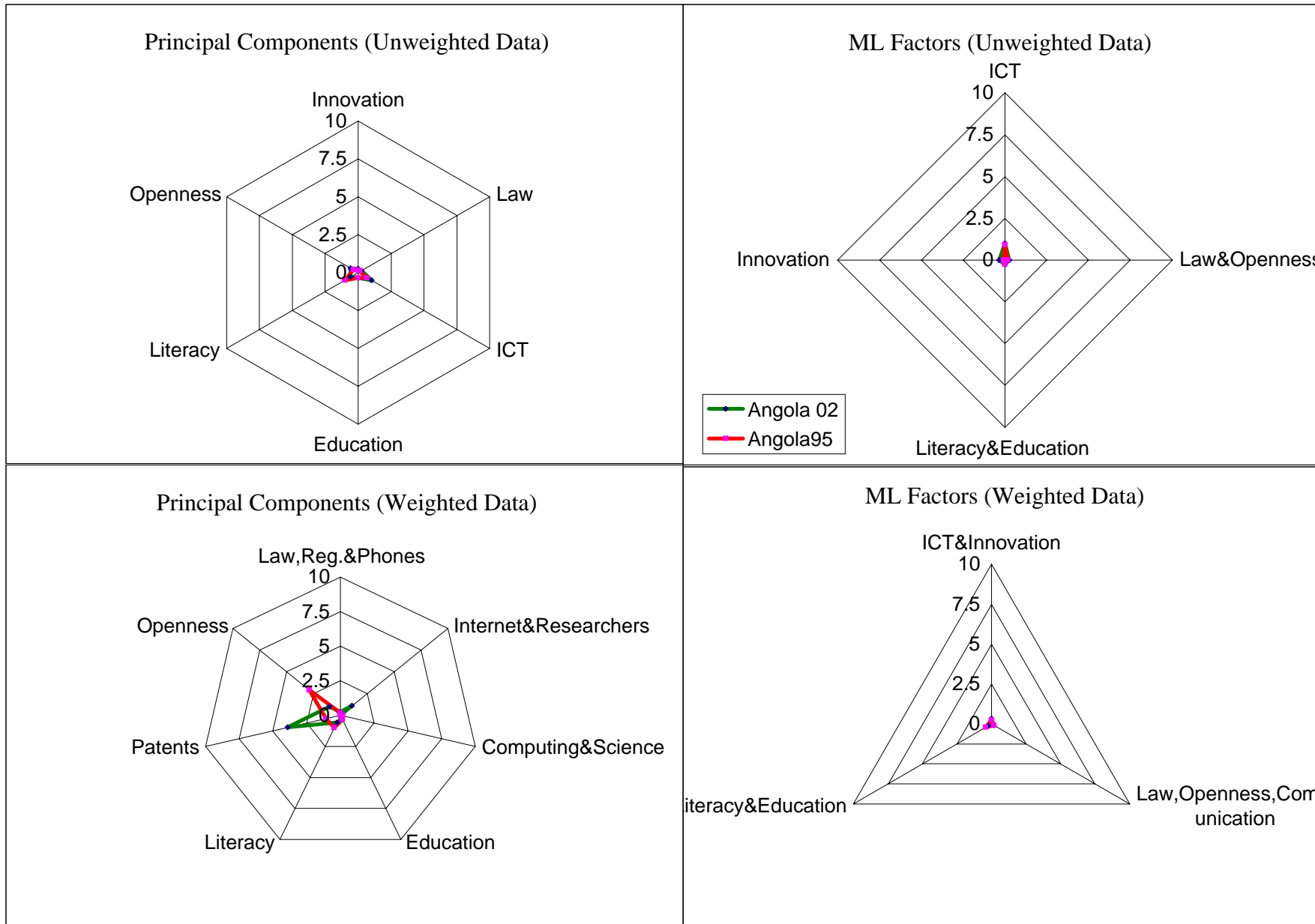


Figure 3: Argentina

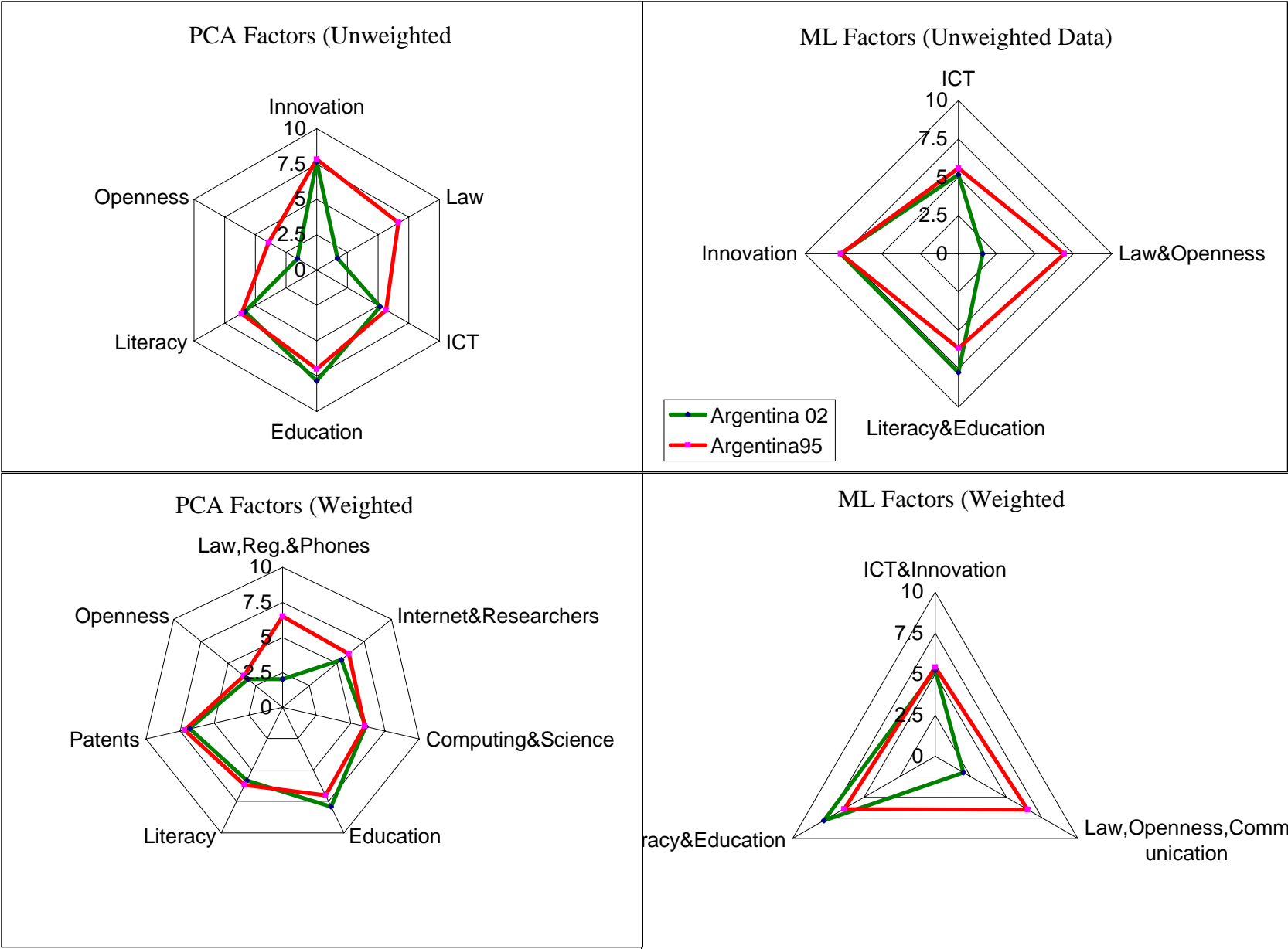


Figure 4: Brazil

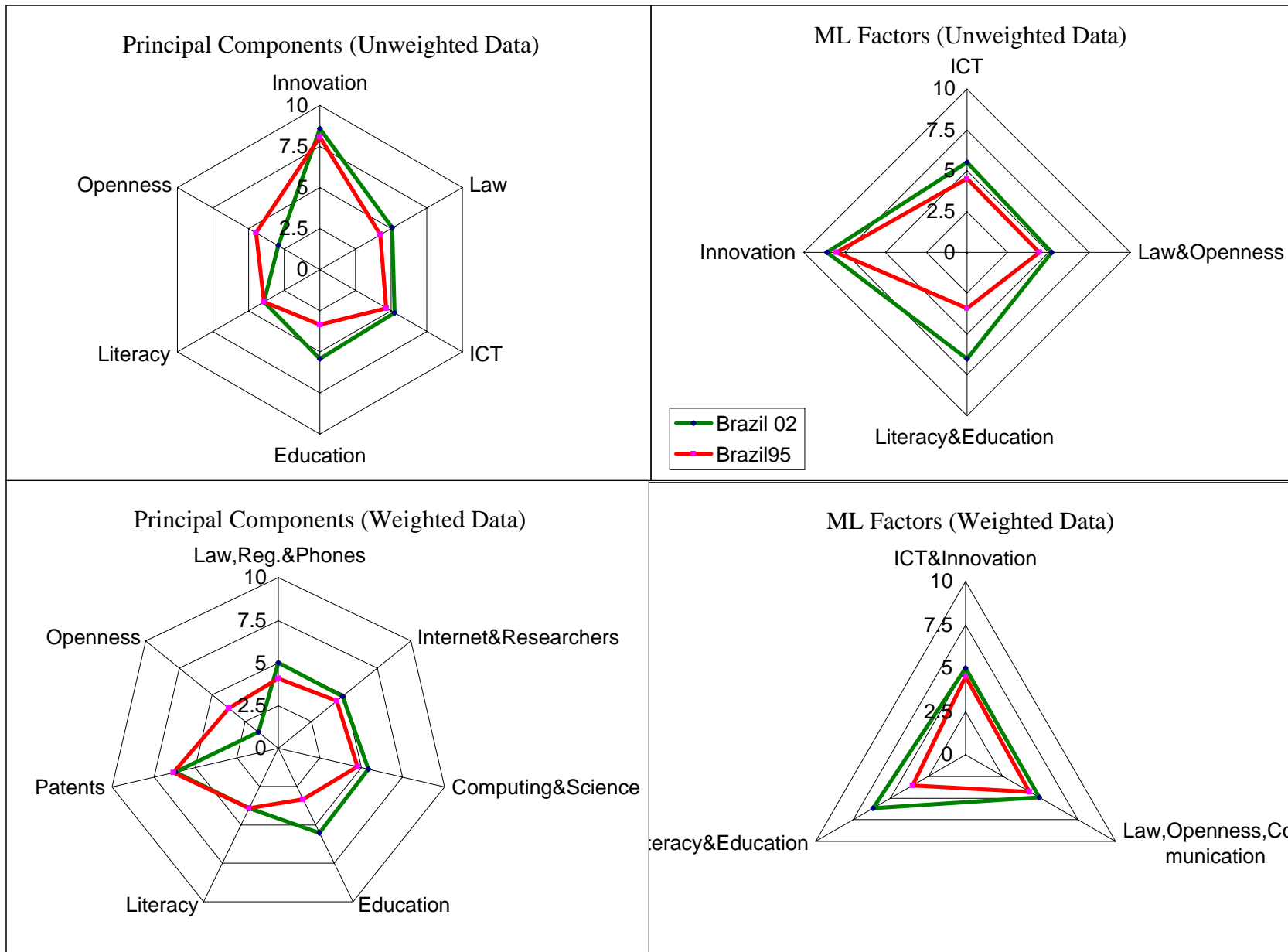
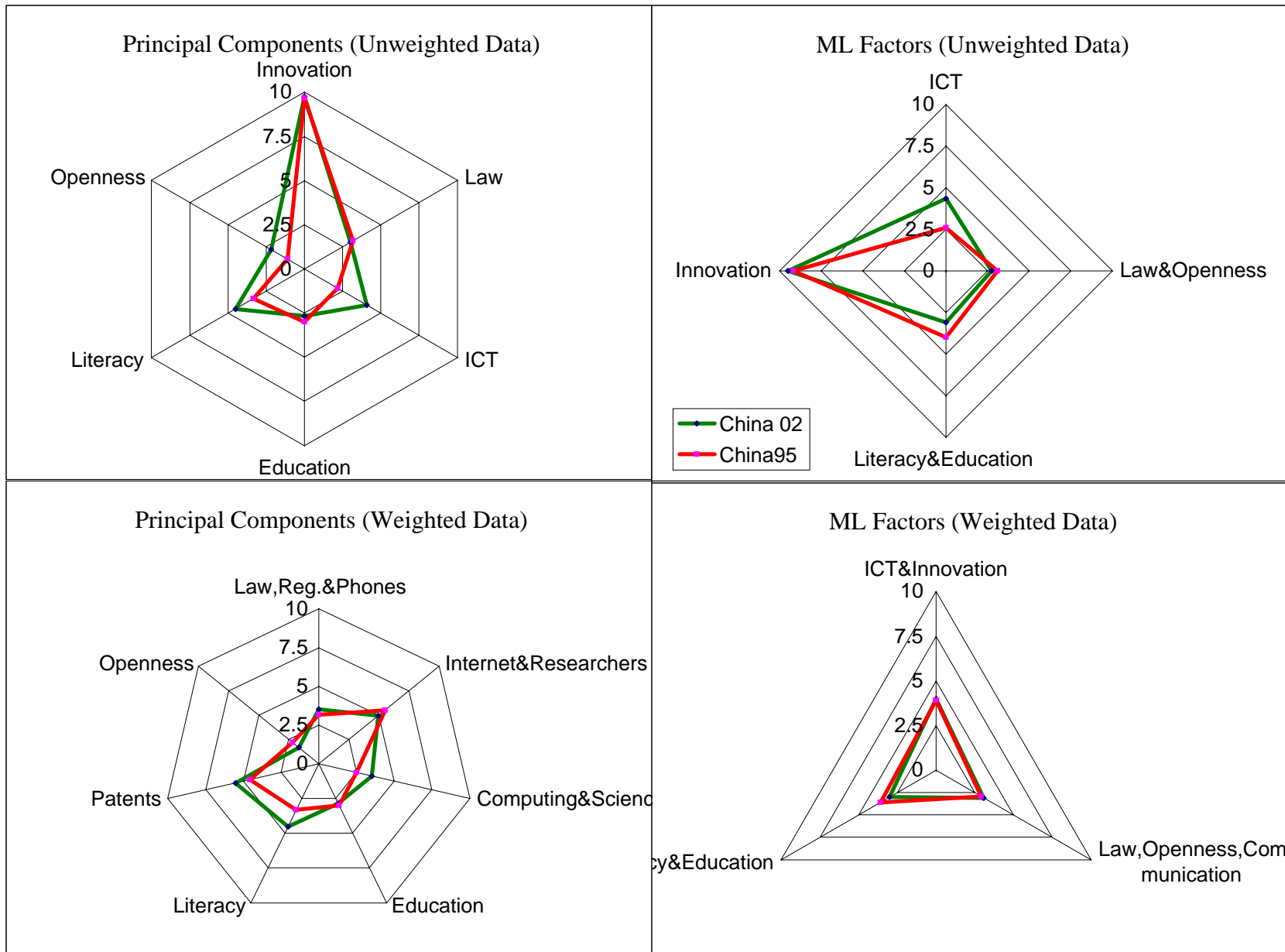


Figure 5: China



Component loadings

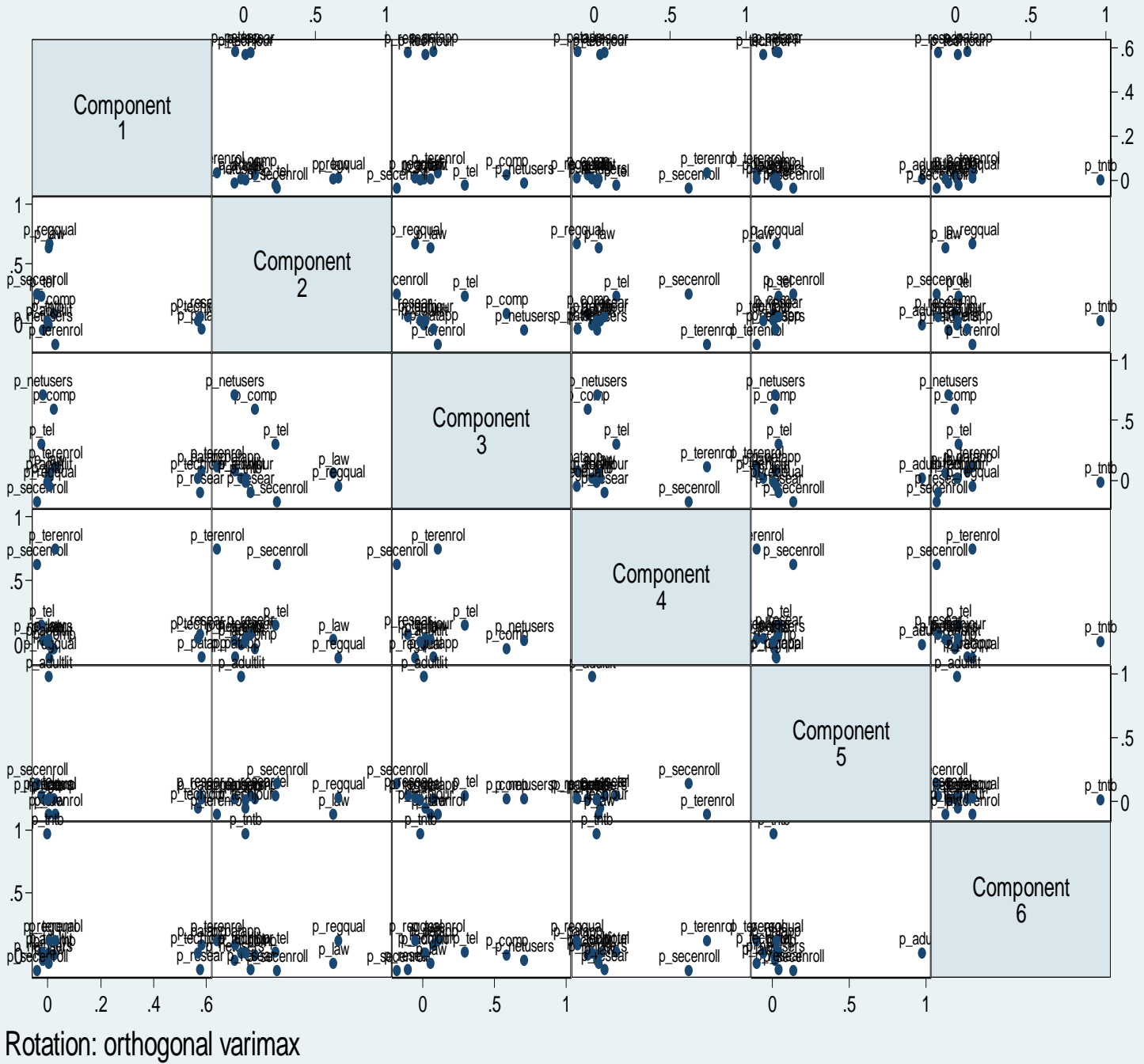


Figure A2: Orthogonal-rotated principal component plots [$k=6$]

