Political Analysis (2004) 12:338–353
doi:10.1093/pan/mph022

Bayesian Factor Analysis for Mixed Ordinal and Continuous Responses

Kevin M. Quinn
Department of Government and CBRSS, 34 Kirkland Street, Harvard University, Cambridge, MA 02138
e-mail: kevin_quinn@harvard.edu

Many situations exist in which a latent construct has both ordinal and continuous indicators. This presents a problem for the applied researcher because standard measurement models are not designed to accommodate mixed ordinal and continuous data. I address this problem by formulating a measurement model that is appropriate for such mixed multivariate responses. This model unifies standard normal theory factor analysis and item response theory models for ordinal data. I detail a Markov chain Monte Carlo algorithm for model fitting. I apply the model to cross-national data on political-economic risk and find that the model works well. Software for fitting this model is publicly available in the *MCMCpack* (Martin and Quinn 2004, "MCMCpack 0.4–8") *R* package.

1 Introduction

A great deal of work in political science makes use of latent concepts such as democracy, autocracy, commitment to the free market, political-economic risk, political efficacy, liberalism, and ethnic identity, to name just a few. Such latent concepts play important roles in the theoretical work of all subfields of political science and are important components of both individual-level and system-level explanations.

Two complementary approaches have dominated the literature dealing with the measurement of latent concepts. The first approach emphasizes methods of data collection. This includes the collection of proxy variables, the elicitation of expert opinion, and the use of theoretical models to guide and focus data collection. The second approach deals with the construction and use of measurement models. Examples include the use of factor analysis models and item response theory models. While both of these research areas are clearly important, the focus of this paper is squarely on the latter approach.

Until now, a major problem with all models used to measure latent political concepts is that they are appropriate only when the observed responses are either all continuous or all ordinal. Nonetheless, in many instances, some of the observed indicators of the latent variable in question are continuous while others are ordinal. In these situations, researchers often view their data-analytic choices as the following: (a) treat the ordinal variables as continuous and use a standard normal theory factor analysis model, (b) discretize the

Author's note: This work was supported under National Science Foundation Grant SES-0350613. In addition, this project benefitted from related collaborative work with Michael Hechter and Erik Wibbels. Wongi Choe provided valuable research assistance. Dan Ho provided helpful comments on an earlier draft. Any errors are mine alone.

continuous variables and use an item response model, (c) discard the ordinal (continuous) observed variables and work with only the continuous (ordinal) variables, or (d) forgo a model-based measurement strategy entirely. Each of these moves results in (possibly serious) negative consequences for accurate inference. Ignoring the ordinal nature of some of the observed variables can result in falsely precise and possibly biased estimates. Discretizing continuous variables throws away information and diminishes the precision of estimates. Discarding some variables similarly reduces the precision of estimates and may make it impossible to uncover complicated, multidimensional structure in the observed data. Non-model-based methods of measurement cannot account for measurement error and do not allow researchers to assess the uncertainty of the resulting measures.

I address these problems by formulating a measurement model that is appropriate for multivariate responses that have some continuous and some ordinal components. This model does not suffer from the problems listed above and can be applied to strictly continuous, strictly ordinal, or combinations of continuous and ordinal data. Further, the model presented here generalizes both standard normal theory factor analysis models and item response theory models for ordinal data in the sense that both of these types of models are special cases of the model presented below. This allows for straightforward interpretation of the model parameters by anyone who has some experience with either traditional factor analysis or item response theory. I take a Bayesian approach to this modeling enterprise and use Markov chain Monte Carlo (MCMC) to fit the model. This has the added benefit that posterior quantities of interest (such as the probability that unit i has a larger value of the latent construct in question than unit j) are easy to calculate. Software for fitting this model is available in the *MCMCpack* (Martin and Quinn 2004) package for the *R* system for data analysis and graphics (Ihaka and Gentleman 1996). This software is publicly available under the GNU public license.

This paper proceeds as follows. In Section 2, I derive the factor analysis model for mixed data and show how it generalizes the normal theory factor analysis model and two parameter item response models. The third section deals with model fitting via Markov chain Monte Carlo and briefly discusses posterior inference. In Section 4, I show how the model can be used to measure what might be called "political-economic risk" in 62 countries. The last section concludes.

2 A Factor Analysis Model for Mixed Data

As in standard factor analysis and item response theory, the goal of the current modeling enterprise is to capture patterns of association among several observed variables via a relatively parsimonious model. Such a model can be given a latent variable interpretation in which the observed patterns of association arise from an unobserved (i.e., latent) variable or variables. Seen in this light, the observed response variables are imperfect indicators of the unobserved variable or variables.

Let $j = 1, \ldots, J$ index response variables and $i = 1, \ldots, N$ index observations. Let $\mathbf{X}$ denote the $N \times J$ matrix of observed responses. Each observed variable can be either ordinal or continuous. If the jth variable is ordinal it has $C_j > 1$ categories.

I assume that the values of the elements of $\mathbf{X}$ are determined by a $N \times J$ matrix $\mathbf{X}^*$ of latent variables and a collection $\boldsymbol{\gamma}$ of cutpoints:

$$x_{ij} = \begin{cases} x_{ij}^* & \text{if variable } j \text{ is continuous} \\ c & \text{if } x_{ij}^* \in (\gamma_{j(c-1)}, \gamma_{jc}] \text{ and variable } j \text{ is ordinal,} \end{cases} \tag{1}$$

where it is assumed that c can take values in $\{1, 2, \ldots, C_j\}$. To identify the model, I make the standard assumption (see Johnson and Albert 1999) that $\gamma_{j0} \equiv -\infty$, $\gamma_{j1} \equiv 0$, and $\gamma_{jC_j} \equiv \infty$ for all j. The remaining cutpoints are free parameters to be estimated.

Patterns of association between the observed variables in $\mathbf{X}$ are modeled via a factor analytic model for the latent $\mathbf{X}^*$:

$$\mathbf{x}_i^* = \mathbf{\Lambda}\phi_i + \varepsilon_i \quad i = 1, \ldots, N, \tag{2}$$

where x_i^* is the J vector of latent responses specific to observation i, $\mathbf{\Lambda}$ is a $J \times K$ matrix of factor loadings, $\boldsymbol{\phi}_i$ is a K vector of factor scores specific to observation i, and $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \Psi)$ is a J vector of disturbances. In what follows, I will make the simplifying assumption that $\mathbf{\Psi}$ is diagonal. This is not necessary and can be easily relaxed in situations in which correlated measurement error is suspected. For example, Green and Citrin (1994) show how the grouping of similar survey questions in batteries can induce correlated measurement error. As they go on to show, failing to adjust for this dependence structure of the measurement error can seriously bias one's point estimates of quantities of interest (such as $\mathbf{\Lambda}$ and $\boldsymbol{\phi}$). It should be noted that one typically needs a fairly large amount of data to make estimation of the diagnonal elements of $\mathbf{\Psi}$ reasonable. Further, the Markov chain Monte Carlo algorithm used to fit the model would need to be changed slightly to accommodate the nonstandard full conditional distribution of the free elements of $\mathbf{\Psi}$.

It will also be convenient at some points below to write the system of N equations given in Eq. 2 together as

$$\mathbf{X}^* = \mathbf{\Phi}\mathbf{\Lambda}' + \mathbf{E}, \tag{3}$$

where $\mathbf{\Phi}$ is an $N \times K$ matrix with row i equal to $\boldsymbol{\phi}_i$ and $\mathbf{E}$ is an $N \times J$ matrix with row i equal to ε_i.

Some brief comments are in order about this model specification. The first element of $\boldsymbol{\phi}_i$ is set equal to 1 for all i. This ensures that the elements in the first column of $\mathbf{\Lambda}$ function as negative item difficulty parameters for the ordinal response variables. The elements in the first column of $\mathbf{\Lambda}$ that correspond to continuous responses represent the mean of these continuous variables. Elements of the first column of $\mathbf{\Lambda}$ that correspond to continuous variables that have been standardized to have mean 0 should be constrained to 0. Standardizing continuous variables is not necessary in this context but it does aid interpretation in that the elements of $\mathbf{\Lambda}$ that correspond to the continuous variable can be interpreted as factor loadings.

The prior specification completes the model. To identify the model, some elements of $\mathbf{\Lambda}$ may be constrained to constants. In addition, some elements of $\mathbf{\Lambda}$ will be constrained to take only positive (negative) values. This eliminates so-called rotational invariance (Clinton et al., forthcoming). In general, to identify the $K-1$-factor model it should be possible to permute the rows of $\mathbf{\Lambda}$ so that

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_{1,1} & \lambda_{1,2} & 0 & & \ldots & 0 \\ \lambda_{2,1} & \lambda_{2,2} & \lambda_{2,3} & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & & & 0 \\ \lambda_{K,1} & \ldots & & & & \lambda_{K,K} \\ \vdots & & & & & \vdots \\ \lambda_{J,1} & \ldots & & & & \lambda_{J,K} \end{bmatrix}.$$

In addition, at least one element in each column of $\mathbf{\Lambda}$ should be constrained to take only positive or negative values. Lopes and West (1999) use a similar prior specification for a factor loading matrix and discuss issues of model identification.

For the elements of $\mathbf{\Lambda}$ constrained to be positive (negative), I assume a normal prior density truncated below (above) at 0. I assume the remaining elements of $\mathbf{\Lambda}$ follow independent normal distributions. For both the truncated and untruncated normal priors for typical element $\mathbf{\Lambda}_{jk}$ I assume that the mean parameter (before truncation) is $l_{0_{jk}}$ and the precision (inverse variance) parameter (before truncation) is $L_{0_{jk}}$. All elements of $\mathbf{\Lambda}$ are assumed independent a priori. As is standard in the item response theory literature, the diagonal elements of $\mathbf{\Psi}$ that correspond to ordinal response variables are constrained to be equal to 1 for reasons of identification. The diagonal elements of $\mathbf{\Psi}$ that correspond to continuous response variables are given independent inverse gamma priors. More specifically, if variable j is continuous, $\psi_{jj} \sim \mathcal{IG}\ (a_{0_j}/2,\ b_{0_j}/2)$.

I assume an improper, uniform prior for all elements of $\boldsymbol{\gamma}$ and that the latent factor scores following independent standard normal distributions a priori: $\boldsymbol{\phi}_{i(2:K)} \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}), i = 1, \ldots, N$ where $\boldsymbol{\phi}_{i(2:K)}$ represents the vector formed from the second through Kth elements of $\boldsymbol{\phi}_i$. In addition, it is assumed that $\mathbf{\Lambda}$, $\boldsymbol{\phi}$, $\mathbf{\Psi}$, and $\boldsymbol{\gamma}$ are all mutually independent a priori.

Special cases of this model are the traditional normal theory factor analysis model (Lawley 1967; Jöreskog 1969; Lawley and Maxwell 1971), which arises when all the responses are continuous; the traditional two-parameter item response model with probit link (Johnson and Albert 1999), which occurs when all the responses are dichotomous; and the two-parameter item response model for ordinal data with probit link (Johnson and Albert 1999; Treier and Jackman 2003), which arises when all the responses are ordinal, polychotomous variables.

To see this, recall that the standard normal theory factor analysis model is

$$\mathbf{x}_i = \mathbf{\Lambda}\phi_i + \varepsilon_i \quad i = 1, \ldots, N \tag{4}$$

with

$$\phi_i \overset{iid}{\sim} \mathcal{N}(0, \mathbf{I}) \quad i = 1, \ldots, N$$

and

$$\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \Psi) \quad i = 1, \ldots, N.$$

It should be obvious that when all the observed Xs are continuous the normal theory factor analysis model given in Eq. 4 is exactly the same as the model given in Eqs. 1 and 2.

To see the equivalence between the mixed data factor analysis model and the two-parameter item response model with probit link, note that this item response model can be written as[1]

$$x_{ij} = c \quad \text{if } x^*_{ij} \in (\gamma_{j(c-1)}, \gamma_{jc}] \quad i = 1, \ldots, N, \quad j = 1, \ldots, J,\ c = 1, \ldots, C_j \tag{5}$$

$$x^*_{ij} = \alpha_j + \beta'_j\theta_i + \varepsilon_{ij} \quad i = 1, \ldots, N,\ j = 1, \ldots, J$$

$$\varepsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, 1) \quad i = 1, \ldots, N, \quad j = 1, \ldots, J, \tag{6}$$

[1]See Johnson and Albert (1999) and Treier and Jackman (2003).

where α_j is the item difficulty parameter for item j, $\boldsymbol{\beta}_j$ is the vector of item discrimination parameters for item j, and $\boldsymbol{\theta}_i$ is the latent subject ability of subject i. Note that when all observed Xs are ordinal the discretization rules in Eqs. 1 and 5 are equivalent, the jth row of $\boldsymbol{\Lambda}$ is equivalent to $(\alpha_j, \boldsymbol{\beta}'_j)$, $\boldsymbol{\phi}_i$ is equivalent to $(1, \boldsymbol{\theta}'_i)'$, and ε_{ij} is equivalent to ϵ_{ij}. Consequently, when all observed variables are ordinal the mixed factor analysis model developed here is equivalent to the two-parameter item response model with probit link.

3 Model Fitting and Inference

A program that uses the algorithm described below to fit the mixed response factor analysis model is available under the GNU public license. The software is part of the *MCMCpack* (Martin and Quinn 2004) package written for the *R* (Ihaka and Gentleman 1996) environment for statistical computing and graphics and features an easy-to-use, formula-based interface along with a suite of functions for MCMC diagnostics and summarization. Example code is provided in the next section.

Writing the model in terms of the latent $\mathbf{X}^*$ is useful not only to build intuition but also for model fitting. Building on the ideas in Tanner and Wong (1987), Albert and Chib (1993), and Johnson and Albert (1999), I choose to treat $\mathbf{X}^*$ as latent data and work with the posterior density

$$\begin{aligned} p(\mathbf{X}^*, \gamma, \boldsymbol{\Lambda}, \phi, \Psi|\mathbf{X}) &\propto p(\mathbf{X}|\mathbf{X}^*, \boldsymbol{\gamma}) p(\mathbf{X}^*|\boldsymbol{\Lambda}, \phi, \Psi) p(\boldsymbol{\gamma}) p(\boldsymbol{\Lambda}) p(\Phi) p(\Psi) \\ &\propto \left\{ \prod_{i=1}^{N} \prod_{j=1}^{J} \left\{ \mathbb{I}\left(x_{ij} = x^*_{ij}\right) \mathbb{I}\left(X_j \text{ continuous}\right) \right.\right. \\ &\quad \left. + \sum_{c=1}^{C_j} \mathbb{I}\left(x_{ij} = c\right) \mathbb{I}\left(x^*_{ij} \in \left(\gamma_{j(c-1)}, \gamma_{jc}\right]\right) \mathbb{I}\left(X_j \text{ ordinal}\right) \right\} \\ &\quad \left. \times p_{\mathcal{N}}(\mathbf{x}^*_i|\boldsymbol{\Lambda}\phi_i, \Psi) \right\} p(\boldsymbol{\Lambda}) p(\Phi) p(\Psi), \end{aligned} \tag{7}$$

where $\mathbb{I}(a)$ is the indicator function, which is equal to 1 if a is true and equal to 0 otherwise, $p_{\mathcal{N}}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a multivariate normal density with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\sum$ evaluated at $\mathbf{z}$, and $p(\boldsymbol{\Lambda})$, $p(\boldsymbol{\Phi})$, and $p(\boldsymbol{\Psi})$ are the prior densities for $\boldsymbol{\Lambda}$, $\boldsymbol{\phi}$, and $\boldsymbol{\Psi}$, respectively. The prior for $\boldsymbol{\gamma}$ drops because it is a constant for all values of $\boldsymbol{\gamma}$.

The advantage of working with the augmented posterior in Eq. 7 is that it is relatively easy to derive full conditional distributions for most of the model parameters from this posterior density. This admits a fairly simple Markov chain Monte Carlo (MCMC) algorithm[2] to be used for model fitting. The algorithm used here works by sampling from the full conditional distributions of $\mathbf{X}^*$, $\boldsymbol{\Lambda}$, $\boldsymbol{\phi}$, and $\boldsymbol{\Psi}$ and using a Metropolis-Hastings step to sample $\boldsymbol{\gamma}$.

It straightforward to show that the full conditional distribution of x^*_{ij} is a normal distribution with mean $\boldsymbol{\lambda}'_j \boldsymbol{\phi}_i$ and variance 1 truncated to the interval $(\gamma_{j(x_{ij}-1)}, \gamma_{j(x_{ij})}]$ if variable j is ordinal and that the full conditional distribution of x^*_{ij} is a point mass at x_{ij} if variable j is continuous. Here $\boldsymbol{\lambda}_j$ denotes the vector formed from the jth row of $\boldsymbol{\Lambda}$.

The full conditional distribution for $\boldsymbol{\phi}_{i(2:K)}$ is normal with mean $(\mathbf{I} + \boldsymbol{\Lambda}'_{\cdot(\mathbf{2:}K)} \boldsymbol{\Psi}^{-1} \boldsymbol{\Lambda}_{\cdot(2:K)})^{-1} (\boldsymbol{\Lambda}'_{\cdot(\mathbf{2:}K)} \boldsymbol{\Psi}^{-1} (\mathbf{x}^*_i - \boldsymbol{\lambda}_1))$ and variance $(\mathbf{I} + \boldsymbol{\Lambda}'_{\cdot(\mathbf{2:}K)} \boldsymbol{\Psi}^{-1} \boldsymbol{\Lambda}_{\cdot(2:K)})^{-1}$. The notation

[2]For more information about MCMC see Gelman et al. (2003), Robert and Casella (1999), Jackman (2000), and Gill (2002).

$\boldsymbol{\Lambda}_{\cdot(2:K)}$ denotes the matrix formed from the second through Kth columns of $\boldsymbol{\Lambda}$ and $\boldsymbol{\lambda}_1$ denotes the first column of $\boldsymbol{\Lambda}$.

The full conditional for $\boldsymbol{\Lambda}$ is a bit more complicated due to the possibility of both equality and inequality constraints on elements of $\boldsymbol{\Lambda}$. In what follows, I use the notation $\boldsymbol{\lambda}_{j\circ}$ to denote the vector formed from the free elements of row j of $\boldsymbol{\Lambda}$, and $\boldsymbol{\lambda}_{j\bullet}$ to denote the vector formed from the fixed elements of row j of $\boldsymbol{\Lambda}$. Similarly, $\boldsymbol{\Phi}_{j\circ}$ and $\boldsymbol{\Phi}_{j\bullet}$ are matrices containing the elements of $\boldsymbol{\Phi}$ that are multiplied by $\boldsymbol{\lambda}_{j\circ}$ and $\boldsymbol{\lambda}_{j\bullet}$, respectively, in the expression $\boldsymbol{\Phi}\boldsymbol{\Lambda}'$. Finally, $\mathbf{l}_{0_{j\circ}}$ denotes the prior mean of $\boldsymbol{\lambda}_{j\circ}$ and $\mathbf{L}_{0_{j\circ}}$ denotes the prior precision matrix of $\boldsymbol{\lambda}_{j\circ}$. As would be expected given the normal priors for $\boldsymbol{\lambda}_{j\circ}$ and the assumption of normal disturbances in Eq. 2, the full conditional for $\boldsymbol{\lambda}_{j\circ}$ is (possibly truncated) normal. When there are no a priori inequality constraints on $\boldsymbol{\lambda}_{j\circ}$ the full conditional is normal with mean $(\mathbf{L}_{0_{j\circ}} + \psi_{jj}^{-1}\,\boldsymbol{\Phi}'_{j\circ}\,\boldsymbol{\Phi}_{j\circ})^{-1}\,(\mathbf{L}_{0_{j\circ}}\mathbf{l}_{0_{j\circ}} + \psi_{jj}^{-1}\,\boldsymbol{\Phi}'_{j\circ}\,(\mathbf{x}_j^* - \boldsymbol{\Phi}_{j\bullet}\boldsymbol{\lambda}_{j\bullet}))$ and variance $(\mathbf{L}_{0_{j\circ}} + \psi_{jj}^{-1}\,\boldsymbol{\Phi}'_{j\circ}\,\boldsymbol{\Phi}_{j\circ})^{-1}$. When inequality constraints are present, a draw from the full conditional distribution of $\boldsymbol{\lambda}_{j\circ}$ can be achieved by sampling from the previously mentioned normal distribution and accepting only draws that satisfy the inequality constraints.

If variable j is continuous, the full conditional distribution of ψ_{jj} is an inverse gamma distribution with shape $(a_{0_j} + N)/2$ and scale $(b_{0_j} + (\mathbf{x}_j^* - \boldsymbol{\Phi}\boldsymbol{\lambda}_j)'\,(\mathbf{x}_j^* - \boldsymbol{\Phi}\boldsymbol{\lambda}_j))$, where $\mathbf{x}_j^*$ is the jth column of $\mathbf{X}^*$ and $\boldsymbol{\lambda}_j$ is the vector formed from the jth row of $\boldsymbol{\Lambda}$. If variable j is ordinal, ψ_{jj} is constrained to 1.

The full conditional distribution for $\boldsymbol{\gamma}$ is not a member of a known parametric family. To sample $\boldsymbol{\gamma}$ we make use of J Metropolis-Hastings steps to sample $\boldsymbol{\gamma}_j$, $j = 1, \ldots, J$. This works as follows. First, a candidate value $\boldsymbol{\gamma}_j^{(can)}$ is drawn. This is done element by element. The cth element $\gamma_{jc}^{(can)}$ of $\boldsymbol{\gamma}_j^{(can)}$ is drawn from a normal distribution with mean γ_{jc} and variance t_j^2 truncated to the interval $(\gamma_{j(c-1)}^{(can)}, \gamma_{j(c+1)})$ for $c = 2, \ldots, C_j - 1$. Here t_j^2 is a user-specified tuning parameter. t_j^2 should be set by the user so that the fraction of candidate values that are accepted is somewhere between 0.20 and 0.50. One typically sets the tuning parameter by trial and error. The *MCMCpack* implementation of this model reports this acceptance rate so that users can adjust the tuning parameter appropriately. Once the candidate value $\boldsymbol{\gamma}_j^{(can)}$ is drawn it is accepted as a value of $\boldsymbol{\gamma}_j$ with probability α, where

$$\alpha = \prod_{i=1}^{N} \frac{\Phi\left(\gamma_{jx_{ij}}^{(can)} - \lambda'_j\phi_i\right) - \Phi\left(\gamma_{j(x_{ij}-1)}^{(can)} - \lambda'_j\phi_i\right)}{\Phi(\gamma_{jx_{ij}} - \lambda'_j\phi_i) - \Phi(\gamma_{j(x_{ij}-1)} - \lambda'_j\phi_i)} \times \prod_{c=2}^{C-1} \frac{\Phi((\gamma_{c+1} - \gamma_c)/t_j) - \Phi\left(\left(\gamma_{c-1}^{(can)} - \gamma_c\right)/t_j\right)}{\Phi\left(\left(\gamma_{c+1}^{(can)} - \gamma_c^{(can)}\right)/t_j\right) - \Phi\left(\left(\gamma_{c-1} - \gamma_c^{(can)}\right)/t_j\right)}.$$

This Metropolis-Hastings step is very similar to that proposed in Cowles (1996) for the cutpoints in an ordinal regression model.

4 Application: Measuring Political-Economic Risk in 62 Countries

Scholars in the fields of international relations and comparative politics make use of numerous latent concepts that can be linked to both ordinal and continuous indicators. Examples include democracy, economic freedom, corruption, and political and economic risk, to name but a few. The mixed data factor analysis model discussed above is well suited to the task of estimating indices of the latent concepts given observed indicators. In what follows, I use the example of political-economic risk to demonstrate how the mixed data factor analysis model can be used for such a task.

Measures of political-economic risk, which I will take to mean the risk of the state "manipulat[ing] economic rules to the advantage of itself and its constituents" (North and Weingast 1989, p. 808), are of substantial interest to scholars (see inter alia Borner et al. 1995; Knack and Keefer 1995; Sobel 1999; Henisz 2002b), policy makers, and investors. Attempts to measure political-economic risk have proceeded in two distinct directions. The first measurement approach, as seen in the work of Henisz (2002a) and Beck et al. (2001), focuses on institutional determinants of risk relating to the number and strength of veto players in the political system. These approaches have a strong theoretical basis. Further, the focus on well-documented institutional features makes it possible to construct these measures well into the past. Nonetheless, these measures are not without problems. Despite prima facie evidence that political risk varies considerably over moderate time spans, these institutional measures exhibit very little temporal variation. The second measurement strategy, typified by the work of Coplin (2003) and Howell and Coplin (2001), relies on subjective expert assessments of political and economic risk in several subcategories that are then added together to form a cumulative scale. While the resulting measures do display more temporal variability than the institutional measures, the aggregation of expert opinion is not handled in a principled manner. As a result, it is not clear what to make of the precision of these measures.

It should be noted that the modeling goal here is primarily data summarization. If the model does a reasonable job of accounting for variability in the observed indicators, then the latent factor scores will provide a convenient lower-dimensional summary of these observed variables. With a stronger theory linking observed indicators to latent concepts and better data, it is also possible to use the model to form more theoretically meaningful scales. The example below does not quite meet this level of theoretical rigor.

4.1 *Data and Model Specification*

The data used to construct the index of political-economic risk come from three sources: Henisz's *Political Constraint Index (POLCON) Dataset* (Henisz 2002a), the *State Failure Task Force Problem Set* (Marshall et al. 2002), and the *ACLP Political and Economic Database* (Alvarez et al. 1999). Five variables were used in the analysis below. Of these five, one variable was dichotomous; two were nondichotomous, ordinal variables; and two were continuous.

The first observed measure of political-economic risk considered here is an indicator variable that measures the independence of the national judiciary. This variable is equal to 1 if the judiciary is judged to be independent and equal to 0 otherwise. This measure of judicial independence is from Henisz (2002a).

The next observed indicator is the black-market premium in each country. This variable is from Marshall et al. (2002) and is operationalized as the black-market exchange rate (local currency per dollar) divided by the official exchange rate minus 1. In the analysis below, this variable is used as a proxy for illegal economic activity more generally. I take the natural log of the black-market premium to transform it to approximate normality. Because the black-market premium was 0 for some country years, 0.001 was added to the black-market premium before taking the log.

The third measure of political-economic risk is an ordinal variable from Marshall et al. (2002) measuring the lack of expropriation risk. As originally coded, this variable ranged from 0 to 10 and was based on a subjective coding by the State Failure Task Force. Because of extremely low cell counts in many of the categories I collapsed the original scale to a six-point scale. Values of the original variable less than or equal to 3.5 were recoded to 0, values of the original variable in (3.5, 4.5] were recoded to 1, values of the

original variable in (4.5, 5.5] were recoded to 2, values of the original variable in (5.5, 6.5] were recoded to 3, values of the original variable in (6.5, 8.5] were recoded to 4, and values of the original variable greater than 8.5 were recoded to 5.

The fourth indicator of political-economic risk is a variable measuring lack of corruption. This variable is from Marshall et al. (2002). As with the expropriation variable above, this is an ordinal variable constructed from expert judgment. The original variable was scaled from 0 to 6 and occasionally took noninteger values. As above, the cell counts in some of these categories were extremely small. As a result I recoded this variable to a 0-to-5 scale by subtracting 1 from the original variable, taking the integer part of this number, and then recoding values of -1 to 0.

The final variable used here is productivity as measured by real gross domestic product (GDP) per worker in 1985 international prices. This variable, from Alvarez et al. (1999), is logged to transform it to approximate normality.

To eliminate complications caused by temporal dependence and to maximize the number of cases with observed data, I focus on measuring political-economic risk in 1987. For that year, 62 countries have fully observed data on all five indicators. The raw data are presented in Table 1.

I treat the independent judiciary, lack of expropriation risk, and lack of corruption variables as ordinal variables and the black-market premium and productivity variables as continuous variables. The continuous variables are standardized to have mean 0 and standard deviation 1. As result I constrain the elements of $\boldsymbol{\lambda}_1$ corresponding to these variables to 0. The prior mean of each element of $\boldsymbol{\Lambda}$ is assumed to be 0 (before any truncation) and the prior precision is assumed to be 0.25 (again, before any truncation). To help identify the model I constrain the element of $\boldsymbol{\lambda}_2$ for the independent judiciary variable to be negative. This implies that an independent judiciary is negatively associated with political-economic risk. To complete the prior specification, I assume that $a_{0_{\text{black-market}}} = a_{0_{\text{productivity}}} = 0.001$ and $b_{0_{\text{black-market}}} = b_{0_{\text{productivity}}} = 0.001$. This is a relatively uninformative prior for the error variances.

4.2 *Results*

Model fitting was accomplished using the MCMC algorithm discussed above. The initial 10,000 MCMC scans were discarded as burn-in. The posterior summaries below are based on a posterior sample of size 10,000 formed by running the chain for an additional 1,000,000 scans and storing every 100th scan. The chain mixed reasonably well and standard diagnostics suggest that the sample is approximately from the stationary distribution of the chain.

To see how easy it is to fit this model in R with *MCMCpack* I have included the code used to fit the model for this application below:

```
library(MCMCpack)
post.samp <- MCMCmixfactanal(~courts+barb2+prsexp2+
                             prscorr2+gdpw2,
                             factors = 1, data=PErisk,
                             lambda.constraints = list(courts=list(2,"-")),
                             burnin=10000, mcmc=1000000, thin=100,
                             verbose=TRUE, L0=0.25,
                             store.lambda=TRUE, store.scores=TRUE, tune=.25)
```

The variables to be modeled (courts, barb2, prsexp2, prscorr2, and gdpw2) are sent to the model-fitting function as the right-hand side of an *R* formula. The user specifies the

Table 1 Political-economic risk data for 62 countries in 1987

Country	*Ind. judic.*	*log(black-market premium)*	*Expropriation*	*Corruption*	*log(GDP/worker)*
Argentina	0	−0.72	1	3	9.69
Australia	1	−6.91	5	4	10.30
Austria	1	−4.91	5	4	10.10
Bangladesh	0	0.78	1	0	8.38
Belgium	1	−4.62	5	4	10.25
Bolivia	0	−2.46	0	0	8.58
Botswana	1	−1.24	4	3	8.78
Brazil	1	−0.46	4	3	9.38
Burma	0	1.60	3	1	7.10
Cameroon	0	−4.23	3	1	8.12
Canada	1	−6.91	5	5	10.41
Chile	1	−1.54	3	2	9.26
Colombia	0	−2.06	3	2	9.19
Congo–Kinshasa	0	−2.32	1	0	7.10
Costa Rica	1	−5.09	3	4	9.17
Cote d'Ivoire	1	−4.23	4	2	8.23
Denmark	1	−6.91	5	5	10.11
Dominican Republic	0	−2.38	2	2	8.90
Ecuador	1	−1.85	3	2	9.12
Finland	1	−6.91	5	5	10.12
Gambia	0	−1.54	4	2	7.50
Ghana	0	−1.01	2	1	7.60
Greece	0	−2.07	3	3	9.70
Hungary	1	−0.90	4	3	9.35
India	0	−2.11	4	2	7.97
Indonesia	0	−2.10	3	0	8.39
Iran	0	2.34	0	2	9.37
Ireland	1	−6.91	5	4	9.89
Israel	0	−2.32	4	4	10.07
Italy	1	−6.91	4	3	10.26
Japan	1	−6.91	5	4	9.89
Kenya	0	−2.33	2	2	7.62
Korea, South	0	−2.66	4	1	9.42
Malawi	0	−1.47	3	3	7.03
Malaysia	1	−3.93	4	3	9.18
Mexico	0	−1.66	2	2	9.66
Morocco	0	−3.16	3	1	8.78
New Zealand	1	−6.91	5	5	10.18
Nigeria	0	0.30	1	1	7.69
Norway	1	−6.91	5	5	10.30
Papua New Guinea	1	−2.64	4	2	8.13
Paraguay	0	−0.97	3	0	8.73
Philippines	0	−2.96	1	1	8.38
Poland	1	1.32	3	3	9.05
Portugal	1	−2.46	4	3	9.44
Sierra Leone	0	1.41	3	1	7.76
Singapore	1	−4.85	5	5	9.88
South Africa	0	−2.18	3	4	9.19

Table 1 Continued

Country	*Ind. judic.*	*log(black-market premium)*	*Expropriation*	*Corruption*	*log(GDP/worker)*
Spain	1	−6.91	5	3	10.05
Sri Lanka	0	−1.86	2	2	8.63
Sweden	1	−6.91	4	5	10.22
Switzerland	1	−6.91	5	5	10.34
Syria	0	1.73	1	1	9.66
Thailand	0	−6.91	3	2	8.55
Togo	0	−4.23	4	1	7.33
Tunisia	0	−2.59	2	2	9.05
Turkey	0	−2.67	3	2	8.98
United Kingdom	1	−6.91	5	5	10.13
Uruguay	0	−2.13	2	2	9.41
Venezuela	1	0.43	3	2	9.85
Zambia	0	0.97	3	1	7.73
Zimbabwe	0	−0.64	3	2	7.97

Note. The *corruption* and *expropriation* variables are measures of the lack of corruption and lack of expropriation risk, respectively. Also, log(black-market premium) is actually log(0.001 + black-market premium).

number of factors to estimated (in this case one), the data set, what constraints are to be applied to $\mathbf{\Lambda}$ (in this case the discrimination parameter on the courts variable is constrained to be negative), the number of burn-in iterations, the number of MCMC scans after burn-in, the thinning interval, the prior precision (in this case 0.25) for the elements of $\mathbf{\Lambda}$, whether to store the samples from the posterior distribution of $\mathbf{\Lambda}$, whether to store the samples from the posterior distribution of the factor scores, and the tuning parameter for the Metropolis-Hastings algorithm. Full details are provided in the *MCMCpack* documentation.

Table 2 displays a summary of the posterior distribution of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$. Looking first at the parameters directly tied to the continuous variables log(black-market premium) and log(GDP/worker) we see that this one-dimensional factor model can account for a sizable portion of variability in these variables. Since both of these variables were standardized to have mean 0 and standard deviation 1, we can interpret the results as we would the results from confirmatory factor analysis. The factor loading of 0.750 on log(black-market premium) is large and indicates a positive association between the black-market premium and the latent factor that we are interpreting to be political-economic risk. This is in line with our prior expectations. Further, the estimated error variance is estimated to be 0.453, which implies that over half of the variability in the log of the black-market premium can be accounted for by the single latent factor. Similarly, the estimated factor loading on log(GDP/worker) is −0.721, which indicates, as expected, a negative association between productivity and the latent factor. Once again, the latent factor accounts for over half the variability in log(GDP/worker).

Turning to the ordinal variables we again see results in line with our prior expectations. The elements of $\boldsymbol{\lambda}_2$ for the ordinal variables are all estimated to be negative and have almost no posterior mass to the right of 0. Again this indicates a strong negative association between the latent factor and an independent judiciary, lack of expropriation threat, and lack of corruption. In the terminology of item response theory, these variables discriminate well.

Table 2 Posterior density summary of the measurement model of political-economic risk

	$\boldsymbol{\lambda}_1$	$\boldsymbol{\lambda}_2$	ψ_{jj}
Independent judiciary	−0.041	−2.930	1.000
	(0.370)	(0.983)	–
log(black-market premium)	0.000	0.750	0.453
	–	(0.114)	(0.099)
Lack of expropriation threat	3.517	−1.963	1.000
	(0.639)	(0.459)	–
Lack of corruption	3.146	−2.278	1.000
	(0.629)	(0.546)	–
Productivity (log(GDP/worker))	0.000	−0.721	0.494
	–	(0.115)	(0.105)

Note: Entries without parentheses are posterior means and entries with parentheses are posterior standard deviations. The column labeled λ_1 (the first column of $\boldsymbol{\Lambda}$) provides information about what can be thought of as negative item difficulty parameters in the ordinal item response theory literature; the column labeled λ_2 (the second column of $\boldsymbol{\Lambda}$) provides information about what can be thought of as the factor loadings or the item discrimination parameters; and the column labeled ψ_{jj} provides information regarding the error variances. The element of λ_2 for *independent judiciary* was constrained to be negative. The chain was run for 1,000,000 scans after 10,000 burn-in scans. Every 100th scan was saved. The Metropolis-Hastings acceptance rate was 0.352.

While the posterior summary of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ provides information about the overall patterns of association among our observed variables, what researchers are often most interested in are the estimates of the latent factor scores (the $\boldsymbol{\phi}_i$s). One of the great advantages of the Bayesian treatment of the model above (as well as measurement models more generally) is that this approach allows one get point estimates and uncertainty estimates about the latent factor scores in a principled fashion. Such point and interval estimates are presented in Fig. 1. Interpreted as a measure of political risk, the latent factor has good face validity. We see that Canada, Switzerland, Norway, New Zealand, Finland, Denmark, and the United Kingdom are all at the very low end of the scale, indicating that they are the least risky. However, Bolivia, Congo-Kinshasa, and Bangladesh are all at the high end of the risk scale.

It is also interesting to look at the marginal 90% credible intervals for the latent factor scores. Typically, when scales are constructed using confirmatory factor analysis or similar techniques, only point estimates are reported. Nonetheless, as we see in Fig. 1, there can be substantial uncertainty in such measures. However, it should be noted that these marginal credible intervals are somewhat misleading due to the positive correlation between the latent factor scores. If one is truly interested in determining whether observation a has a higher value of the latent factor than observation b then one should calculate the posterior probability that ϕ_a is greater than ϕ_b. This is easily accomplished by simply taking the MCMC output and calculating the fraction of the scans in which ϕ_a was greater than ϕ_b.

As an example, to calculate the posterior probability that $\phi_{\text{Singapore}}$ is greater than ϕ_{Denmark} we calculate the fraction of draws for which $\phi_{\text{Singapore}} > \phi_{\text{Denmark}}$. Doing this, we

→

Fig. 1 Plot of estimates of latent political-economic risk, 1987. Dots are posterior means and the thick line segments depict the (marginal) central 90% credible intervals for each country.

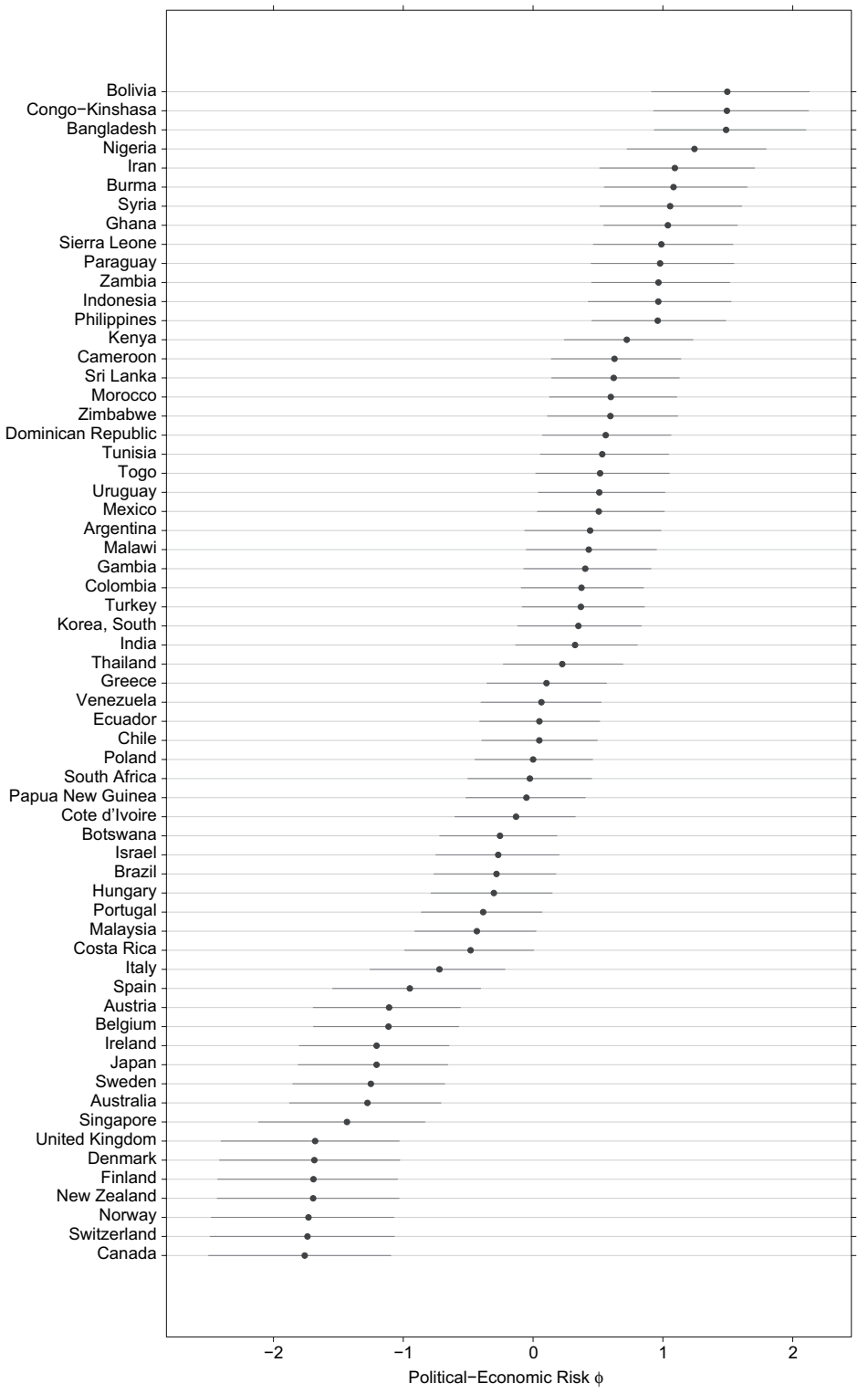


Table 3 Factor loadings and uniquenesses from a normal theory maximum likelihood factor analysis of political-economic risk

	Loadings	*Uniquenesses*
Independent judiciary	−0.796	0.366
log(black-market premium)	0.742	0.450
Lack of expropriation threat	−0.803	0.356
Lack of corruption	−0.901	0.189
Productivity (log(GDP/worker))	−0.747	0.442

find that the posterior probability that Singapore is "riskier" (as judged by its latent factor score) than Denmark is 0.67.

It is also possible to calculate the posterior probability that a particular country has the highest latent risk score. To do this we calculate the fraction of the time that the ϕ draw for the country in question was the maximum value of ϕ across all the countries. Doing this for the three highest-ranked countries (Bolivia, Congo-Kinshasa, and Bangladesh), we find that there is an approximately 26% chance that Bolivia is the riskiest country, a 25% chance that Congo-Kinshasa is the riskiest country, and a 24% chance that Bangladesh is the riskiest country. The probability that one of these three countries has the highest risk rating is simply $0.26 + 0.25 + 0.24 = 0.75$.

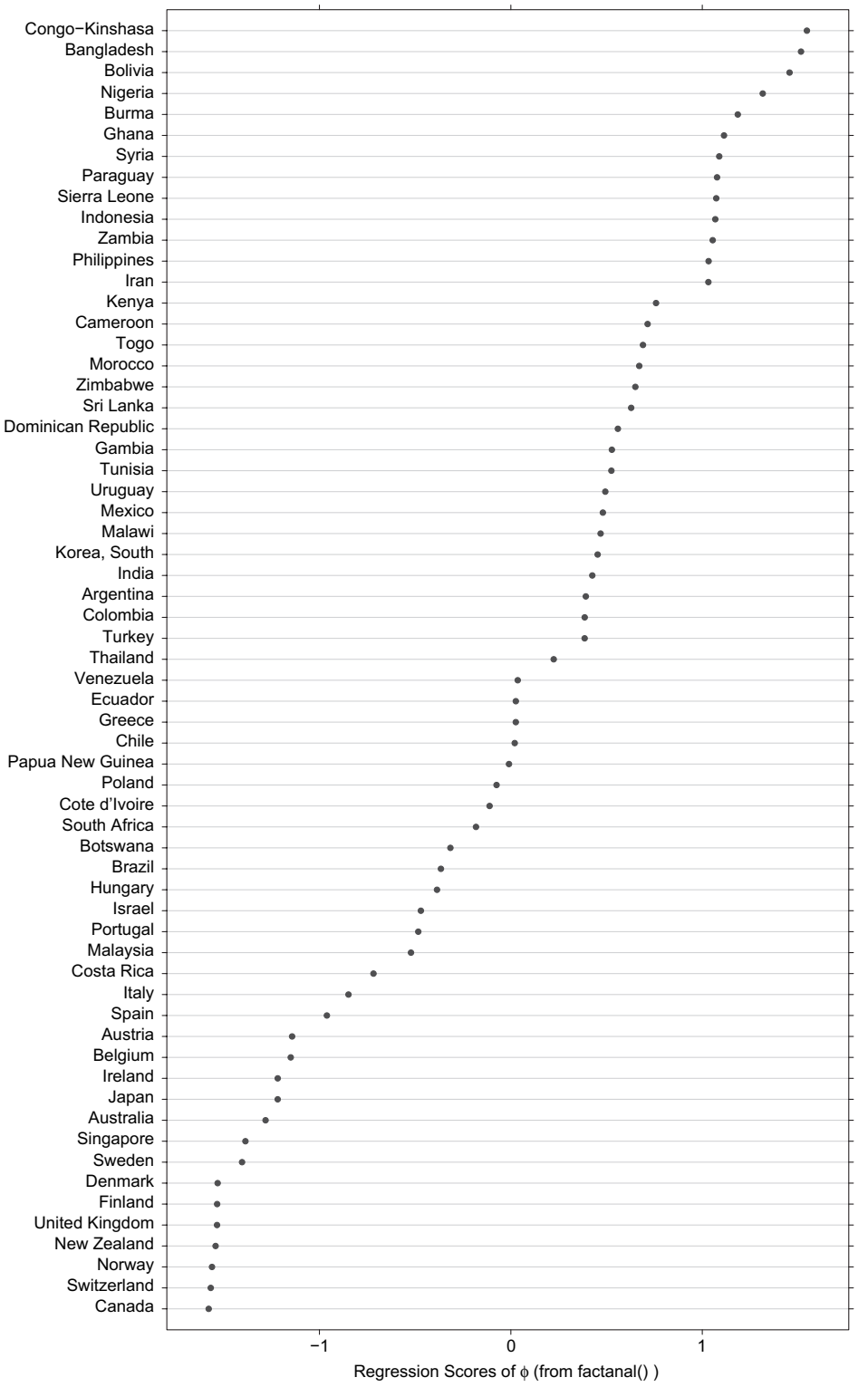
One might reasonably wonder how the mixed data factor analysis results compare with those from a cruder classical normal theory factor analysis. To gauge this, I used the *R* *factanal()* function to fit such a model to these data under the (clearly false) assumption that the five manifest variables follow a multivariate normal distribution. The estimated factor loadings and uniquenesses from this fit are presented in Table 3. It should be noted that the results in this table are not directly comparable to the results in Table 2 for essentially the same reason that linear regression coefficients are not comparable to probit coefficients.

What *are* comparable are estimates of the latent factor scores from the mixed data factor analysis model and the predictions of the latent factor scores from the classical model. In the classical framework, the factor scores are not treated as parameters and, as such, there is no way to uniquely estimate them. What can be done in the classical framework is to construct predictions of the latent factor scores. Several approaches have been proposed for this task (see Lawley and Maxwell 1971 for a discussion). Here I use what are commonly referred to as the regression scores as my predictions. Note that the classical method does not provide any uncertainty statements about the latent factor scores. However, the Bayesian framework described in this article almost automatically provides a sense of estimation uncertainty.

Figure 2 presents the predicted factor scores from the classical model. As we would expect, we see general agreement in the ranking of countries on their latent factor scores across models. Nonetheless, there are numerous small-to-moderate changes in the ordering of the countries on the latent scores across the two sets of results. This can most easily be seen by plotting the ranks of the classical scores on the ranks of the Bayesian scores. This is done in Fig. 3. Here we see some moderate changes in the ordering of the countries on

→

Fig. 2 Plot of predictions of latent political-economic risk, 1987. Dots are the regression scores from a normal theory maximum likelihood factor analysis of the data.



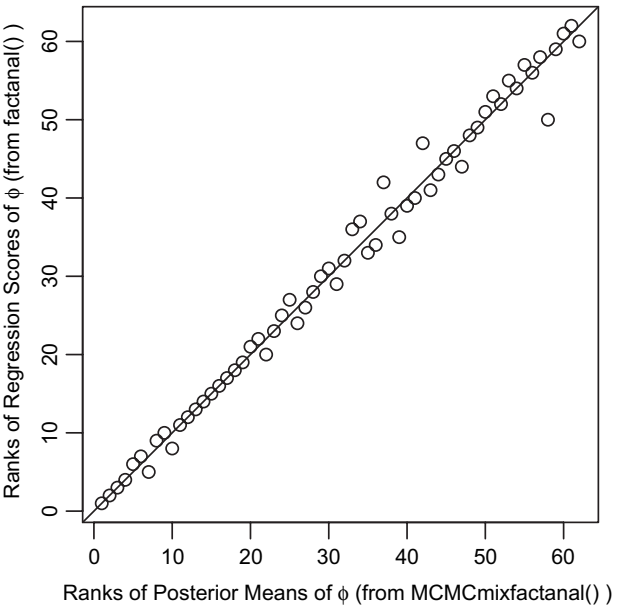


Fig. 3 Comparison of ranks of Bayesian posterior means of ϕ and ranks of classical regression scores from a normal theory factor analysis.

the risk scale. Two countries (Gambia and Togo) move five positions on the scale, while one country (Iran) moves eight positions. Given the ease with which the more principled mixed data factor analysis model can now be fitted, there is currently no real reason to accept this amount of bias.

5 Conclusion

Political scientists routinely deal with latent concepts that have both continuous and ordinal indicators. The model presented above provides a principled means to go forward with model-based measurement in such situations. Not only is it a principled approach, but it is also easy for researchers to use and interpret. Software that implements the MCMC algorithm discussed in the paper is freely available at CRAN (http://cran.r-project.org) as the *mixfactanal()* function in the *MCMCpack* package. The software uses a fairly transparent formula-based interface that should be relatively easy to use for anyone with some experience with *R*. Further, since the model generalizes traditional normal theory factor analysis and item response theory models, the interpretation of the model parameters is fairly straightforward.

It is also possible to extend the model presented above in a number of directions. First, as noted in the body of the paper, allowing for correlated measurement error is quite simple. A different prior for $\mathbf{\Psi}$ is required and the simulation of $\mathbf{\Psi}$ given the other model parameters will also change, but the rest of the MCMC algorithm would remain intact. Second, it is possible to allow for an even wider range of response types. Poisson-distributed counts, censored and/or truncatated variables, and exponential and/or gamma-distributed continuous variables are all possible to accommodate in this general modeling framework. Finally, it is possible to extend the model to allow for temporal, spatial, or spatiotemporal dependence. This is especially important for applications in comparative politics and international relations.

References

Albert, James H., and Siddhartha Chib. 1993. "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association* 88:669–679.

Alvarez, Mike, José Antonio Cheibub, Fernando Limongi, and Adam Przeworski. 1999. "ACLP Political and Economic Database." (Available from http://www.ssc.upenn.edu/~cheibub/data/.)

Beck, Thorsten, George Clarke, Alberto Groff, Philip Keefer, and Patrick Walsh. 2001. "New Tools and New Tests in Comparative Political Economy: The Database of Political Institutions." *World Bank Economic Review* 15:165–176.

Borner, Silvio, Aymo Brunetti, and Beatrice Weder. 1995. *Political Credibility and Economic Development.* New York: St. Martins.

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Voting: A Unified Approach." *American Political Science Review* 98:355–370.

Coplin, William D., ed. 2003. *Political Risk Yearbook, 2003.* East Syracuse, NY: Political Risk Services.

Cowles, M. K. 1996. "Accelerating Monte Carlo Markov Chain Convergence for Cumulative Link Generalized Linear Models." *Statistics and Computing* 6:101–111.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2003. *Bayesian Data Analysis*, 2nd ed. London: Chapman & Hall.

Gill, Jeff. 2002. *Bayesian Methods: A Social and Behavioral Sciences Approach.* Boca Raton, FL: Chapman & Hall/CRC.

Green, Donald P., and Jack Citrin. 1994. "Measurement Error and the Structure of Attitudes: Are Positive and Negative Judgements Opposites?" *American Journal of Political Science* 38:256–281.

Henisz, Witold J. 2002a. "The Political Constraint Index (POLCON) Dataset." (Available from http://www-management.wharton.upenn.edu/henisz/POLCON/ContactInfo.html.)

Henisz, Witold J. 2002b. *Politics and International Investment: Measuring Risk and Protecting Profits.* London: Edward Elgar.

Howell, Llewelyn D., and William D. Coplin, eds. 2001. *The Handbook of Country and Political Risk Analysis,* 3rd ed. East Syracuse, NY: Political Risk Services.

Ihaka, Ross, and Robert Gentleman. 1996. "*R*: A Language for Data Analysis and Graphics." *Journal of Computational and Graphical Statistics* 5:299–314.

Jackman, Simon. 2000. "Estimation and Inference via Bayesian Simulation: An Introduction to Markov Chain Monte Carlo." *American Journal of Political Science* 44:375–404.

Jöreskog, Karl G. 1969. "A General Approach to Confirmatory Maximum Likelihood Factor Analysis." *Psychometrika* 34:183–220.

Johnson, Valen, and James Albert. 1999. *Ordinal Data Modeling.* New York: Springer.

Knack, Stephen, and Philip Keefer. 1995. "Institutions and Economic Performance: Cross-Country Tests Using Alternative Institutional Measures." *Economics and Politics* 7:207–227.

Lawley, D. N. 1967. "Some New Results in Maximum Likelihood Factor Analysis." *Proceedings of the Royal Society of Edinburgh A* 67:256–264.

Lawley, D. N., and A. E. Maxwell. 1971. *Factor Analysis as a Statistical Method.* London: Butterworth.

Lopes, Hedibert Freitas, and Mike West. 1999. "Model Uncertainty in Factor Analysis." Discussion Paper No. 98-38, Durham, NC: Institute of Statistics and Decision Sciences, Duke University.

Marshall, Monty G., Ted Robert Gurr, and Barbara Harff. 2002. "State Failure Task Force Problem Set." (Available from http://www.cidcm.umd.edu/inscr/stfail/index.htm.)

Martin, Andrew D., and Kevin M. Quinn. 2004. "MCMCpack 0.4–8." (Available from http://mcmcpack.wustl.edu/.)

North, Douglass C., and Barry R. Weingast. 1989. "Constituents and Commitment: The Evolution of Institutions Governing Public Choice in Seventeenth Century England." *Journal of Economic History* 49:803–832.

Robert, Christian P., and George Casella. 1999. *Monte Carlo Statistical Methods.* New York: Springer.

Sobel, Andrew C. 1999. *State Institutions, Private Incentives, Global Capital.* Ann Arbor: University of Michigan Press.

Tanner, M. A., and W. Wong. 1987. "The Calculation of Posterior Distributions by Data Augmentation." *Journal of the American Statistical Association* 82:528–550.

Treier, Shawn, and Simon Jackman. 2003. "Democracy as a Latent Variable." Unpublished manuscript, Stanford University.