Chapter 2

Factor analysis

2.1 Introduction

Methodological innovations and real-world applications of factor analysis, and latent structure models more generally, have developed rapidly in recent years, partly due to increased access to appropriate computational tools. In particular, iterative MCMC simulation methods have very naturally opened up access to fully Bayesian treatments of factor analytic models, as developed and applied in, for example, Geweke and Zhou (1996), Polasek (1997), Arminger and Muthén (1998) and, with extensions to dynamic factor components in financial time series modelling (Aguilar and West, 2000; Pitt and Shephard, 1999b). The growing range of developments and creative applications in increasingly complex models, and with larger data sets in higher dimensions, justifies the view that computational advances have been critically enabling; the near future will very likely see much broader use of factor analysis in routine applied statistical work.

The above studies, and others, explore fully Bayesian inference in latent factor models in which the number of factors is a modelling choice; applied work typically studies sensitivity of predictions and variations/ambiguities of interpretations as the number of factors is varied as a control parameter. Formal inference on the number of factors itself has been relatively ignored in the Bayesian literature, though there are ranges of standard likelihood and frequentist methods available. Some key additional references, Bayesian and non-Bayesian, include (in order of appearance) Lawley and Maxwell (1963), Joreskog (1967), Martin and McDonald (1981), Bartholomew (1981), Press (1982) (chapter 10), Lee (1981), Akaike (1987), Bartholomew (1987), Press and Shigemasu (1989), Press and Shigemasu (1994). The book by Bartholomew (1987) is an excellent overview of the field up to about ten years ago.

In this chapter we formally introduce the factor model along with some of its basic properties. Section 2.2 introduces the basic notation and the probabilistic framework in a k-factor model. Sections 2.3 and 2.4 discuss identification issues, invariance to linear transformation and the independence assumption of common factors in some details. We see, for instance, that assuming a nondiagonal covariance structure for unobserved common factor scores is irrelevant from an estimation viewpoint. The incorporation of prior information is touched in Section 2.5, while posterior analysis through Markov chain Monte Carlo is introduced in Section 2.6.

2.2 Basic model form

Data on m related variables are considered to arise through random sampling from a zero-mean multivariate normal distribution denoted by $N(\mathbf{0}, \mathbf{\Omega})$ where $\mathbf{\Omega}$ denotes the $m \times m$ non-singular variance matrix [1]. A random sample of size T is denoted by $\{\boldsymbol{y}_t, t = 1, \ldots, T\}$. For any specified positive integer $k \leq m$, the standard $k-$factor model relates each $\boldsymbol{y}_t$ to an underlying $k-$vector of random variables $\boldsymbol{f}_t$, the common factors, via

$$\boldsymbol{y}_t = \boldsymbol{\beta} \boldsymbol{f}_t + \boldsymbol{\epsilon}_t \tag{2.1}$$

[1] See Appendix A for the definition and some properties of the multivariate normal distribution.

where

- the factors $\boldsymbol{f}_t$ are independent with $\boldsymbol{f}_t \sim N(\mathbf{0}, \boldsymbol{I}_k)$,
- the $\boldsymbol{\epsilon}_t$ are independent normal $m-$vectors with $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, and $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \cdots, \sigma_m^2)$,
- $\boldsymbol{\epsilon}_t$ and $\boldsymbol{f}_s$ are independent for all t and s,
- $\boldsymbol{\beta}$ is the $m \times k$ factor loadings matrix.

Under this model, the variance-covariance structure of the data distribution is constrained; we have $\boldsymbol{\Omega} = V(\boldsymbol{y}_t|\boldsymbol{\Omega}) = V(\boldsymbol{y}_t|\boldsymbol{\beta}, \boldsymbol{\Sigma})$ given by

$$\boldsymbol{\Omega} = \boldsymbol{\beta}\boldsymbol{\beta}' + \boldsymbol{\Sigma}. \tag{2.2}$$

The model implies that, conditional on the common factors, the observable variables are uncorrelated: hence the common factors explain all the dependence structure among the m variables. For any elements y_{it} and y_{jt} of $\boldsymbol{y}_t$, we have the characterising moments:

$$\begin{aligned}
\text{var}(y_{it}|\boldsymbol{\beta}, \boldsymbol{f}, \boldsymbol{\Sigma}) &= \sigma_i^2, && \forall i, \\
\text{cov}(y_{it}, y_{jt}|\boldsymbol{\beta}, \boldsymbol{f}, \boldsymbol{\Sigma}) &= 0, && \forall i, j, i \neq j, \\
\text{var}(y_{it}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) &= \sum_{l=1}^{k} \beta_{il}^2 + \sigma_i^2, && \forall i, \\
\text{cov}(y_{it}, y_{jt}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) &= \sum_{l=1}^{k} \beta_{il}\beta_{jl}, && \forall i, j, i \neq j.
\end{aligned}$$

In practical problems, especially with larger values of m, the number of factors k will often be small relative to m, so that much of the variance-covariance structure is explained by the common factors. The *uniquenesses,* or *idiosyncratic variances,* σ_i^2

measure the residual variability in each of the data variables once that contributed by the factors is accounted for.

The model (2.1) can be written as

$$\boldsymbol{y} = \boldsymbol{F}\boldsymbol{\beta}' + \boldsymbol{\epsilon} \tag{2.3}$$

where $\boldsymbol{y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_T)'$, $\boldsymbol{F} = (\boldsymbol{f}_1, \cdots, \boldsymbol{f}_T)'$ and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_T)'$ are matrices of dimension $(T \times m), (T \times k)$ and $(T \times m)$, respectively. The elements $\boldsymbol{\epsilon}$ and $\boldsymbol{F}$ are mutually independent matrix variate normal random variables, as in Dawid (1981), Press (1982) and West and Harrison (1997) [2]. The notation, as in Dawid (1981), is simply $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{I}_T, \boldsymbol{\Sigma})$. We then have densities

$$p(\boldsymbol{y}|\boldsymbol{F}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-T/2} \boldsymbol{etr}(-0.5\boldsymbol{\Sigma}^{-1}\boldsymbol{\epsilon}\boldsymbol{\epsilon}') \tag{2.4}$$

and, marginalising over $\boldsymbol{F}$,

$$p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Omega}|^{-T/2} \boldsymbol{etr}(-0.5\boldsymbol{\Omega}^{-1}\boldsymbol{y}'\boldsymbol{y}) \tag{2.5}$$

where $\boldsymbol{etr}(\boldsymbol{A}) = \exp(\text{trace}(\boldsymbol{A}))$ for any square matrix $\boldsymbol{A}$. The likelihood function (2.4) will be subsequently used in Gibbs sampling for the parameters of a factor model with k fixed, whereas the likelihood form (2.5) will be extensively used in the RJMCMC algorithms and other techniques that also treat uncertainty about k to be presented in Chapter 3.

2.3 Model structure and identification issues

As is well-known, the $k-$factor model must be further constrained to define a unique model free from identification problems. First we address the standard issue that the model is invariant under transformations of the form $\boldsymbol{\beta}^* = \boldsymbol{\beta}\boldsymbol{P}'$ and $\boldsymbol{f}_t^* = \boldsymbol{P}\boldsymbol{f}_t$,

[2]See Appendix A for the definition and some properties of the matrix variate normal distribution.

where $\boldsymbol{P}$ is any orthogonal $k \times k$ matrix. There are many ways of identifying the model by imposing constraints on $\boldsymbol{\beta}$, including constraints to orthogonal $\boldsymbol{\beta}$ matrices, and constraints such that $\boldsymbol{\beta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}$ is diagonal (see Seber (1984), for example). The alternative preferred here is to constrain $\boldsymbol{\beta}$ to be a block lower triangular matrix, assumed to be of full rank. That is,

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_{11} & 0 & 0 & \cdots & 0 & 0 \\ \beta_{21} & \beta_{22} & 0 & \cdots & 0 & 0 \\ \beta_{31} & \beta_{32} & \beta_{33} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \beta_{k-1,1} & \beta_{k-1,2} & \beta_{k-1,3} & \cdots & \beta_{k-1,k-1} & 0 \\ \beta_{k,1} & \beta_{k,2} & \beta_{k,3} & \cdots & \beta_{k,k-1} & \beta_{k,k} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \beta_{m,1} & \beta_{m,2} & \beta_{m,3} & \cdots & \beta_{m,k-1} & \beta_{m,k} \end{pmatrix} \tag{2.6}$$

where the diagonal elements β_{ii} are strictly positive. This form is used, for example, in Geweke and Zhou (1996) and Aguilar and West (2000), and provides both identification and, often, useful interpretation of the factor model. In this form, the loadings matrix has $r = mk - k(k-1)/2$ free parameters. With m non-zero σ_i parameters, the resulting factor form of $\boldsymbol{\Omega}$ has $m(k+1) - k(k-1)/2$ parameters, compared with the total $m(m+1)/2$ in an unconstrained (or $k = m$) model. This leads to the constraint that

$$m(m+1)/2 - m(k+1) + k(k-1)/2 \geq 0 \tag{2.7}$$

which provides an upper bound on k. For example, $m = 6$ implies $k \leq 3$, $m = 12$ implies $k \leq 7$, $m = 20$ implies $k \leq 14$, $m = 50$ implies $k \leq 40$, and so on. Even for small m, the bound will often not matter as relevant k values will not be so large. In realistic problems, with m in double digits or more, the resulting bound will rarely matter. Finally, note that the number of factors can be increased beyond such bounds by setting one or more of the residual variances σ_i to zero. This is similar to rank restrictions usually present in simultaneous equations estimation of econometric data. When k is larger than the maximum number of factors we have an overidentified model, in econometric terms, and Ω from equation 2.2 is not well defined.

A question arises about the full-rank assumption for $\boldsymbol{\beta}$. This was addressed in Geweke and Singleton (1980) who shown that, if $\boldsymbol{\beta}$ is rank deficient, then the model is unidentified. Specifically, if $\boldsymbol{\beta}$ has rank $r < k$ there exists a matrix $\boldsymbol{Q}$ such that $\boldsymbol{\beta Q} = \boldsymbol{0}$, $\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}$ and, for any orthogonal matrix $\boldsymbol{M}$,

$$\boldsymbol{\Omega} = \boldsymbol{\beta\beta}' + \boldsymbol{\Sigma} = (\boldsymbol{\beta} + \boldsymbol{MQ}')'(\boldsymbol{\beta} + \boldsymbol{MQ}') + (\boldsymbol{\Sigma} - \boldsymbol{MM}'). \tag{2.8}$$

This translation invariance of $\boldsymbol{\Omega}$ under the factor model implies lack of identification and, in application, induces symmetries and potential multimodalities in resulting likelihood functions. This issue relates intimately to the question of uncertainty of the number of factors, discussed further below.

A final question concerns the ordering of the y_{it} variables and the connection between a chosen ordering and the specific form of the factor loading matrix above. The order of variables is a modelling decision that has no effect on the resulting theoretical model nor on predictive inferences under the model. Given the k-factor model (2.1) specified and appropriate for the $\boldsymbol{y}$ with variables in a specific order, alternative orderings are trivially produced via $\boldsymbol{A}\boldsymbol{y}_t$ for some rotation matrix $\boldsymbol{A}$. Model (2.1) then transforms to a similar factor model for the reordered data $\boldsymbol{A}\boldsymbol{y}_t$ with the same latent factors but transformed loadings matrix $\boldsymbol{A\beta}$. This new loadings matrix does not have the lower triangular structure. However, we can always find an orthonormal matrix $\boldsymbol{P}$ such that $\boldsymbol{A\beta P}'$ is lower triangular, and so simply recover the factor model in precisely the form (2.1) with the same probability structure for the underlying latent factors $\boldsymbol{P}\boldsymbol{f}_t$. This result confirms that the order of the variables in $\boldsymbol{y}_t$ is theoretically irrelevant assuming that k is properly chosen. However, when it comes to model estimation, the order of variables has a determining effect on the choice of k, and the interaction between variable order and model fitting can be quite subtle, as we illustrate in examples below.

2.4 Independent common factors

In this section we show that whether $E(\boldsymbol{f}\boldsymbol{f}')$ is diagonal or not is irrelevant, as far as a static factor model is concerned.

Let us start assuming that $\boldsymbol{y}$ follows a k-factor model with dependent common factors, ie. $\boldsymbol{y} = \boldsymbol{\beta}\boldsymbol{f} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{f} \sim N(\mathbf{0}, \boldsymbol{H})$ where $\boldsymbol{H} > 0$ is not restricted to be diagonal.

Since $\boldsymbol{H} > 0$, $\boldsymbol{L}$ we can found such that $\boldsymbol{H} = \boldsymbol{L}\boldsymbol{L}'$ and $\boldsymbol{L}^{-1}\boldsymbol{H}(\boldsymbol{L}')^{-1} = \boldsymbol{L}^{-1}\boldsymbol{H}(\boldsymbol{L}^{-1})' = \boldsymbol{I}$. Then, the new factor model with $\boldsymbol{\beta}$ replaced by $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}\boldsymbol{L}$ and the common factors replaced by $\tilde{\boldsymbol{f}} = \boldsymbol{L}^{-1}\boldsymbol{f}$, has independent common factor structure.

To recover the lower triangular property of $\boldsymbol{\beta}$, the following fact is used; there exists $\boldsymbol{P}$ such that $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{P}\boldsymbol{P}' = \boldsymbol{I}$ and

$$\boldsymbol{\beta}^* = \tilde{\boldsymbol{\beta}}\boldsymbol{P}' = \boldsymbol{\beta}\boldsymbol{L}\boldsymbol{P}'$$

is lower triangular with positive real numbers on the main diagonal.

A particular expression for $\boldsymbol{P}$ is $\boldsymbol{U}^{-1}\tilde{\boldsymbol{\beta}}_1$, where

$$\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1', \tilde{\boldsymbol{\beta}}_2')'$$

and

$$\boldsymbol{U}\boldsymbol{U}' = \tilde{\boldsymbol{\beta}}_1\tilde{\boldsymbol{\beta}}_1'$$

It follows that,

$$\begin{aligned} \boldsymbol{P}'\boldsymbol{P} &= (\boldsymbol{U}^{-1}\tilde{\boldsymbol{\beta}}_1)'(\boldsymbol{U}^{-1}\tilde{\boldsymbol{\beta}}_1) = \tilde{\boldsymbol{\beta}}_1'(\boldsymbol{U}')^{-1}\boldsymbol{U}^{-1}\tilde{\boldsymbol{\beta}}_1 \\ &= \tilde{\boldsymbol{\beta}}_1'(\boldsymbol{U}\boldsymbol{U}')^{-1}\tilde{\boldsymbol{\beta}}_1 = \tilde{\boldsymbol{\beta}}_1'(\tilde{\boldsymbol{\beta}}_1\tilde{\boldsymbol{\beta}}_1')^{-1}\tilde{\boldsymbol{\beta}}_1 = \boldsymbol{I}, \end{aligned}$$

by the definition of $\boldsymbol{P}$; $\boldsymbol{P}$, and

$$\begin{aligned}\boldsymbol{P}\boldsymbol{P}' &= \boldsymbol{U}^{-1}\tilde{\boldsymbol{\beta}}_1\tilde{\boldsymbol{\beta}}_1'(\boldsymbol{U}^{-1})' \\ &= \boldsymbol{U}^{-1}\boldsymbol{U}\boldsymbol{U}'(\boldsymbol{U}')^{-1} = \boldsymbol{I},\end{aligned}$$

This result has been overlooked by most researchers in factor analysis and simplifies matters considerably. For the rest of the next few chapters we will assume that the common factors are, *a priori*, independent. In the next section we set up the prior information.

2.5 Elements of prior specification

To complete the model specification we require classes of priors for the model parameters $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. Our reported analyses are based on very diffuse but proper priors with the following ingredients. For the factor loadings, we take independent priors such that $\beta_{ij} \sim N(0, C_0)$ when $i \neq j$, and $\beta_{ii} \sim TN(0, C_0)$[3] for the upper-diagonal elements of positive loadings $i = 1, \cdots, k$. The latter simply truncates the basic normal prior to restrict the diagonal elements to positive values. Analysis now requires only that we specify the variance parameter C_0, which we take to be rather large in the studies below.

For each of the idiosyncratic variances σ_i^2 we assume a common inverse gamma prior, and take the variances to be independent. Specifically, the σ_i^2 are independently modelled as $\sigma_i^2 \sim IG(\nu/2, \nu s^2/2)$ with specified hyperparameters ν and s^2. Here s^2 is the prior mode of each σ_i^2 and ν the prior degrees of freedom hyperparameter[4]. Our examples below assume values of ν to produce diffuse though proper priors. Note that we eschew the use of standard improper reference priors $p(\sigma_i^2) \propto 1/\sigma_i^2$. Such

[3]See Appendix A for the definition and some properties of the truncated normal distribution.

[4]See Appendix A for the definition and some properties of the inverse gamma distribution.

priors lead to the Bayesian analogue of the so-called *Heywood problem* (Martin and McDonald, 1981; Ihara and Kano, 1995). In terms of these variance parameters, likelihood functions in factor models are bounded below away from zero as σ_i^2 tends to zero, so inducing singularities in the posterior at zero. Proper priors that decay to zero at the origin obviate this problem and induce proper posteriors.

2.6 MCMC methods in a $k-$factor model

With a specified $k-$factor model, Bayesian analyses using MCMC methods are straightforward. We simply summarise the main ingredients here, referring to Geweke and Zhou (1996), Polasek (1997), and Aguilar and West (2000) for further details. MCMC analysis involves iteratively simulating from sets of conditional posterior distributions which, in this model, are standard forms. A basic method simulates from the conditional posteriors for each of $\boldsymbol{F}$, $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ in turn, utilising the following sets of full conditional posteriors arising from our model as specified. These are as follows.

First, the factor model in (2.3) can be seen as a standard multivariate regression model with "parameters" $\boldsymbol{F}$ when $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$ and k are fixed (e.g., Press (1982), Box and Tiao (1973), Broemeling (1985) and Zellner (1971)). It easily follows that the full conditional posterior for $\boldsymbol{F}$ factors into independent normal distributions for the $\boldsymbol{f}_t$, namely

$$\boldsymbol{f}_t \sim N((\boldsymbol{I}_k + \boldsymbol{\beta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_t, (\boldsymbol{I}_k + \boldsymbol{\beta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta})^{-1})$$

independently for $t = 1, \ldots .T$.

Second, the full conditional posterior for $\boldsymbol{\beta}$ also factors into independent margins for the non-zero elements of the rows of $\boldsymbol{\beta}$, as follows. For rows $i = 1, \ldots, k$, write $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{ii})'$ for just these non-zero elements. For the remaining rows $i = k+1, \ldots, m$, write $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{ik})$. Similarly, for $i = 1, \ldots, k$ denote by $\boldsymbol{F}_i$ the $T \times i$ matrix containing the first i columns of $\boldsymbol{F}$, and for all i let $\boldsymbol{y}_i$ be the column i

of $\boldsymbol{y}$.

Finally, it is trivially deduced that full conditional posterior for the elements of $\boldsymbol{\Sigma}$ reduces to a set of m independent inverse gammas, with $\sigma_i^2 \sim IG((\nu+T)/2, (\nu s^2 + d_i)/2)$ where $d_i = (\boldsymbol{y}_i - \boldsymbol{F}\boldsymbol{\beta}_i')'(\boldsymbol{y}_i - \boldsymbol{F}\boldsymbol{\beta}_i')$.

Then we have full conditionals as follows:

- for $i = 1, \ldots, k$, $\boldsymbol{\beta}_i \sim N(\boldsymbol{m}_i, \boldsymbol{C}_i)\mathbf{1}(\beta_{ii} > 0)$ where $\boldsymbol{m}_i = \boldsymbol{C}_i\left(C_0^{-1}\mu_0\mathbf{1}_i + \sigma_i^{-2}\boldsymbol{F}_i'\boldsymbol{y}_i\right)$ and $\boldsymbol{C}_i^{-1} = C_0^{-1}\boldsymbol{I}_i + \sigma_i^{-2}\boldsymbol{F}_i'\boldsymbol{F}_i$;

- for $i = k+1, \ldots, m$, $\boldsymbol{\beta}_i \sim N(\boldsymbol{m}_i, \boldsymbol{C}_i)$ where $\boldsymbol{m}_i = \boldsymbol{C}_i\left(C_0^{-1}\mu_0\mathbf{1}_k + \sigma_i^{-2}\boldsymbol{F}'\boldsymbol{y}_i\right)$ and $\boldsymbol{C}_i^{-1} = C_0^{-1}\boldsymbol{I}_k + \sigma_i^{-2}\boldsymbol{F}'\boldsymbol{F}$.

These distributions are easily simulated.

2.7 Summary

In this chapter we reviewed the state of the art in Bayesian factor models along with model issues, such as invariance to linear transformation, identifiability constraints, prior information and MCMC methods for posterior inference analysis when the number of common factors, k is fixed. Chapter 3 explores simulated and real data applications that rely on the methodology here.

We have explored a variety of simulated and real datasets to test the MCMC algorithm. Some of them are fully explored in the next chapter, when model uncertainty for the number of factors is also considered. However, before moving forward, the following comments are worth making:

- When the number of factors is known to be correct and the factor loadings' prior information is relatively scarce (represented by large values for C_0, for instance), the MCMC algorithm converges quickly for fairly large datasets and posterior first moments converge to the classical maximum likelihood estimators.