Parsimonious Bayesian factor analysis when the number of factors is unknown[1]

Hedibert Freitas Lopes

Booth School of Business
University of Chicago

7th ERPEM
Santa Fe, Argentina, December 3rd 2010

[1] Joint work with Frühwirth-Schnatter.

Example 1
Normal factor model
Variance structure
Identification issues
Number of parameters
Ordering of the variables
Reduced-rank loading matrix
Prior specification
Posterior inference
Example 2
Varimax rotation and correlated factors
Parsimonious BFA
New identification conditions
Theorem
Example 3
Conclusion

Example 1. Early origins of health

Conti, Heckman, Lopes and Remi (2011) Constructing economically justified aggregates: an application of the early origins of health. *Journal of Econometrics*.

Here we focus on a subset of the The British Cohort Study started in 1970.

7 continuous cognitive tests (Picture Language Comprehension,Friendly Math, Reading, Matrices, Recall Digits, Similarities, Word Definition),

10 continuous noncognitive measurements (Child Developmental Scale).

A total of $m = 17$ measurements on $T = 2397$ 10-year old individuals.

Correlation matrix (rounded)

$$
\begin{array}{ccccccc|cccccccccc}
1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1
\end{array}
$$

Normal factor model

For any specified positive integer $k \leq m$, the standard $k-$factor model relates each y_t to an underlying $k-$vector of random variables f_t, the common factors, via

$$y_t | f_t \sim N(\beta f_t, \Sigma)$$

where

$$f_t \sim N(0, I_k)$$

and

$$\Sigma = \text{diag}(\sigma_1^2, \cdots, \sigma_m^2)$$

Unconditional covariance matrix

$$V(y_t | \beta, \Sigma) = \Omega = \beta\beta' + \Sigma$$

Variance structure

Conditional	Unconditional
$\mathrm{var}(y_{it}\mid f) = \sigma_i^2$	$\mathrm{var}(y_{it}) = \sum_{l=1}^k \beta_{il}^2 + \sigma_i^2$
$\mathrm{cov}(y_{it}, y_{jt}\mid f) = 0$	$\mathrm{cov}(y_{it}, y_{jt}) = \sum_{l=1}^k \beta_{il}\beta_{jl}$

Common factors explain all the dependence structure among the m variables.

Identification issues

A rather trivial non-identifiability problem is *sign-switching*.

A more serious problem is factor rotation: invariance under any transformation of the form

$$\beta^* = \beta P' \quad \text{and} \quad f_t^* = P f_t,$$

where P is any orthogonal $k \times k$ matrix.

Solutions

1. $\beta'\Sigma^{-1}\beta = I$.

   Pretty hard to impose!

2. β is a block lower triangular[2].

$$\beta = \begin{pmatrix} \beta_{11} & 0 & 0 & \cdots & 0 \\ \beta_{21} & \beta_{22} & 0 & \cdots & 0 \\ \beta_{31} & \beta_{32} & \beta_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta_{k1} & \beta_{k2} & \beta_{k3} & \cdots & \beta_{kk} \\ \beta_{k+1,1} & \beta_{k+1,2} & \beta_{k+1,3} & \cdots & \beta_{k+1,k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \beta_{m1} & \beta_{m2} & \beta_{m3} & \cdots & \beta_{mk} \end{pmatrix}$$

Somewhat restrictive, but useful for estimation.

[2] Geweke and Zhou (1996) and Lopes and West (2004).

Number of parameters

The resulting factor form of Ω has

$$m(k+1) - k(k-1)/2$$

parameters, compared with the total

$$m(m+1)/2$$

in an unconstrained (or $k = m$) model, leading to the constraint that

$$k \leq m + \frac{3}{2} - \sqrt{2m + \frac{9}{4}}.$$

For example,

- $m = 6$ implies $k \leq 3$,
- $m = 7$ implies $k \leq 4$,
- $m = 10$ implies $k \leq 6$,
- $m = 17$ implies $k \leq 12$.

Ordering of the variables

Alternative orderings are trivially produced via

$$y_t^* = Ay_t$$

for some switching matrix A.

The new rotation has the same latent factors but transformed loadings matrix $A\beta$.

$$y^* = A\beta f + \varepsilon_t = \beta^* f + \varepsilon_t$$

Problem: β^* not necessarily block lower triangular.

Solution

We can always find an orthonormal matrix P such that

$$\tilde{\beta} = \beta^* P' = A\beta P'$$

is block lower triangular and common factors

$$\tilde{f}_t = P f_t$$

still $N(0, I_k)$ (Lopes and West, 2004).

The order of the variables in y_t is immaterial when k is properly chosen, i.e. when β is full-rank.

Reduced-rank loading matrix

Geweke and Singleton (1980) show that, if β has rank $r < k$ then there exists a matrix Q such that

$$\beta Q = 0 \quad \text{and} \quad Q'Q = I$$

and, for any orthogonal matrix M,

$$\beta\beta' + \Sigma = (\beta + MQ')'(\beta + MQ') + (\Sigma - MM').$$

This translation invariance of Ω 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.

Prior specification

Loading matrix:

$$
\begin{aligned}
\beta_{ij} &\sim N(0, C_0) & \text{when} \quad i \neq j, \\
\beta_{ii} &\sim N(0, C_0)1(\beta_{ii} > 0) & \text{when} \quad i = 1, \ldots, k
\end{aligned}
$$

Idiosyncratic variances

$$\sigma_i^2 \sim IG(\nu/2, \nu s^2/2)$$

where s^2 is the prior mode of each σ_i^2 and ν the prior degrees of freedom hyperparameter.

We eschew the use of standard improper reference priors $p(\sigma_i^2) \propto 1/\sigma_i^2$, since such priors lead to the Bayesian analogue of the so-called *Heywood problem* (Martin and McDonald, 1975, and Ihara and Kano, 1995).

Gibbs sampler

Factor scores

$$f_t \sim N(V_f \beta' \Sigma^{-1} y_t, V_f)$$

where $V_f = (I_k + \beta' \Sigma^{-1} \beta)^{-1}$.

Idiosyncrasies

$$\sigma_i^2 \sim IG((\nu + T)/2, (\nu s^2 + d_i)/2)$$

where $d_i = (y_i - f\beta_i')'(y_i - f\beta_i')$.

First k rows of β

$$\beta_i \sim N(M_i, C_i)1(\beta_{ii} > 0)$$

where

$$\begin{aligned} M_i &= C_i \left(C_0^{-1} \mu_0 1_i + \sigma_i^{-2} f_i' y_i \right) \\ C_i^{-1} &= C_0^{-1} I_i + \sigma_i^{-2} f_i' f_i. \end{aligned}$$

Last $m - k$ rows of β

$$\beta_i \sim N(M_i, C_i)$$

where

$$\begin{aligned} M_i &= C_i \left(C_0^{-1} \mu_0 1_k + \sigma_i^{-2} f' y_i \right) \\ C_i^{-1} &= C_0^{-1} I_k + \sigma_i^{-2} f' f. \end{aligned}$$

Example 2. Exchange rate data

- Monthly international exchange rates.
- The data span the period from 1/1975 to 12/1986 inclusive.
- Time series are the exchange rates in British pounds of
  - US dollar (US)
  - Canadian dollar (CAN)
  - Japanese yen (JAP)
  - French franc (FRA)
  - Italian lira (ITA)
  - (West) German (Deutsch)mark (GER)
- Example taken from Lopes and West (2004)

Posterior inference of (β, Σ)

1st ordering

$$E(\beta|y) = \begin{array}{l} \text{US} \\ \text{CAN} \\ \text{JAP} \\ \text{FRA} \\ \text{ITA} \\ \text{GER} \end{array} \begin{pmatrix} 0.99 & 0.00 \\ 0.95 & 0.05 \\ 0.46 & 0.42 \\ 0.39 & 0.91 \\ 0.41 & 0.77 \\ 0.40 & 0.77 \end{pmatrix} \qquad E(\Sigma|y) = \text{diag}\begin{pmatrix} 0.05 \\ 0.13 \\ 0.62 \\ 0.04 \\ 0.25 \\ 0.28 \end{pmatrix}$$

2nd ordering

$$E(\beta|y) = \begin{array}{l} \text{US} \\ \text{JAP} \\ \text{CAN} \\ \text{FRA} \\ \text{ITA} \\ \text{GER} \end{array} \begin{pmatrix} 0.98 & 0.00 \\ 0.45 & 0.42 \\ 0.95 & 0.03 \\ 0.39 & 0.91 \\ 0.41 & 0.77 \\ 0.40 & 0.77 \end{pmatrix} \qquad E(\Sigma|y) = \text{diag}\begin{pmatrix} 0.06 \\ 0.62 \\ 0.12 \\ 0.04 \\ 0.25 \\ 0.26 \end{pmatrix}$$

Posterior inference of f_t

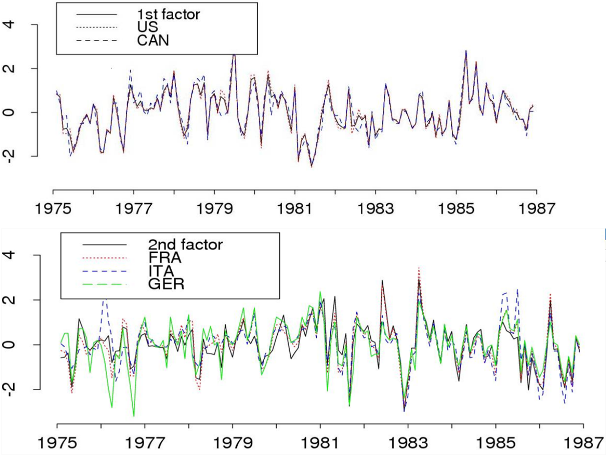
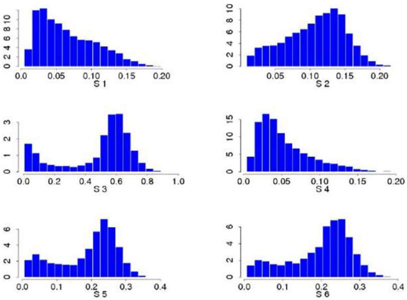


Illustration of multimodality

Variance decomposition

Country	Factor 1	Factor 2
US	95.1	0
CAN	87.6	0.2
JAP	20.5	17.6
FRA	14.7	81.8
ITA	16.4	58.6
GER	16.1	58.5



Marginal posteriors of the idiosyncratic variances when fitting a three-factor structure to the international exchange rates.

Varimax rotation

Proposed by Kaiser (1958) the varimax method of orthogonal rotation aims at providing axes with as few large loadings and as many near-zero loadings as possible.

They are also known as dedicated factors.

Notice that the method can be applied to classical or Bayesian estimates in order to obtain more interpretable factors.

Example 1: ML estimates

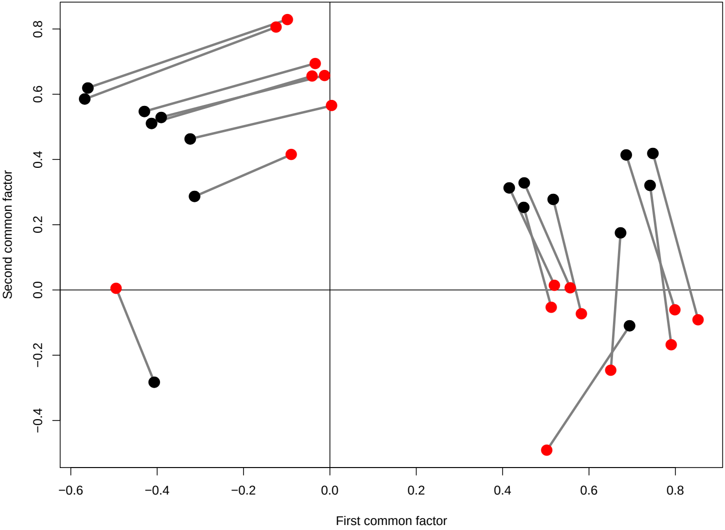
i	Measurement	$\hat{\beta}_{i1}$	$\hat{\beta}_{i2}$	$\hat{\sigma}_i^2$
1	Picture Language	0.32	0.46	0.68
2	Friendly Math	0.57	0.59	0.34
3	Reading	0.56	0.62	0.30
4	Matrices	0.41	0.51	0.57
5	Recall digits	0.31	0.29	0.82
6	Similarities	0.39	0.53	0.57
7	Word definition	0.43	0.55	0.52
8	Child Dev 2	-0.67	0.18	0.52
9	Child Dev 17	-0.69	-0.11	0.51
10	Child Dev 19	-0.74	0.32	0.35
11	Child Dev 20	-0.45	0.33	0.69
12	Child Dev 23	-0.75	0.42	0.26
13	Child Dev 30	-0.52	0.28	0.66
14	Child Dev 31	-0.45	0.25	0.73
15	Child Dev 33	-0.42	0.31	0.73
16	Child Dev 52	0.41	-0.28	0.76
17	Child Dev 53	-0.69	0.41	0.36

Example 1: varimax rotation

i	Measurement	$\hat{\beta}_{i1}$	$\hat{\beta}_{i2}$	$\hat{\sigma}_i^2$
1	Picture Language	0.00	0.56	0.68
2	Friendly Math	0.12	0.81	0.34
3	Reading	0.10	0.83	0.30
4	Matrices	0.04	0.66	0.57
5	Recall digits	0.09	0.41	0.82
6	Similarities	0.01	0.66	0.57
7	Word definition	0.03	0.69	0.52
8	Child Dev 2	-0.65	-0.25	0.52
9	Child Dev 17	-0.50	-0.49	0.51
10	Child Dev 19	-0.79	-0.17	0.35
11	Child Dev 20	-0.56	0.01	0.69
12	Child Dev 23	-0.85	-0.09	0.26
13	Child Dev 30	-0.58	-0.07	0.66
14	Child Dev 31	-0.51	-0.05	0.73
15	Child Dev 33	-0.52	0.02	0.73
16	Child Dev 52	0.49	0.01	0.76
17	Child Dev 53	-0.80	-0.06	0.36

Example 1: Bayesian estimation

i	Measurement	$\tilde{\beta}_{i1}$	$\tilde{\beta}_{i2}$	$\tilde{\sigma}_i^2$
1	Picture Language	0	0.57	0.68
2	Friendly Math	0	0.81	0.34
3	Reading	0	0.83	0.31
4	Matrices	0	0.66	0.56
5	Recall digits	0	0.42	0.83
6	Similarities	0	0.66	0.56
7	Word definition	0	0.7	0.51
8	Child Dev 2	-0.68	0	0.54
9	Child Dev 17	-0.56	0	0.68
10	Child Dev 19	-0.81	0	0.35
11	Child Dev 20	-0.55	0	0.70
12	Child Dev 23	-0.86	0	0.27
13	Child Dev 30	-0.59	0	0.66
14	Child Dev 31	-0.51	0	0.74
15	Child Dev 33	-0.51	0	0.74
16	Child Dev 52	0.49	0	0.76
17	Child Dev 53	-0.8	0	0.36


Second common factor
First common factor
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8

Correlated factors

Assume that

$$y \sim N(\beta f, \Sigma) \quad \text{and} \quad f \sim N(0, H)$$

where H is a non-diagonal covariance matrix.

We can always decompose $LHL' = I$ such that

$$y \sim N(\beta^* f^*, \Sigma) \quad \text{and} \quad f^* = L'f \sim N(0, I_k)$$

where $\beta^* = \beta L$.

When β is "dedicated", there is no guarantee that β^* is also dedicated.

Only when there is lots of prior information regarding β one can entertain correlated factors.

Modern structured factor analysis

- Time-varying factor loadings: dynamic correlations
  Lopes and Carvalho (2007)
- Spatial dynamic factor analysis
  Lopes, Salazar and Gamerman (2008)
- Spatial hierarchical factors: ranking vulnerability
  Lopes, Schmidt, Salazar, Gomez and Achkar (2010)
- Sparse factor models
  Frühwirth-Schnatter and Lopes (2010)

Parsimonious BFA

Frühwirth-Schnatter and Lopes (2010) introduce a more general set of identifiability conditions for the basic model

$$\mathbf{y}_t = \mathbf{\Lambda}\mathbf{f}_t + \boldsymbol{\epsilon}_t \qquad \boldsymbol{\epsilon}_t \sim N_m(\mathbf{0}, \mathbf{\Sigma}_0),$$

which handles the ordering problem in a more flexible way:

C1. $\mathbf{\Lambda}$ has full column-rank, i.e. $r = \mathrm{rank}(\mathbf{\Lambda})$.

C2. $\mathbf{\Lambda}$ is a *generalized lower triangular* matrix, i.e. $l_1 < \ldots < l_r$, where l_j denotes for $j = 1, \ldots, r$ the row index of the top non-zero entry in the jth column of $\mathbf{\Lambda}$, i.e. $\Lambda_{l_j,j} \neq 0; \Lambda_{ij} = 0, \forall\, i < l_j$.

C3. $\mathbf{\Lambda}$ does not contain any column j where $\Lambda_{l_j,j}$ is the only non-zero element in column j.

The regression-type representation

Assume that data $\mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$ were generated by the previous model and that the number of factors r, as well as $\boldsymbol{\Lambda}$ and $\boldsymbol{\Sigma}_0$, should be estimated.

The usual procedure is to fit a model with k factors,

$$\mathbf{y}_t = \beta \mathbf{f}_t + \epsilon_t, \qquad \epsilon_t \sim N_m(\mathbf{0}, \boldsymbol{\Sigma}),$$

where β is a $m \times k$ coefficient matrix with elements β_{ij} and $\boldsymbol{\Sigma}$ is a diagonal matrix.

Theorem

Theorem. Assume that the data were generated by a basic factor model obeying conditions **C1** − **C3** and that a regression-type representation with $k \geq r$ potential factors is fitted. Assume that the following condition holds for $\boldsymbol{\beta}$:

B1 The row indices of the top non-zero entry in each non-zero column of $\boldsymbol{\beta}$ are different.

Then $(r, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}_0)$ are related in the following way to $(\boldsymbol{\beta}, \boldsymbol{\Sigma})$:

(a) r columns of $\boldsymbol{\beta}$ are identical to the r columns of $\boldsymbol{\Lambda}$.

(b) If $\mathrm{rank}(\boldsymbol{\beta}) = r$, then the remaining $k - r$ columns of $\boldsymbol{\beta}$ are zero columns. The number of factors is equal to r and $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}$,

(c) If $\mathrm{rank}(\boldsymbol{\beta}) > r$, then only $k - \mathrm{rank}(\boldsymbol{\beta})$ of the remaining $k - r$ columns are zero columns, while $s = \mathrm{rank}(\boldsymbol{\beta}) - r$ columns with column indices $j_1, \ldots, j_s$ differ from a zero column for a single element lying in s different rows $r_1, \ldots, r_s$. The number of factors is equal to $r = \mathrm{rank}(\boldsymbol{\beta}) - s$, while $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma} + \mathbf{D}$, where $\mathbf{D}$ is a $m \times m$ diagonal matrix of rank s with non-zero diagonal elements $D_{r_l, r_l} = \beta^2_{r_l, j_l}$ for $l = 1, \ldots, s$.

Indicator matrix

Since the identification of r and $\boldsymbol{\Lambda}$ from $\boldsymbol{\beta}$ by Theorem 1 relies on identifying zero and non-zero elements in $\boldsymbol{\beta}$, we follow previous work on parsimonious factor modeling and consider the selection of the elements of $\boldsymbol{\beta}$ as a variable selection problem.

We introduce for each element β_{ij} a binary indicator δ_{ij} and define β_{ij} to be 0, if $\delta_{ij} = 0$, and leave β_{ij} unconstrained otherwise.

In this way, we obtain an indicator matrix $\boldsymbol{\delta}$ of the same dimension as $\boldsymbol{\beta}$.

Number of factors

Theorem 1 allows the identification of the true number of factors r directly from the indicator matrix $\boldsymbol{\delta}$:

$$r = \sum_{j=1}^{k} I\{\sum_{i=1}^{m} \delta_{ij} > 1\},$$

where $I\{\cdot\}$ is the indicator function, by taking spurious factors into account.

This expression is invariant to permuting the columns of $\boldsymbol{\delta}$ which is helpful for MCMC based inference with respect to r.

Our approach provides a principled way for inference on r, as opposed to previous work which are based on rather heuristic procedures to infer this quantity (Carvalho et al., 2008; Bhattacharya and Dunson, 2009).

Model size

The model size d is defined as the number of non-zero elements in $\boldsymbol{\Lambda}$, i.e.

$$d = \sum_{j=1}^{k} d_j I\{d_j > 1\}, \qquad d_j = \sum_{j=1}^{m} \delta_{ij}.$$

Model size could be estimated by the posterior mode $\tilde{d}$ or the posterior mean $\mathrm{E}(d|\mathbf{y})$ of $p(d|\mathbf{y})$.

The Prior of the Indicator Matrix

To define $p(\boldsymbol{\delta})$, we use a hierarchical prior which allows different occurrence probabilities $\boldsymbol{\tau} = (\tau_1, \dots, \tau_k)$ of non-zero elements in the different columns of $\boldsymbol{\beta}$ and assume that all indicators are independent *a priori* given $\boldsymbol{\tau}$:

$$\Pr(\delta_{ij} = 1|\tau_j) = \tau_j, \qquad \tau_j \sim \mathcal{B}\left(a_0, b_0\right).$$

A priori r may be represented as $r = \sum_{j=1}^{k} I\{X_j > 1\}$, where $X_1, \dots, X_k$ are independent random variables and X_j follows a Beta-binomial distribution with parameters $N_j = m - j + 1$, a_0, and b_0.

We recommend to tune for a given data set with known values m and k the hyperparameters a_0 and b_0 in such a way that $p(r|a_0, b_0, m, k)$ is in accordance with prior expectations concerning the number of factors.

The Prior on the Idiosyncratic Variances

Heywood problems typically occur, if the constraint

$$\frac{1}{\sigma_i^2} \geq (\mathbf{\Omega}^{-1})_{ii} \quad \Leftrightarrow \quad \sigma_i^2 \leq \frac{1}{(\mathbf{\Omega}^{-1})_{ii}}$$

is violated, where the matrix $\mathbf{\Omega}$ is the covariance matrix of $\mathbf{y}_t$.

Improper priors on the idiosyncratic variances such as

$$p(\sigma_i^2) \propto 1/\sigma_i^2,$$

are not able to prevent Heywood problems.

We assume instead a proper inverted Gamma prior

$$\sigma_i^2 \sim \mathcal{G}^{-1}(c_0, C_{i0})$$

for each of the idiosyncratic variances σ_i^2.

c_0 is large enough to bound the prior away from 0, typically $c_0 = 2.5$.

C_{i0} is chose by controlling the probability of a Heywood problem

$$\Pr(X \leq C_{i0}(\mathbf{\Omega}^{-1})_{ii})$$

where $X \sim \mathcal{G}(c_0, 1)$. So, C_{i0} as the largest value for which

$$C_{i0}/(c_0 - 1) \leq \frac{1}{(\mathbf{S}_y^{-1})_{ii}}$$

or

$$\sigma_i^2 \sim \mathcal{G}^{-1}\left(c_0, (c_0 - 1)/(\mathbf{S}_y^{-1})_{ii}\right).$$

The Prior on the Factor Loadings

We assume that the rows of the coefficient matrix $\boldsymbol{\beta}$ are independent *a priori* given the factors $\mathbf{f}_1, \ldots, \mathbf{f}_T$.

Let $\boldsymbol{\beta}_{i\cdot}^{\boldsymbol{\delta}}$ be the vector of unconstrained elements in the ith row of $\boldsymbol{\beta}$ corresponding to $\boldsymbol{\delta}$. For each $i = 1, \ldots, m$, we assume that

$$\boldsymbol{\beta}_{i\cdot}^{\boldsymbol{\delta}} | \sigma_i^2 \sim N\left(\mathbf{b}_{i0}^{\boldsymbol{\delta}}, \mathbf{B}_{i0}^{\boldsymbol{\delta}} \sigma_i^2\right).$$

The prior moments are either chosen as in Lopes and West (2004) or Ghosh and Dunson (2009) who considered a "unit scale" prior where $\mathbf{b}_{i0}^{\boldsymbol{\delta}} = \mathbf{0}$ and $\mathbf{B}_{i0}^{\boldsymbol{\delta}} = \mathbf{I}$.

Fractional prior

We use a fractional prior (O'Hagan, 1995) which was applied by several authors for variable selection in latent variable models[3]

The fractional prior can be interpreted as the posterior of a non-informative prior and a fraction b of the data.

We consider a conditionally fractional prior for the "regression model"

$$\tilde{\mathbf{y}}_i = \mathbf{X}_i^\delta \beta_{i\cdot}^\delta + \tilde{\boldsymbol{\epsilon}}_i,$$

where $\tilde{\mathbf{y}}_i = (y_{i1} \cdots y_{iT})'$ and $\tilde{\boldsymbol{\epsilon}}_i = (\epsilon_{i1} \cdots \epsilon_{iT})'$. $\mathbf{X}_i^\delta$ is a regressor matrix for $\beta_{i\cdot}^\delta$ constructed from the latent factor matrix $\mathbf{F} = (\mathbf{f}_1 \cdots \mathbf{f}_T)'$.

[3] Smith & Kohn, 2002; Frühwirth-Schnatter & Tüchler, 2008; Tüchler, 2008.

Using

$$p(\beta_{i\cdot}^{\delta}|\sigma_i^2) \propto p(\tilde{\mathbf{y}}_i|\beta_{i\cdot}^{\delta}, \sigma_i^2)^b$$

we obtain from regression model:

$$\beta_{i\cdot}^{\delta}|\sigma_i^2 \sim N\left(\mathbf{b}_{iT}, \mathbf{B}_{iT}\sigma_i^2/b\right),$$

where $\mathbf{b}_{iT}$ and $\mathbf{B}_{iT}$ are the posterior moments under an non-informative prior:

$$\mathbf{B}_{iT} = \left((\mathbf{X}_i^{\delta})'\mathbf{X}_i^{\delta}\right)^{-1}, \qquad \mathbf{b}_{iT} = \mathbf{B}_{iT}(\mathbf{X}_i^{\delta})'\tilde{\mathbf{y}}_i.$$

It is not entirely clear how to choose the fraction b for a factor model. If the regressors $\mathbf{f}_1, \ldots, \mathbf{f}_T$ were observed, then we would deal with m independent regression models for each of which T observations are available and the choice $b = 1/T$ would be appropriate.

The factors, however, are latent and are estimated together with the other parameters. This ties the m regression models together.

If we consider the multivariate regression model as a whole, then the total number $N = mT$ of observations has to be taken into account which motivates choosing $b_N = 1/(Tm)$.

In cases where the number of regressors d is of the same magnitude as the number of observations, Ley & Steel (2009) recommend to choose instead the risk inflation criterion $b_R = 1/d^2$ suggested by Foster & George (1994), because b_N implies a fairly small penalty for model size and may lead to overfitting models.

In the present context this implies choosing $b_R = 1/d(k,m)^2$ where $d(k,m) = (km - k(k-1)/2)$ is the number of free elements in the coefficient matrix $\boldsymbol{\beta}$.

Most visited configuration

Let $\mathbf{l}^\star = (l_1^\star, \ldots, l_{r_M}^\star)$ the most visited configuration.

We may then estimate for each indicator the marginal inclusion probability

$$\Pr(\delta_{ij}^\Lambda = 1 | \mathbf{y}, \mathbf{l}^\star)$$

under $\mathbf{l}^\star$ as the average over the elements of $(\boldsymbol{\delta}^\Lambda)^\star$.

Note that $\Pr(\delta_{l_j,j}^\Lambda = 1 | \mathbf{y}, \mathbf{l}^\star) = 1$ for $j = 1, \ldots, r_M$.

Following Scott and Berger (2006), we determine the median probability model (MPM) by setting the indicators δ_{ij}^Λ in $\boldsymbol{\delta}_M^\Lambda$ to 1 iff $\Pr(\delta_{ij}^\Lambda = 1 | \mathbf{y}, \mathbf{l}^\star) \geq 0.5$.

The number of non-zero top elements r_M in the identifiability constraint $\mathbf{l}^\star$ is a third estimator of the number of factors, while the number d_M of non-zero elements in the MPM is yet another estimator of model size.

A discrepancy between the various estimators of the number of factors r is often a sign of a weakly informative likelihood and it might help to use a more informative prior for $p(r)$ by choosing the hyperparameters a_0 and b_0 accordingly.

Also the structure of the indicator matrices $\boldsymbol{\delta}_H^\Lambda$ and $\boldsymbol{\delta}_M^\Lambda$ corresponding, respectively, to the HPM and the MPM may differ, in particular if the frequency p_H is small and some of the inclusion probabilities $\Pr(\delta_{ij}^\Lambda = 1|\mathbf{y}, \mathbf{l}^\star)$ are close to 0.5.

MCMC

(a) Sample $\boldsymbol{\delta}$ from $p(\boldsymbol{\delta}|\mathbf{f}_1,\ldots,\mathbf{f}_T,\boldsymbol{\tau},\mathbf{y})$:

(b) Sample $\boldsymbol{\beta}$ and $\sigma_1^2,\ldots,\sigma_m^2$ from $p(\boldsymbol{\beta},\sigma_1^2,\ldots,\sigma_m^2|\boldsymbol{\delta},\mathbf{f}_1,\ldots,\mathbf{f}_T,\mathbf{y})$.

(c) Sample $\mathbf{f}_1,\ldots,\mathbf{f}_T$ from $p(\mathbf{f}_1,\ldots,\mathbf{f}_T|\boldsymbol{\beta},\sigma_1^2,\ldots,\sigma_m^2,\mathbf{y})$.

(d) Perform an acceleration step.

(e) For each $j=1,\ldots,k$, perform a random sign switch: substitute the draws of $\{f_{jt}\}_{t=1}^T$ and $\{\beta_{ij}\}_{i=j}^m$ with probability 0.5 by $\{-f_{jt}\}_{t=1}^T$ and $\{-\beta_{ij}\}_{i=j}^m$, otherwise leave these draws unchanged.

(f) Sample τ_j for $j=1,\ldots,k$ from $\tau_j|\boldsymbol{\delta}\sim\mathcal{B}\left(a_0+d_j,b_0+p_j\right)$, where p_j is the number of free elements and $d_j=\sum_{i=1}^m\delta_{ij}$ is number of non-zero elements in column j.

To generate sensible values for the latent factors in the initial model specification, we found it useful to run the first few steps without variable selection.

Example 2 (cont.)

Table: Exchange rate data; posterior distribution $p(r|\mathbf{y})$ of the number r of factors (bold number corresponding to the posterior mode $\tilde{r}$) and number of visited models N_v for various priors; frequency p_H, number of factors r_H, and model size d_H of the highest probability model (HPM).

	$p(r\|\mathbf{y})$						
Prior	1	2	3	N_v	p_H	r_H	d_H
$b = 0.01$	0	**0.799**	0.201	37	0.50	2	10
$b_N = 0.0022$	0	**0.984**	0.016	24	0.85	2	10
$b = 0.001$	0	**0.974**	0.026	15	0.89	2	10
GD	0	**0.966**	0.034	36	0.71	2	10
LW	0	**0.944**	0.056	37	0.69	2	10

Table: Exchange rate data; marginal inclusion posterior probabilities $\Pr(\delta_{ij}^{\Lambda} = 1|\mathbf{y}, \mathbf{l}^{\star})$ for a fractional prior with $b = b_N$ (bold elements correspond to the HPM).

	$\boldsymbol{\delta}_{\cdot 1}^{\Lambda}$	$\boldsymbol{\delta}_{\cdot 1}^{\Lambda}$
US	**1.**	0
Can	**1.**	0
Yen	**1.**	**1.**
FF	**1.**	**1.**
Lira	**1.**	**1.**
DM	**1.**	**1.**

Table: Exchange rate data; posterior means of the factor loading matrix, the idiosyncratic variances and the communalities for a two-factor model under the constraint **C2** with $l_1 = 1$ and $l_2 = 3$; bold numbers correspond to non-zero elements in the factor loading matrix of the 2-factor HPM.

Currency	$\mathrm{E}(\Lambda_{i1}\|\mathbf{y}, \mathbf{l}^\star)$	$\mathrm{E}(\Lambda_{i2}\|\mathbf{y}, \mathbf{l}^\star)$	$\mathrm{E}(\sigma_i^2\|\mathbf{y}, \mathbf{l}^\star)$
US	**0.960**	0	0.081
Can	**0.951**	0	0.098
Yen	**0.449**	**0.418**	0.615
FF	**0.395**	**0.889**	0.053
Lira	**0.415**	**0.764**	0.241
DM	**0.408**	**0.765**	0.245

Example 3: Maxwell's Children Data

Scores on 10 tests for a sample of $T = 148$ children attending a psychiatric clinic as well as a sample of $T = 810$ normal children.

The first five tests are cognitive tests – (1) verbal ability, (2) spatial ability, (3) reasoning, (4) numerical ability and (5) verbal fluency. The resulting tests are inventories for assessing orectic tendencies, namely (6) neurotic questionaire, (7) ways to be different, (8) worries and anxieties, (9) interests and (10) annoyances.

While psychological theory suggests that a 2-factor model is sufficient to account for the variation between the test scores, the significance test considered in Maxwell (1961) suggested to fit a 3-factor model to the first and a 4-factor model to the second data set.

Table: Maxwell's Children Data - neurotic children; posterior distribution $p(r|\mathbf{y})$ of the number r of factors (bold number corresponding to the posterior mode $\tilde{r}$) and highest posterior identifiability constraint $\mathbf{l}^\star$ with corresponding posterior probability $p(\mathbf{l}^\star|\mathbf{y})$ for various priors; number of visited models N_v; frequency p_H, number of factors r_H, and model size d_H of the HPM; number of factors r_M and model size d_M of the MPM corresponding to $\mathbf{l}^\star$; inefficiency factor τ_d of the posterior draws of the model size d.

	$p(r\|\mathbf{y})$					
Prior	2	3	4	5 - 6	$\mathbf{l}^\star$	$p(\mathbf{l}^\star\|\mathbf{y})$
$b = 10^{-3}$	**0.755**	0.231	0.014	0	(1,6)	0.532
$b_N = 6.8 \cdot 10^{-4}$	**0.828**	0.160	0.006	0	(1,6)	0.623
$b_R = 4.9 \cdot 10^{-4}$	**0.871**	0.127	0.001	0	(1,6)	0.589
$b = 10^{-4}$	**0.897**	0.098	0.005	0	(1,6)	0.802
GD	0.269	**0.482**	0.246	0.003	(1,2,3)	0.174
LW	0.027	0.199	**0.752**	0.023	(1,2,3,6)	0.249

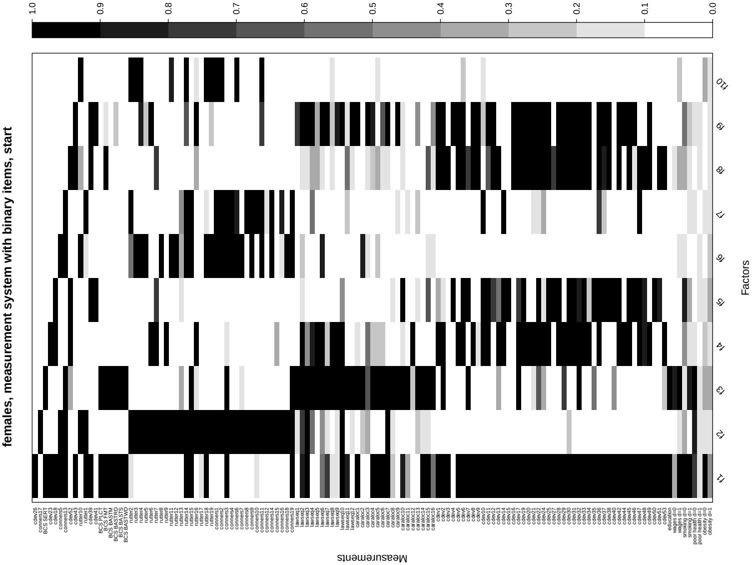
Prior	N_v	p_H	r_H	d_H	r_M	d_M	τ_d
$b = 10^{-3}$	1472	0.20	2	12	2	12	30.9
$b_N = 6.8 \cdot 10^{-4}$	976	0.27	2	12	2	12	27.5
$b_R = 4.9 \cdot 10^{-4}$	768	0.34	2	12	2	12	22.6
$b = 10^{-4}$	421	0.45	2	12	2	12	18.1
GD	4694	0.06	2	15	3	19	40.6
LW	7253	0.01	4	24	4	24	32.5

Table: Maxwell's Children Data - normal children; posterior distribution $p(r|\mathbf{y})$ of the number r of factors (bold number corresponding to the posterior mode $\tilde{r}$) and highest posterior identifiability constraint $\mathbf{l}^\star$ with corresponding posterior probability $p(\mathbf{l}^\star|\mathbf{y})$ for various priors; number of visited models N_v; frequency p_H, number of factors r_H, and model size d_H of the HPM; number of factors r_M and model size d_M of the MPM corresponding to $\mathbf{l}^\star$; inefficiency factor τ_d of the posterior draws of the model size d.

	$p(r\|\mathbf{y})$					
Prior	3	4	5	6	$\mathbf{l}^\star$	$p(\mathbf{l}^\star\|\mathbf{y})$
$b = 10^{-3}$	0	0.391	**0.604**	0.005	(1,2,4,5,6)	0.254
$b_N = 1.2 \cdot 10^{-4}$	0	**0.884**	0.116	0	(1,2,4,5)	0.366
$b = 10^{-4}$	0	**0.891**	0.104	0.005	(1,2,4,5)	0.484
GD	0	0.396	**0.594**	0	(1,2,4,5,6)	0.229
LW	0	0.262	**0.727**	0.011	(1,2,4,5,6)	0.259

Prior	N_v	p_H	r_H	d_H	r_M	d_M	τ_d
$b = 10^{-3}$	4045	1.79	5	26	5	27	30.1
$b_N = 1.2 \cdot 10^{-4}$	1272	11.41	4	23	4	23	28.5
$b = 10^{-4}$	1296	12.17	4	23	4	23	29.1
GD	4568	1.46	5	29	5	28	32.7
LW	5387	1.56	5	28	5	28	30.7

Example 1: 130 measurements.



Conclusion

- Summary
  - New take on factor model identifiability
  - Generalized triangular parametrization
  - Customized MCMC scheme
  - Highly efficient model search strategy
  - Posterior distribution for the number of common factors
- Extensions
  - Nonlinear (discrete choice) factor models
  - Non-normal (mixture of) factor models