Principal components analysis (PCA) & factor analysis (FA)

Hedibert F. Lopes & Paulo Marques
INSPER Institute of Education and Research
São Paulo, Brazil

Examples

Example	Data description	Observations	Features
3	Monthly exchange rates data	144	6
6	Weekly measurements of SO_2	342	22
4	Daily exchange rates data	2650	26
1	Human development index data	5481	22
2	BOVESPA data	21802	15
7	Quarterly US economic data	202	224
5	Daily returns for S&P500 firms	2000	300

Outline

Principal components analysis

- Example 1: `hdi` data (5481×22)
- Full OLS regression
- Eigenvalues and eigenvectors
- Singular value and spectral decompositions
- Principal components analysis
- Full regression vs PCA-based regression
- Example 2: `ibovespa` data (21802×15)

Factor analysis

- Early days
- Basic model
- Example 3: Monthly exchange rates data (144×6)
- Example 4: Daily exchange rates data (2650×26)
- Example 5: Daily returns for S&P500 firms (2000×300)
- Example 6: Weekly measurements of SO_2 (342×22)

Related models

- Factor regression
- Partial least squares
- Canonical correlation analysis
- Example 7: Quarterly data on the US economy (202×224)

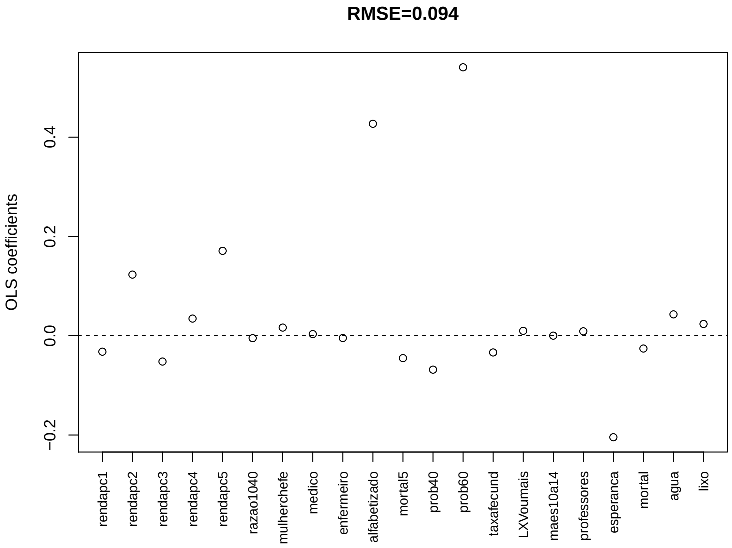
Outline

hdi data

```
RESPONSE:
14. Indice de Desenvolvimento Humano

21 PREDICTORS:
 4. Renda domiciliar per capita - media do 1o quintil (R$ por mes)
 5. Renda domiciliar per capita - media do 2o quintil (R$ por mes)
 6. Renda domiciliar per capita - media do 3o quintil (R$ por mes)
 7. Renda domiciliar per capita - media do 4o quintil (R$ por mes)
 8. Renda domiciliar per capita - media do 5o quintil (R$ por mes)
 9. Razao entre a renda dos 10% mais ricos e 40% mais pobres
10. Mulheres chefes de familia sem conjuge e com filhos menores de 15 anos (%)
11. Medicos residentes (por mil habitantes)
12. Enfermeiros residentes com curso superior (%)
13. Alfabetizados - pessoas 15 anos e mais (%)
18. Mortalidade ate cinco anos de idade (por mil nascidos vivos)
19. Probabilidade de sobrevivencia ate 40 anos (%)
20. Probabilidade de sobrevivencia ate 60 anos (%)
21. Taxa de fecundidade (%)
22. Pessoas 65 anos ou mais - morando sozinhas (%)
23. Pessoas 10 e 14 anos - mulheres com filhos (%)
24. Professores do fundamental residentes com curso superior (%)
25. Esperanca de vida ao nascer
26. Mortalidade infantil (por mil nascidos vivos)
27. Domicilios - com agua encanada - pessoas (%)
28. Domicilios - com servico de coleta de lixo - pessoas (%)
```

Full OLS regression ($p = 21$)

Full regression: $\text{hdi} = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \varepsilon$



Eigenvalues and eigenvectors[1]

Let x be a zero-mean p-dimensional vector of features with variance $\Sigma > 0$.

Eigenvalues: If Σ $(p \times p)$ is any square matrix then

$$q(\lambda) = |\Sigma - \lambda I_p|$$

is a pth order polynomial in λ. The roots $\lambda_1, \ldots, \lambda_p$ are called eigenvalues of Σ.

Eigenvectors: For each $i = 1, \ldots, p$, $|\Sigma - \lambda_i I_p| = 0$, so $\Sigma - \lambda_i I_p$ is singular. Hence, there exists a non-zero vector γ satisfying

$$\Sigma\gamma = \lambda_i \gamma.$$

Any such vector is called an eigenvector of Σ for the eigenvalue λ_i.

Symmetric matrices: All the eigenvalues of a symmetric matrix are real
Rank of a matrix:. The rank of Σ equals the number of non-zero eigenvalues.

[1] Mardia, Kent and Bibby (1979) *Multivariate Analysis*. Academic Press. Page 466-469.

Singular value and spectral decompositions

Singular value decomposition theorem

If A is an $(n \times p)$ matrix of rank r, then A can be written as

$$A = ULV' = \sum_{i=1}^{r} \ell_i u_{(i)} v'_{(i)}$$

where $U = (u_{(1)}, \ldots, u_{(r)})$ is $(n \times r)$ and $V = (v_{(1)}, \ldots, v_{(r)})$ is $(p \times r)$ are column orthonormal matrices $(U'U = V'V = I_r)$ and L is a diagonal matrix with positive elements, i.e. $L = \text{diag}(\ell_1, \ldots, \ell_r)$.

Spectral decomposition theorem, or Jordan decomposition theorem

Any symmetric matrix Σ $(p \times p)$ can be written as

$$\Sigma = \Gamma \Lambda \Gamma' = \sum_{i=1}^{p} \lambda_i \gamma_{(i)} \gamma'_{(i)}$$

where Λ is a diagonal matrix of eigenvalues of Σ, and Γ is an orthogonal matrix whose columns are standardized eigenvectors.

Principal components analysis

Recall that $E(x) = 0$ and $V(x) = \Sigma = \Gamma\Lambda\Gamma'$.

Let $y = \Gamma' x$.

By using the spectral decomposition theorem, we can see that

$$\begin{aligned} E(y) &= 0 \\ V(y) &= \Lambda \\ V(y_1) &\geq V(y_2) \geq \cdots \geq V(y_p) \geq 0 \end{aligned}$$

Result 1: No linear combination of x has variance larger than λ_1, the variance of the first principal component.

Result 2: If $\alpha = a'x$ is a linear combination of x which is uncorrelated with the first k principal components of x, then the variance of α is maximized when α is the $(k+1)$th principal component of x.

Sample principal components

Let X be our `hdi` data, i.e. a $(n \times p)$ matrix with $n = 5,481$ municipalities and its $p = 21$ features.

The sample covariance matrix is $\hat{\Sigma}$, which can be decomposed as

$$\hat{\Sigma} = GLG',$$

such that the ith principal component can be written as

$$y_{(i)} = Xg_{(i)}$$

or simply $Y = XG$, such that the sample covariance matrix of Y is L.

The vector $g_{(i)}$ are the loadings of the ith principal component.
The vector $y_{(i)}$ are the scores of the ith principal component.

A few references

Pearson (1901) On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2, 559-572.

Hotelling (1933) Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6), 417.

Tipping and Bishop (1999) Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B*, 61(3), 611-622.

Jolliffe (200) *Principal component analysis*. Wiley Online Library.

Hoff (2007) Model averaging and dimension selection for the singular value decomposition. *Journal of the American Statistical Association*, 102(478), 674-685.

Zhang and El Ghaoui (2011) Large-scale sparse principal component analysis with application to text data. In *Advances in Neural Information Processing Systems*, 532-539.

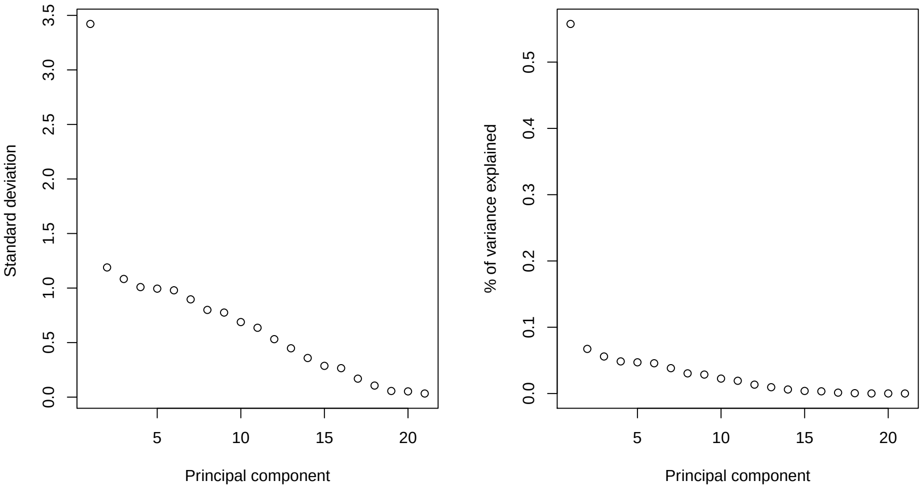
Jolliffe and Cadima (2016) Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 374.

Toy example

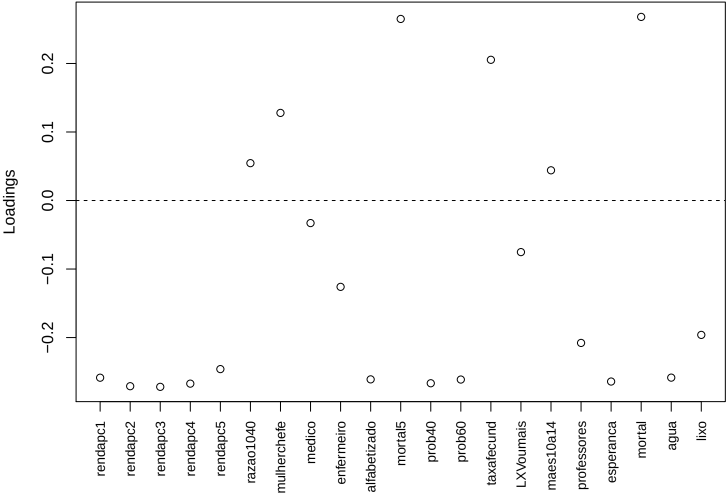
```
S = var(dados[,2:6])
          rendapc1  rendapc2  rendapc3  rendapc4  rendapc5
rendapc1 1.0000000 0.9694697 0.9368180 0.8922824 0.7548276
rendapc2 0.9694697 1.0000000 0.9871980 0.9567946 0.8208739
rendapc3 0.9368180 0.9871980 1.0000000 0.9859019 0.8535538
rendapc4 0.8922824 0.9567946 0.9859019 1.0000000 0.8876594
rendapc5 0.7548276 0.8208739 0.8535538 0.8876594 1.0000000

svd(S)
$d
[1] 4.62446396 0.28228144 0.07769785 0.01265673 0.00290002

$u
           [,1]       [,2]         [,3]        [,4]        [,5]
[1,] -0.4411378  0.4941891  0.608777172 -0.42710055  0.09027785
[2,] -0.4586014  0.2724745 -0.007324998  0.66916141 -0.51731208
[3,] -0.4613133  0.1018765 -0.368647838  0.23092054  0.76654138
[4,] -0.4571075 -0.1142474 -0.575268235 -0.55792360 -0.36849418
[5,] -0.4163232 -0.8112358  0.403115279  0.07214636  0.02940206

$v
           [,1]       [,2]         [,3]        [,4]        [,5]
[1,] -0.4411378  0.4941891  0.608777172 -0.42710055  0.09027785
[2,] -0.4586014  0.2724745 -0.007324998  0.66916141 -0.51731208
[3,] -0.4613133  0.1018765 -0.368647838  0.23092054  0.76654138
[4,] -0.4571075 -0.1142474 -0.575268235 -0.55792360 -0.36849418
[5,] -0.4163232 -0.8112358  0.403115279  0.07214636  0.02940206

round(100*svd(S)$d/sum(svd(S)$d),2)
[1] 92.49  5.65  1.55  0.25  0.06
```

Back to the `hdi` dataset

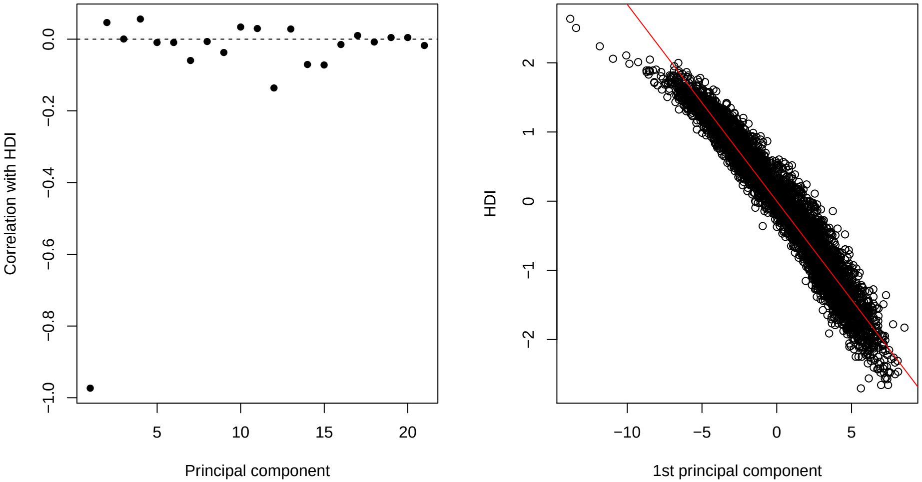
Using the `R` function `princomp`



First principal component: loadings $g_{(1)}$



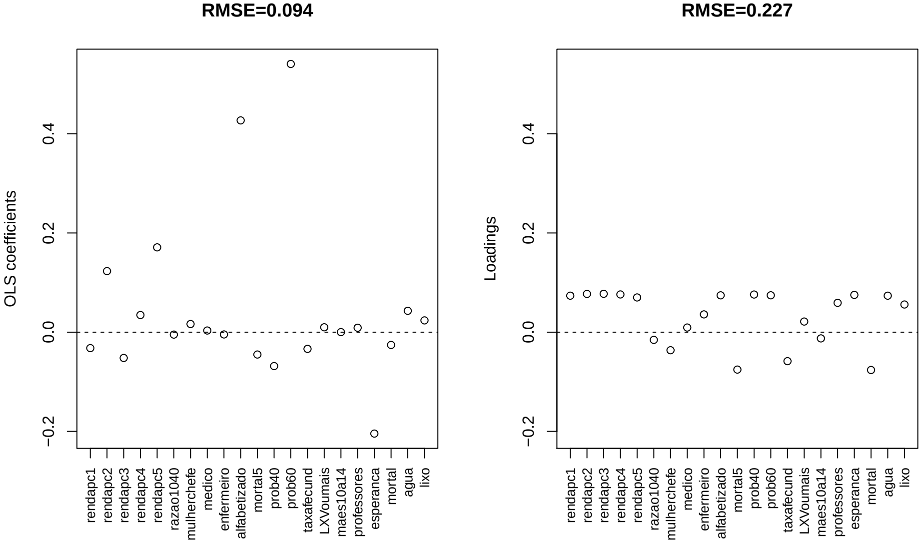
HDI vs 1st principal component



Full regression vs PCA-based regression

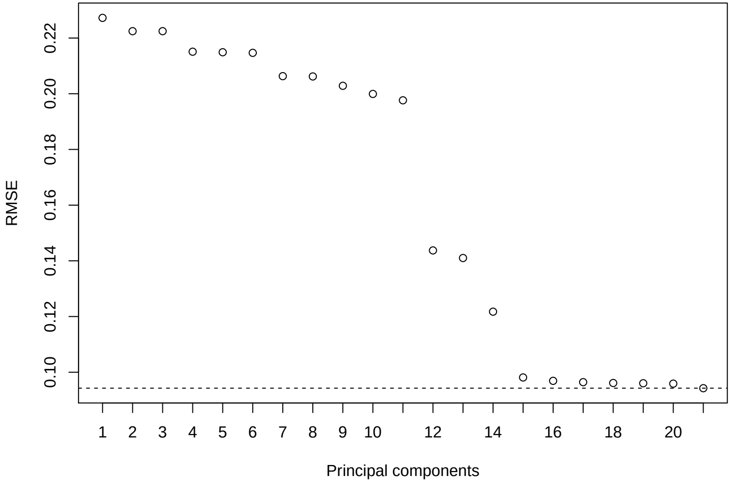
Full regression: $\text{hdi} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \varepsilon$

PCA-based regression: $\text{hdi} = \gamma_0 + \gamma_1 y_1 + \epsilon = \gamma_0 + \sum_{j=1}^{p} (\gamma_1 g_j) x_j + \epsilon$



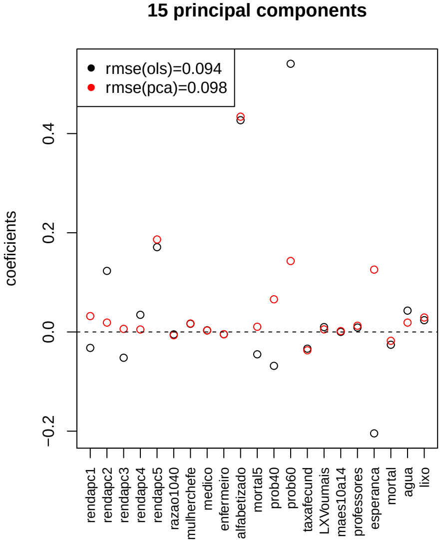
PCA-based regressions: 1 to p principal components

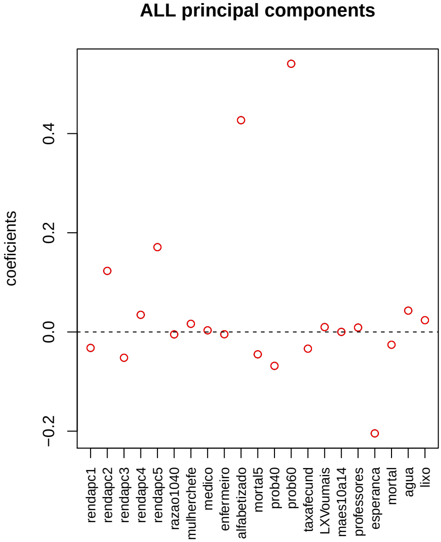
PCA regression: $\text{hdi} = \gamma_0 + \sum_{j=1}^{k} \gamma_j y_j + \epsilon$, for $k = 1, \ldots, p$.



Dashed line is RMSE of full regression.

PCA-based regressions: 15 and ALL principal components





Remarks from the `hdi` exercise

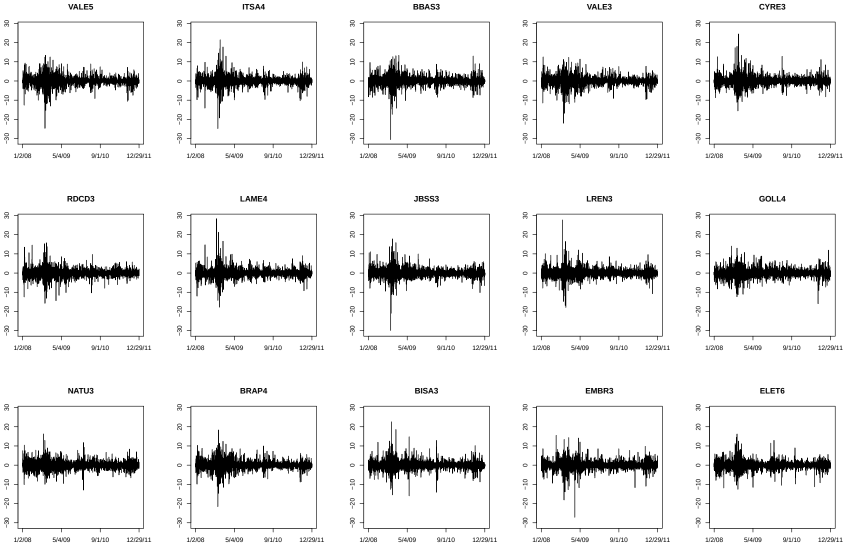
A single principal component is responsible for 50% of the variability of X.

A total of 15 principal components are needed to fit `hdi`.

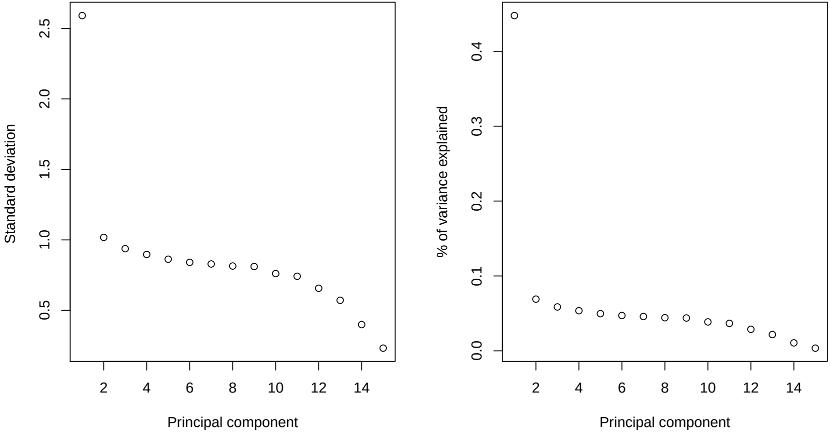
ibovespa data: January 2nd, 2008

	Date	Time.of.Day	VALE5	BVMF3	GGBR4	BBDC4	OGXP3	ITUB4	ITSA4	BBAS3
1	1/2/08	11.38	50.70	11.77	51.26	56.55	1274.9	30.4	11.78	29.91
2	1/2/08	11.63	50.40	11.77	51.46	56.60	1274.9	30.4	11.78	30.03
3	1/2/08	11.88	50.39	11.77	51.41	56.41	1274.9	30.4	11.78	30.18
4	1/2/08	12.13	50.50	11.77	51.43	56.30	1274.9	30.4	11.77	30.28
5	1/2/08	12.38	50.39	11.77	51.49	56.31	1274.9	30.4	11.76	30.42
6	1/2/08	12.63	50.31	11.77	51.60	56.15	1274.9	30.4	11.79	30.50
7	1/2/08	12.88	50.30	11.77	51.68	55.90	1274.9	30.4	11.76	30.51
8	1/2/08	13.13	50.03	11.77	51.60	55.49	1274.9	30.4	11.70	30.50
9	1/2/08	13.38	49.79	11.77	51.35	55.30	1274.9	30.4	11.67	30.46
10	1/2/08	13.63	49.93	11.77	51.40	55.70	1274.9	30.4	11.66	30.55
11	1/2/08	13.88	49.87	11.77	51.30	55.80	1274.9	30.4	11.63	30.60
12	1/2/08	14.13	49.77	11.77	50.70	55.65	1274.9	30.4	11.62	30.55
13	1/2/08	14.38	49.65	11.77	50.60	55.33	1274.9	30.4	11.55	30.55
14	1/2/08	14.63	49.35	11.77	50.55	55.01	1274.9	30.4	11.45	30.50
15	1/2/08	14.88	49.70	11.77	50.78	55.14	1274.9	30.4	11.49	30.55
16	1/2/08	15.13	49.64	11.77	50.65	55.03	1274.9	30.4	11.45	30.61
17	1/2/08	15.38	49.45	11.77	50.56	54.62	1274.9	30.4	11.44	30.67
18	1/2/08	15.63	49.45	11.77	50.76	54.60	1274.9	30.4	11.46	30.72
19	1/2/08	15.88	49.45	11.77	50.71	54.20	1274.9	30.4	11.47	30.51
20	1/2/08	16.13	49.50	11.77	50.57	54.34	1274.9	30.4	11.45	30.51
21	1/2/08	16.38	49.48	11.77	50.80	54.47	1274.9	30.4	11.44	30.60
22	1/2/08	16.63	49.50	11.77	50.61	54.48	1274.9	30.4	11.44	30.65

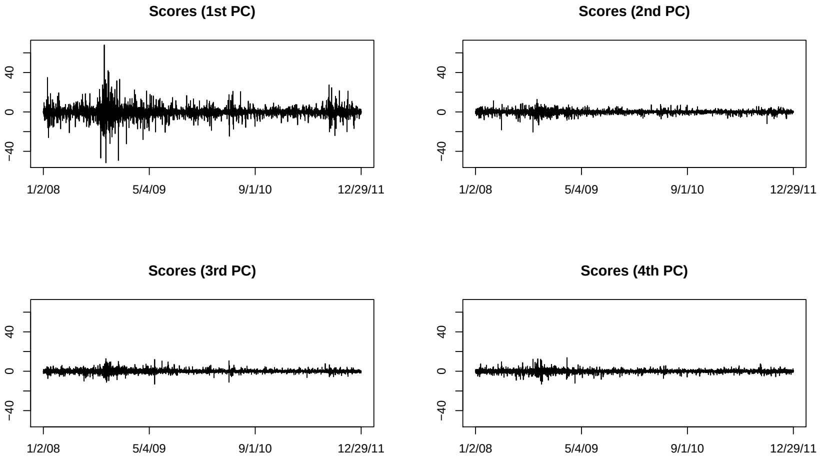
Standardized returns



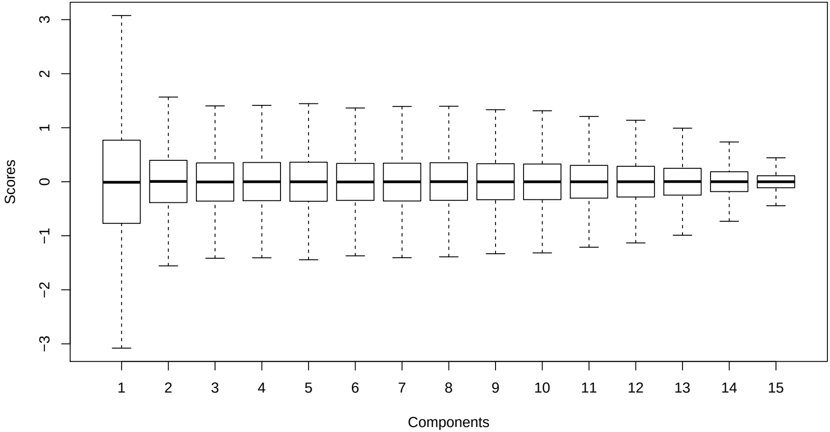
% of explained variance



Principal components: scores



Principal components: scores

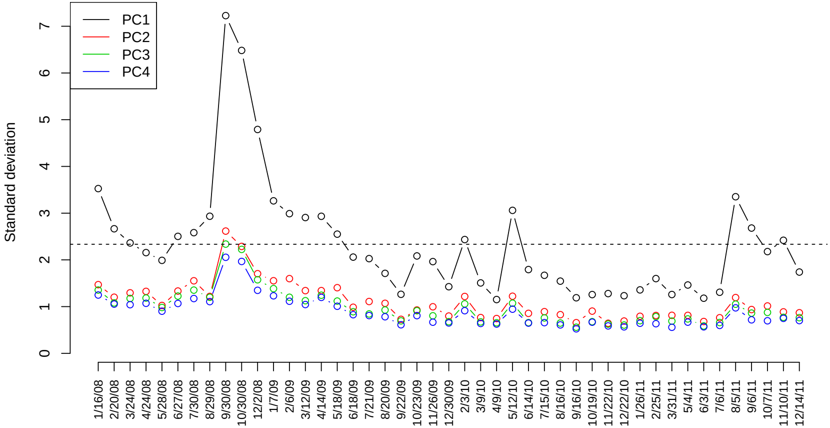


PCA over partition of the data

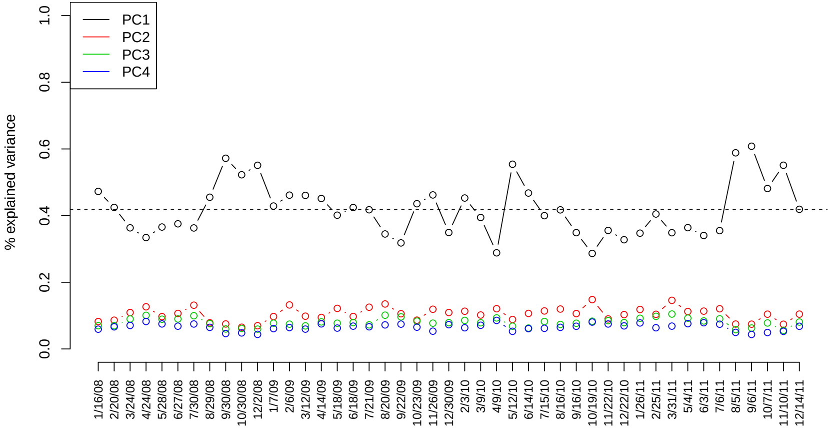
Each subset has 482 observations, but the last which has 503

1	1/2/08	2/1/08
2	2/6/08	3/6/08
3	3/7/08	4/8/08
4	4/9/08	5/12/08
5	5/13/08	6/12/08
6	6/13/08	7/15/08
7	7/16/08	8/14/08
8	8/15/08	9/15/08
9	9/16/08	10/15/08
10	10/16/08	11/14/08
11	11/17/08	12/17/08
12	12/18/08	1/22/09
13	1/23/09	2/25/09
14	2/26/09	3/27/09
15	3/30/09	4/30/09
16	5/4/09	6/2/09
17	6/3/09	7/3/09
18	7/6/09	8/5/09
19	8/6/09	9/4/09
20	9/8/09	10/7/09
21	10/8/09	11/10/09
22	11/11/09	12/11/09
23	12/14/09	1/18/10
24	1/19/10	2/22/10
25	2/23/10	3/24/10
26	3/25/10	4/27/10
27	4/28/10	5/27/10
28	5/28/10	6/29/10
29	6/30/10	7/30/10
30	8/2/10	8/31/10
31	9/1/10	10/1/10
32	10/4/10	11/4/10
33	11/5/10	12/7/10
34	12/8/10	1/10/11
35	1/11/11	2/10/11
36	2/11/11	3/16/11
37	3/17/11	4/15/11
38	4/18/11	5/19/11
39	5/20/11	6/20/11
40	6/21/11	7/21/11
41	7/22/11	8/22/11
42	8/23/11	9/22/11
43	9/23/11	10/25/11
44	10/26/11	11/28/11
45	11/29/11	12/29/11

First 4 principal components: standard deviations



First 4 principal components: % of explained variance



PCA and factor analysis

Principal components analysis is a dimension-reduction and projection tool for high dimensional matrices.

Factor analysis is a modeling framework for high dimensional and highly structure data.

Outline

Factor analysis: early days

Bartholomew (1995)[2] starts his paper by saying that

> *Spearman invented factor analysis but his almost exclusive concern with the notion of a general factor prevented him from realizing its full potential.*

Factor analysis, however, has flourished ever since Spearman's (1904) seminal paper on the American Journal of Psychology (Vol 15, pp. 201-292) entitled "General Inteligente objectively determined and measured".

[2] Spearman and the origin and development of factor analysis, *British Journal of Mathematical and Statistical Psychology*, 48, 211-220.

Spearman's general intelligence

Psychologists were trying to define intelligence by a single, all-encompassing unobservable entity, the g factor.

Spearman studied the influence of the g factor on examinees test scores on several domains:

- Pitch
- Light
- Weight
- Classics
- French
- English
- Mathematics

End of the day: Postulating g provides a mechanism to detect common correlations among such variables.

Spearman's one-factor model

One-factor model:

$$
\begin{aligned}
y_{i1} &= \mu_1 + \lambda_1 g_i + \epsilon_{i1} \\
y_{i2} &= \mu_2 + \lambda_2 g_i + \epsilon_{i2} \\
&\vdots \\
y_{im} &= \mu_m + \lambda_m g_i + \epsilon_{im}
\end{aligned}
$$

where

- y_{ij}: score of examinee i on test domain j.
- μ_j: mean of test domain j.
- g_i: value of the intelligence factor for person i.
- λ_j: loading of test domain j onto the intelligence factor g.
- ϵ_{ij}: random error term for person i and test domain j.

Multiple factor analysis

Factor models are mainly applied in two major situations:

1. Identifying underlying structures.
2. Data reduction.

Basic model

The Gaussian linear factor model relates a m-vector of observables y_t to a r-vector of latent variables f_t via

$$y_t | f_t, \Lambda, \Sigma \sim N(\Lambda f_t, \Sigma).$$

Common factors:

$$f_t | \Lambda, \Sigma \sim N(0, I_r).$$

Specific/idiosyncratic factor variances:

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

Variance structure

Conditional variance

The common latent factors explain all the dependence structure among the m variables:

$$\text{cov}(y_{it}, y_{jt}|f_t, \Lambda, \Sigma) = \begin{cases} \sigma_i^2 & i = j \\ 0 & i \neq j \end{cases}$$

Unconditional variance

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

Invariance

The factor model is invariant to orthogonal transformations, i.e.

$$\tilde{\Lambda} = \Lambda P' \quad \text{and} \quad \tilde{f}_t = P f_t,$$

for any orthogonal matrix P, such that

$$V(y_t|\Lambda, \Sigma) = \Lambda\Lambda' + \Sigma = \tilde{\Lambda}\tilde{\Lambda}' + \Sigma$$

Dealing with invariance

Classical approach: Orthogonality of the columns of Λ

$$\Lambda' \Sigma^{-1} \Lambda = I_r$$

Bayesian approach: Λ is block lower triangular

$$\Lambda = \begin{pmatrix} \lambda_{11} & 0 & 0 & \cdots & 0 \\ \lambda_{21} & \lambda_{22} & 0 & \cdots & 0 \\ \lambda_{31} & \lambda_{32} & \lambda_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda_{r1} & \lambda_{r2} & \lambda_{r,r-1} & \cdots & \lambda_{rr} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda_{m1} & \lambda_{m2} & \lambda_{m,r-1} & \cdots & \lambda_{mr} \end{pmatrix}$$

Frühwirth-Schnatter and Lopes (2009/2017): Λ is generalized block lower triangular. Our generalization provides both identification and, often, useful interpretation of the factor model.

Classical literature

- Thurstone (1934)
- Bartlett (1937)
- Lawley (1940,1941)
- Kendall and Smith (1950)
- Anderson and Rubin (1956)
- Jöreskog (1969,1970)
- Rubin and Thayer (1982)
- Bentler and Tanaka (1983)
- Rubin and Thayer (1983)
- Akaike (1987)
- Anderson and Amemiya (1988)
- Amemiya and Anderson (1990)

Bayesian literature

Pre-MCMC

- Press (1972)
- Martin and McDonald (1975)
- Geweke and Singleton (1980)
- Bartholomew (1981)
- Lee (1981)
- Press and Shigemasu (1989)

Post-MCMC

- Geweke and Zhou (1996)
- Aguilar and West (2000)
- Lopes, Aguilar and West (2000)
- Lopes and Migon (2002)
- West (2003)
- Wang and Wall (2003)
- Lopes and West (2004)
- Quinn (2004)
- Hogan and Tchernis (2004)
- Lopes, Salazar and Gamerman (2008)
- Carvalho *et al.* (2008)
- Chib and Ergashev (2009)
- Frühwirth-Schnatter and Lopes (2009/2017)
- Carvalho, Lopes and Aguilar (2011)
- Lopes, Schmidt, Salazar, Gomez and Achkar (2012)
- Bhattacharya and Dunson (2011)
- Lopes, Conti, Heckman and Piatek (2012)
- Hahn, He and Lopes (2017)
- Kastner, Frühwirth-Schnatter and Lopes (2017)

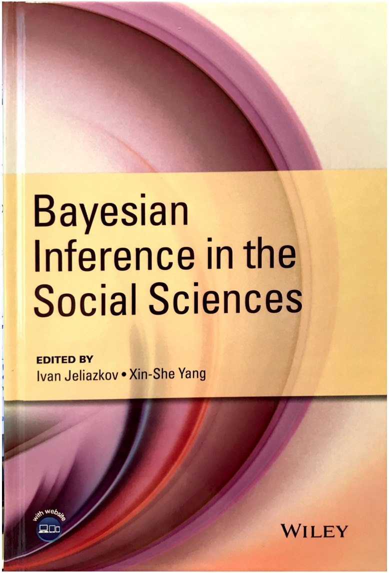
Factor analysis by area[3]

	1904 -1980	1981 -1985	1986 -1990	1991 -1995	1995 -2000	2000 -2004	Total
Biology	18	17	20	23	47	41	166
Chemistry	12	14	36	53	88	77	244
Chromatography	4	7	16	22	24	15	88
Ecology	2	4	11	15	61	45	138
Economics	14	12	9	4	20	26	85
Food	1	4	5	2	17	21	50
Geriatry	8	5	10	9	25	31	88
Image Processing	2	7	22	27	38	51	151
Industry	4	0	2	6	38	28	78
Magnetic Resonance	1	1	3	6	25	13	49
Medicine	30	32	64	67	109	116	418
Methodology	10	25	31	49	125	151	391
Operational Research	1	1	1	9	42	41	95
Physiology	20	26	38	39	51	29	203
Psychiatry	15	14	39	61	137	99	365
Psychology	93	86	159	219	379	344	1287
Spectroscopy	11	27	40	50	108	90	326
(a) Total FA-papers	196	242	408	545	1065	1002	3460
(b) All papers($*10^3$)	5186	1518	1890	2117	2430	1999	14707
(c) FA/All($*10^{-6}$)	38	159	216	257	438	501	235

Table 1. Distribution of papers on factor analysis in the Internet.

[3] Kaplunovsky (2004) Why using factor analysis?

Modern Bayesian factor analysis



CHAPTER 5

MODERN BAYESIAN FACTOR ANALYSIS

Hedibert Freitas Lopes

George Washington University, D.C., USA

5.1 Introduction

The origin of factor analysis can be traced back to Spearman's (1904) seminal paper on general intelligence. At the time, psychologists were trying to define intelligence by a single, all-encompassing unobservable entity, the g factor. Spearman studied the influence of the g factor on examinees' test scores on several domains: pitch, light, weight, classics, French, English, and mathematics. At the end of the day, the g factor would provide a mechanism to detect common correlations among such imperfect measurements. More precisely, Spearman's (1904) one-factor model based on p test domains (measurements) and n examinees (individuals) can be written as

$$y_{ij} = \mu_j + \beta_j g_i + \varepsilon_{ij}, \tag{5.1}$$

Prior specification

Loadings

For $i = 1, \ldots, r$

$$\lambda_{ii} \sim N_{(0,\infty)}(m_0, C_0)$$

For $i = 2, \ldots, r$ and $j = 1, \ldots, \min\{i-1, r\}$

$$\lambda_{ij} \sim N(m_0, C_0)$$

Idiosyncratic variances

For $i = 1, \ldots, m$

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

The hyperparameters m_0, C_0, ν and s^2 are known.

Posterior inference via Gibbs sampler

A factor model is a multivariate regression model when deriving the full conditional distributions of $p(\Lambda, f, \Sigma | y)$.

The two easiest ones are the full conditional distributions of the common factors $f_1, \ldots, f_n$ and the idiosyncratic variances $\sigma_1^2, \ldots, \sigma_m^2$.

Let $y = (y_1', \ldots, y_n')' = (y_{(1)}, \ldots, y_{(m)})$, $\Lambda = (\lambda_1, \ldots, \lambda_m)'$ and $F_i = (f_{(1)}, \ldots, f_{(i)})$ for $i = 1, \ldots, r$, with $F = F_r$.

$(f_i | \Lambda, \sigma, y)$: For $i = 1, \ldots, n$,

$$(f_i | \Lambda, \Sigma, y) \sim N((I_k + \Lambda'\Sigma^{-1}\Lambda)^{-1}\Lambda'\Sigma^{-1}y_i, (I_k + \Lambda'\Sigma^{-1}\Lambda)^{-1}).$$

$(\sigma_i^2 | \Lambda, f, y)$:. For $i = 1, \ldots, m$,

$$(\sigma_i^2 | \Lambda, f, y) \sim IG\left(\frac{\nu + n}{2}, \frac{\nu s^2 + (y_{(i)} - F\lambda_i)'(y_{(i)} - F\lambda_i)}{2}\right).$$

Full conditional of Λ, $(\Lambda|\sigma, f, y)$

The identifiability constraints are such that, for $i = 1, \ldots, r-1$,

$$\lambda_i = (\tilde{\lambda}_i', 0_{r-i}')'.$$

For $i = 1, \ldots, r$:

$$(\tilde{\lambda}_i|\Sigma, f, y) \sim N(m_i, C_i)1\{\tilde{\lambda}_{ii} > 0\},$$

where

$$\begin{aligned} m_i &= C_i(C_0^{-1}m_0 1_i + \sigma_i^{-2}F_i' y_{(i)}) \\ C_i^{-1} &= C_0^{-1}I_i + \sigma_i^{-2}F_i' F_i \end{aligned}$$

For $i = r+1, \ldots, m$:

$$(\lambda_i|\Sigma, f, y) \sim N(m_i, C_i),$$

where

$$\begin{aligned} m_i &= C_i\left(C_0^{-1}m_0 1_r + \sigma_i^{-2}F' y_{(i)}\right) \\ C_i^{-1} &= C_0^{-1}I_r + \sigma_i^{-2}F' F \end{aligned}$$

Exchange rate data (Lopes and West, 2004)

- Monthly exchange rates from January 1975 to December 1986.
- Time series are the exchange rates in British pounds of
  - US dollar (US) and Canadian dollar (CAN)
  - Japanese yen (JAP)
  - French franc (FRA), Italian lira (ITA) and German Deutschmark (GER)
- The prior hyperparameters are
  - Informative prior: $(m_0, C_0^{-1}, \nu_0, s^2) = (0, 1, 2.2, 0.0455)$
    Prior mode of σ_i is 0.154.
  - Noninformative prior: $(m_0, C_0^{-1}, \nu_0, s^2) = (0, 0, 0.001, 1)$
    Prior mode of σ_i^2 is 0.032.
- We burn-in the Gibbs sampler for 10,000 iterations, and then save equally spaced samples of 5,000 draws from a longer run of 100,000.
- It takes about one minute to run a two-factor model (in `R`) on my MacBook Pro with a 2.6GHz Intel Core i7 processor, 8 GB 1600 MHz DDR3 Memory running a Mac OS X Lion 10.7.5.

Posterior means

The posterior means of Σ and Λ' in a two-factor model are

$$E(\Sigma|y) = \text{diag}(0.05, 0.13, 0.62, 0.04, 0.25, 0.26),$$

and

$$E(\Lambda'|y) = \begin{pmatrix} 1.00 & 0.96 & 0.46 & 0.39 & 0.42 & 0.41 \\ 0.00 & 0.05 & 0.43 & 0.92 & 0.78 & 0.78 \end{pmatrix},$$

respectively.

One can argue that

- The first common factor groups North American currencies, and
- The second common factor groups European currencies.

Ordering of the variables

1st ordering

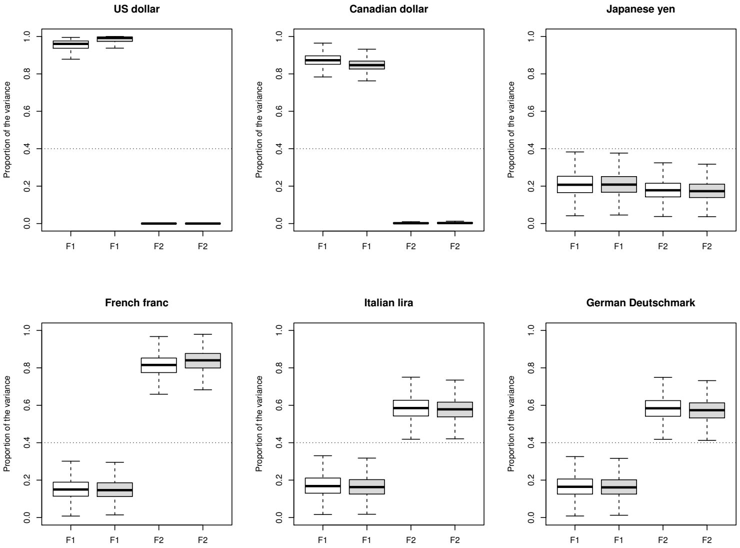
$$\widehat{\Lambda} = \begin{pmatrix} \text{US} & 1.00 & 0.00 \\ \text{CAN} & 0.96 & 0.05 \\ \text{JAP} & 0.46 & 0.43 \\ \text{FRA} & 0.39 & 0.92 \\ \text{ITA} & 0.42 & 0.78 \\ \text{GER} & 0.41 & 0.78 \end{pmatrix} \qquad \widehat{\Sigma} = \text{diag}\begin{pmatrix} 0.05 \\ 0.13 \\ 0.62 \\ 0.04 \\ 0.25 \\ 0.26 \end{pmatrix}$$

2nd ordering

$$\widehat{\Lambda} = \begin{pmatrix} \text{US} & 0.98 & 0.00 \\ \text{JAP} & 0.45 & 0.42 \\ \text{CAN} & 0.95 & 0.03 \\ \text{FRA} & 0.39 & 0.91 \\ \text{ITA} & 0.41 & 0.77 \\ \text{GER} & 0.40 & 0.77 \end{pmatrix} \qquad \widehat{\Sigma} = \text{diag}\begin{pmatrix} 0.06 \\ 0.62 \\ 0.12 \\ 0.04 \\ 0.25 \\ 0.26 \end{pmatrix}$$

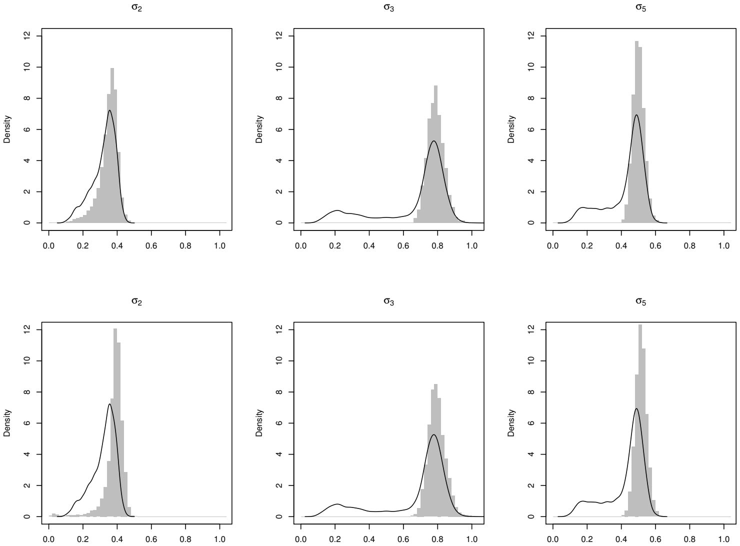
Variance decomposition

The proportion of the variance of currency i attributed to factor j is given by $\nu_{ij} = \beta_{ij}^2/(\lambda_{i1}^2 + \lambda_{i2}^2 + \sigma_i^2)$. Informative prior (white boxplots). Noninformative prior (grey boxplots).



Marginal posteriors based on 2- and 3-factor models

Two-factor model (histograms) and a (overfitted) three-factor model (solid lines).
Informative prior (top row) and noninformative prior (bottom row).



Factor stochastic volatility[4]

For each point in time $t = 1, \ldots, T$,

- m observed returns: $y_t = (y_{1t}, \ldots, y_{mt})'$
- r unobserved factors: $f_t = (f_{1t}, \ldots, f_{rt})'$
- Volatilities: $h_t = (h_t^U, h_t^V)$, $h_t^U = (h_{1t}, \ldots, h_{mt})'$ and $h_t^V = (h_{1,m+1}, \ldots, h_{m+r,t})'$.

Our factor stochastic volatility model is

$$
\begin{aligned}
y_t | f_t &\sim N(\Lambda f_t, U_t) \\
f_t &\sim N(0, V_t)
\end{aligned}
$$

where

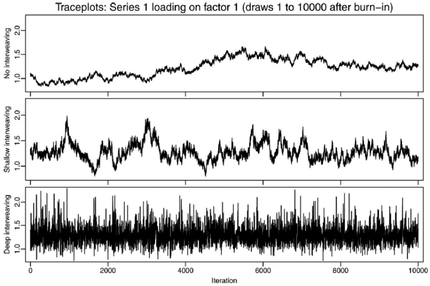
- Factor loadings: Λ is $m \times r$
- Idiosyncratic variance: $U_t = \text{diag}(\exp(h_{1t}, \ldots, h_{mt}))$
- Factor variance: $V_t = \text{diag}(\exp(h_{m+1,t}, \ldots, h_{m+r,t}))$
- Log-volatilities:

$$
\begin{aligned}
h_{it} &= (1 - \phi_i)\mu_i + \phi_i h_{i,t-1} + \sigma_i \eta_{it} \qquad i = 1, \ldots, m \\
h_{it} &= \phi_i h_{i,t-1} + \sigma_i \eta_{it} \qquad i = m+1, \ldots, m+r.
\end{aligned}
$$

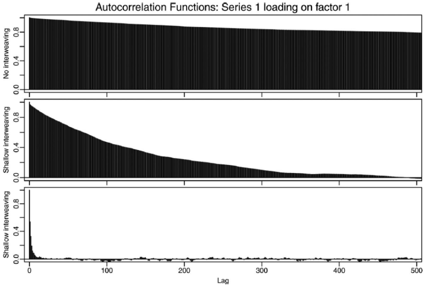
[4] Kastner, Frühwirth-Schnatter & Lopes (2017) Efficient Bayesian inference for multivariate FSV models. *Journal of Computational and Graphical Statistics*.

Shallow/deep interweaving[5]

Trace plots (10,000)



ACF (5,000,000)

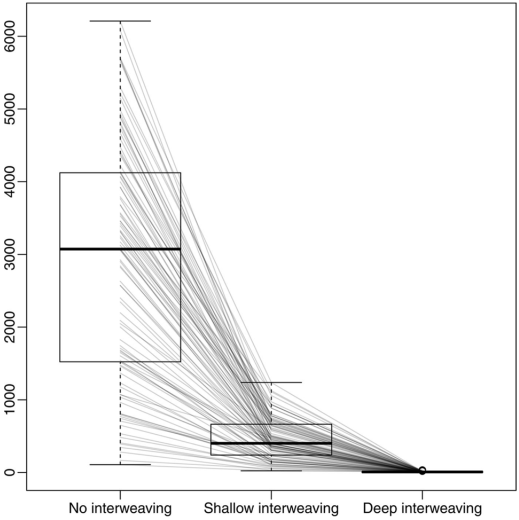


Top: Standard sampler
Middle: Shallow interweaving
Bottom: Deep interweaving

[5] Yu and Meng (2011) To Center or not to Center: That is not the Question – An Ancillarity-Suffiency Interweaving Strategy (ASIS) for Boosting MCMC Efficiency, *Journal of Computational and Graphical Statistics*, 20, 531-570.

Inefficiency factor (based on Λ_{11})

Estimated inefficiency factors for draws from $p(\Lambda_{11}|y^{[i]})$, where $y^{[i]}$, $i \in \{1, \ldots, 100\}$,denote artificially generated datasets whose underlying parameters are identical.



Application to exchange rate data

We analyze exchange rates with respect to EUR.

Data were obtained from the European Central Bank's Statistical Data Warehouse and ranges from April 1, 2005 to August 6, 2015.

It contains $m = 26$ daily exchange rates on 2650 days.

Table 3. Currency abbreviations.

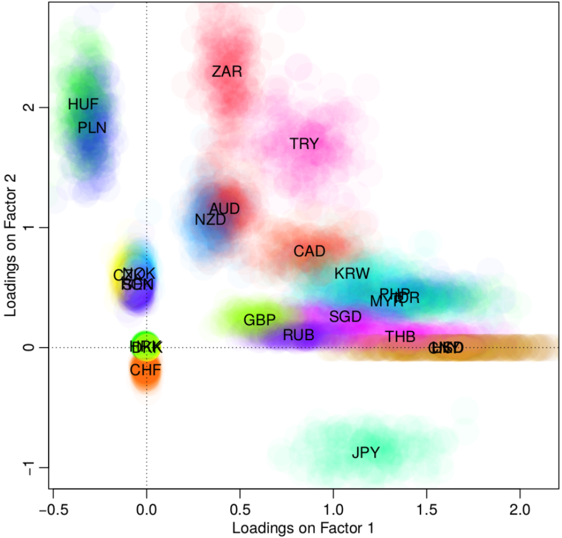
AUD	Australia dollar	CAD	Canada dollar	CHF	Switzerland franc
CNY	China yuan renminbi	CZK	Czech R. koruna	DKK	Denmark krone
GBP	UK pound	HKD	Hong Kong dollar	HRK	Croatia kuna
HUF	Hungary forint	IDR	Indonesia rupiah	JPY	Japan yen
KRW	South Korea won	MYR	Malaysia ringgit	NOK	Norway krone
NZD	New Zealand dollar	PHP	Philippines peso	PLN	Poland zloty
RON	Romania fourth leu	RUB	Russia ruble	SEK	Sweden krona
SGD	Singapore dollar	THB	Thailand baht	TRY	Turkey lira
USD	US dollar	ZAR	South Africa rand		

Posterior means of factor loadings

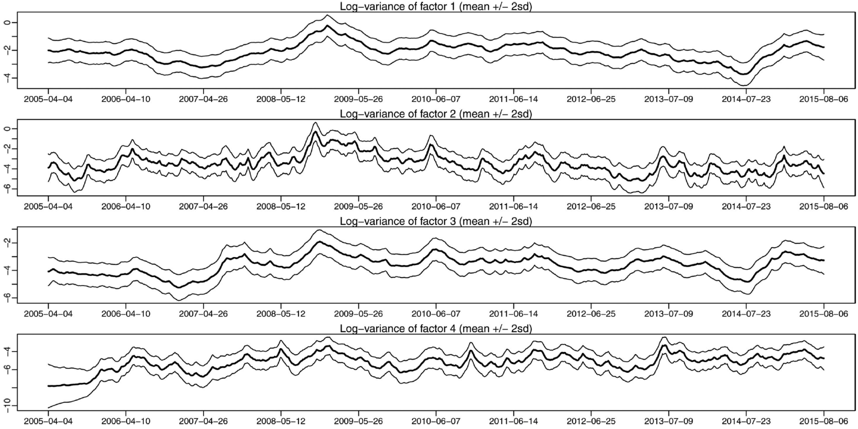
Table 4. Posterior means of $p(\mathbf{\Lambda}|y)$, in alphabetical order. Blank entries signify that the respective marginal distribution is not bound away from zero with at least 99% posterior probability. Starred entries are those which have been set to zero a priori.

	$\Lambda_{\bullet,1}$	$\Lambda_{\bullet,2}$	$\Lambda_{\bullet,3}$	$\Lambda_{\bullet,4}$		$\Lambda_{\bullet,1}$	$\Lambda_{\bullet,2}$	$\Lambda_{\bullet,3}$	$\Lambda_{\bullet,4}$
AUD	0.418	1.156	2.772	*	MYR	1.285	0.391	0.587	2.439
CAD	0.873	0.805	1.389		NOK		0.619	0.704	
CHF		−0.184			NZD	0.342	1.066	2.665	
CNY	1.592			0.076	PHP	1.330	0.449	0.389	1.702
CZK	− 0.099	0.605			PLN	− 0.292	1.835	*	*
DKK	0.002				RON	− 0.051	0.530		
GBP	0.605	0.230	0.627		RUB	0.813	0.104	0.138	0.237
HKD	1.611		0.003	0.005	SEK	− 0.049	0.529	0.527	
HRK					SGD	1.065	0.260	0.642	1.463
HUF	− 0.339	2.028			THB	1.358	0.092	0.273	1.049
IDR	1.395	0.419	0.347	1.153	TRY	0.845	1.702	0.549	0.920
JPY	1.176	− 0.875	0.310	0.904	USD	1.614	*	*	*
KRW	1.100	0.617	0.750	1.935	ZAR	0.431	2.303	1.219	1.390

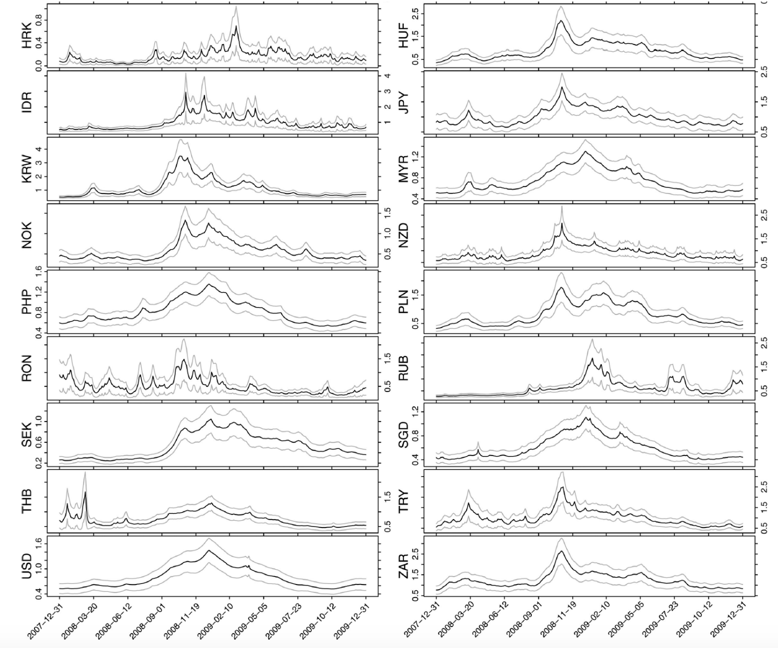
Marginal posteriors of the first two factor loadings



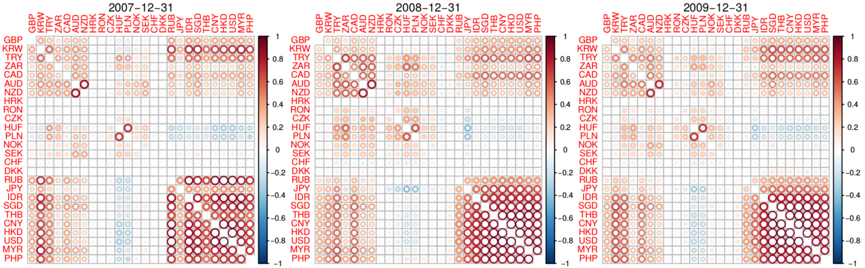
Marginal posteriors of the factor log-variances



Posterior volatilities of log returns (18 of 26 countries)



Posterior correlation matrices



R-package factorstochvol containing code to run the samplers described in the article. Available at https://cran.rproject.org/package=factorstochvol.

Sparse FSV[6]

Kastner (2017) applies a sparse 4-factor SV to model stock prices listed in the Standard & Poor's 500 index.

A total of $m = 300$ firms were continuously included in the index.
Time span: 5/3/2006 to 12/31/2013 ($T = 2000$).

GICS sector	Members
Consumer Discretionary	45
Consumer Staples	28
Energy	23
Financials	54
Health Care	30
Industrials	42
Information Technology	27
Materials	23
Telecommunications Services	3
Utilities	25

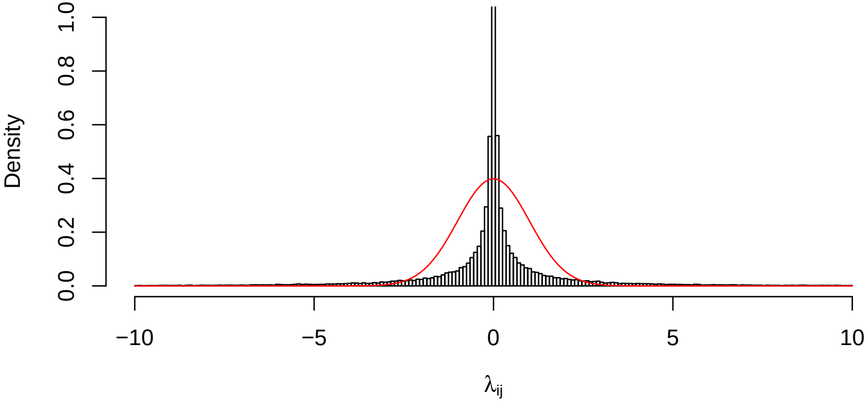
[6] Kastner (2018) Bayesian Time-Varying Covariance Estimation in Many Dimensions. *Journal of Econometrics*.

Sparse loadings matrix

Prior for loadings Λ with row-wise shrinkage with element-wise adaption:

$$\Lambda_{ij}|\tau_{ij}^2 \sim N(0,\tau_{ij}^2), \quad \tau_{ij}^2 \sim G(a_i, a_i\lambda_i/2) \quad \text{and} \quad \lambda_i^2 \sim G(c_i, d_i).$$

They used $a_i = 0.1$ and $c_i = d_i = 1$, for all $i = 1, \ldots, m = 300$.



Density goes all the way to 4. Standard normal in red.

MCMC set up

Normal-Gamma prior with row-wise shrinkage for 110,000 draws.

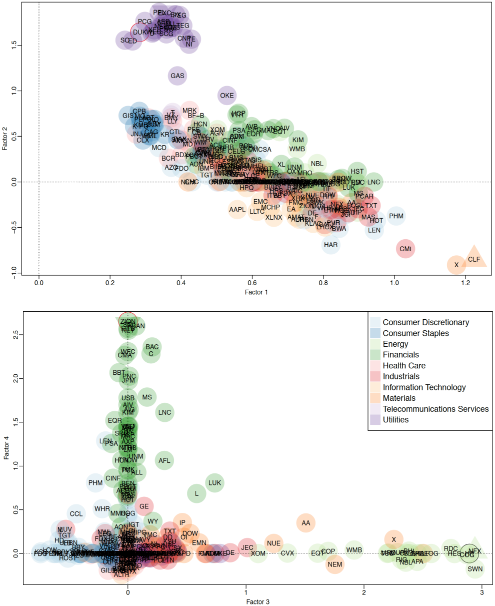
Discard the first 10,000 draws as burn-in.

Of the remaining 100,000 draws every 10th draw is kept.

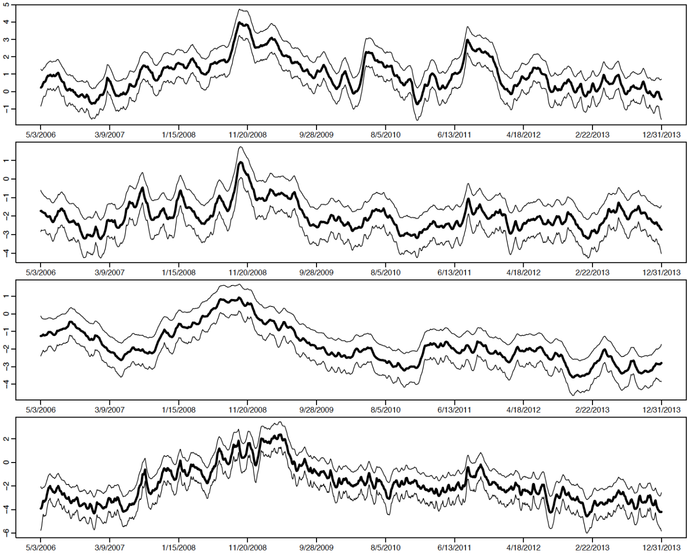
Posterior inference based on $10,000$ draws.

Factor loadings

F1: overall; F2: Utilities; F3: Energy & Materials; F4: Financial

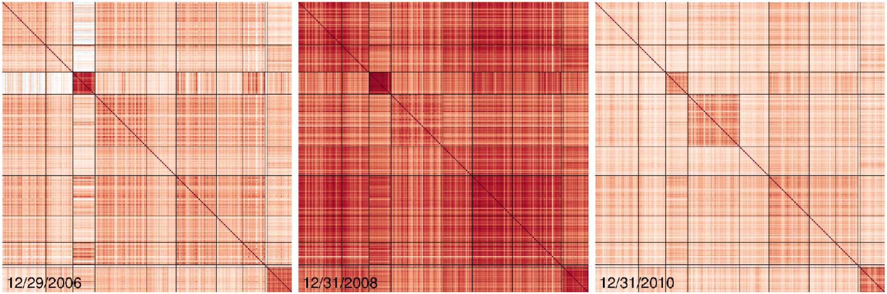


Log-variances of common factors



Posterior mean of the time-varying correlation matrix

Last trading day in 2006, 2008, 2010.



Spatial dynamic factor models

Lopes, Salazar and Gamerman (2008) introduces the following spatio-temporal model for $y_t = (y_{1t}, \ldots, y_{mT})'$, measurements on m spatial locations and over T time periods:

Dimension reduction:

$$y_t \sim N(\Lambda f_t, \Sigma)$$

Time series component:

$$f_t \sim N(\Gamma f_{t-1}, \Gamma)$$

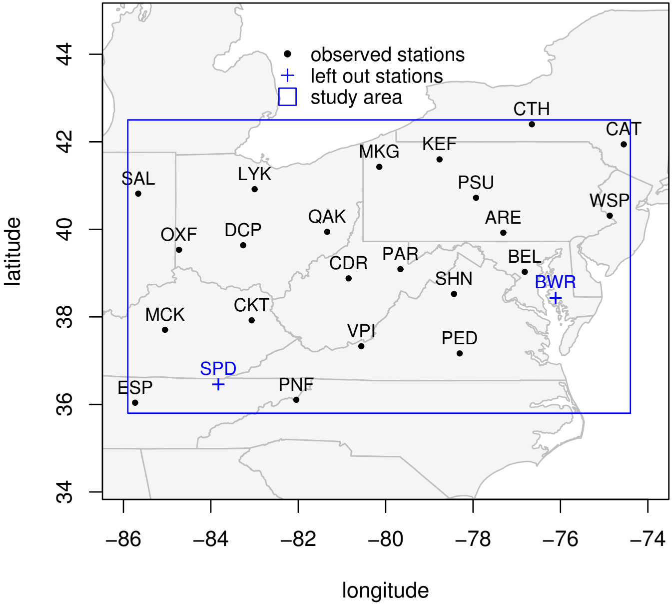
Spatial component:

$$\lambda_j \sim GP(\mu_j, \tau_j^2 R_{\phi_j})$$

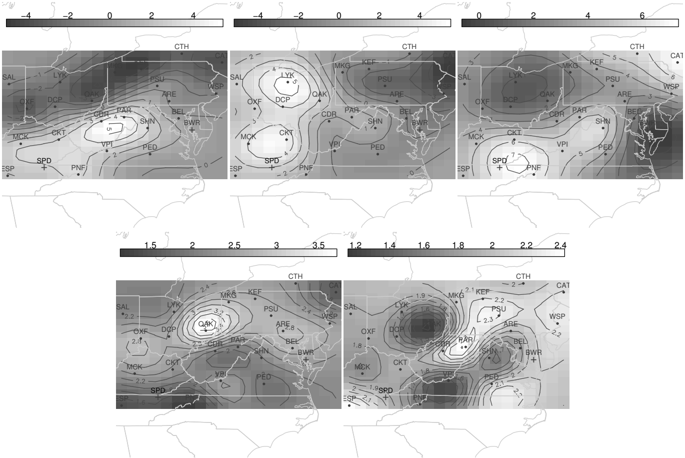
where $\Lambda = (\lambda_1, \ldots, \lambda_r)$ and R_{ϕ_j} spatial correlation matrix.

A Reversible Jump MCMC is proposed to select r, as in Lopes and West (2004).

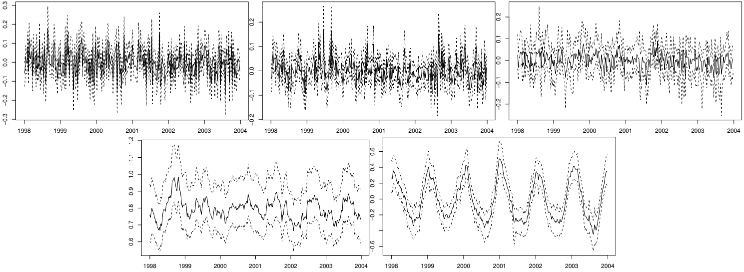
SO_2 in Eastern US



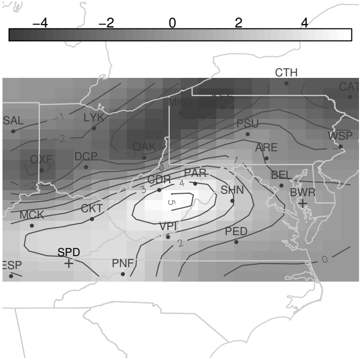
Spatial loadings

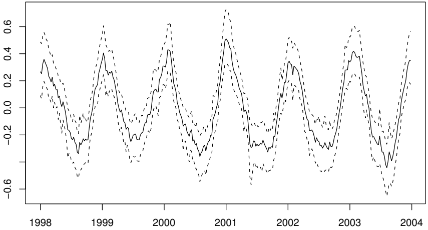


Dynamic factors

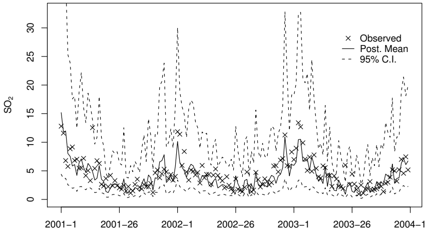


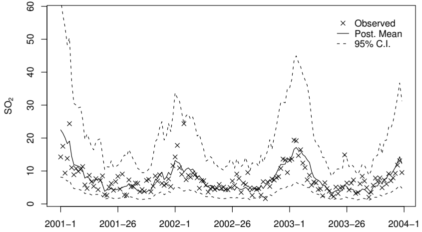
Seasonal loadings and factor





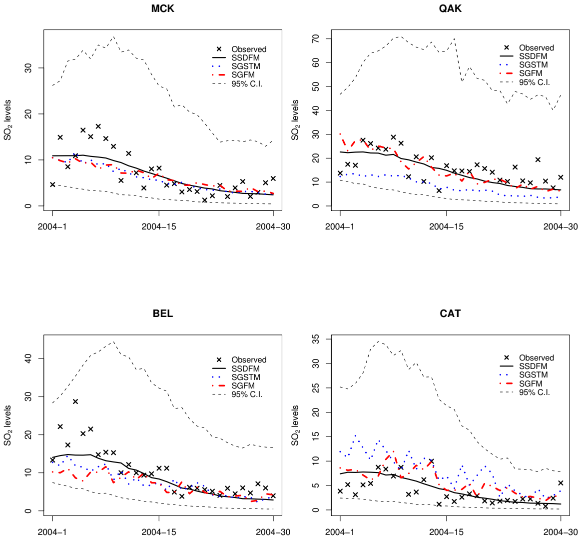
Spatial interpolation





Interpolated values at stations SPD and BWR.

Forecasting



Outline

Latent factor regression[7]

Recall the `hdi` data, where y_i is the DHI and x_i is the m-dimensional vector of characteristics for municipality i. A latent r-factor regression model is:

$$
\begin{aligned}
p(f_i) &= N(0, I_r) \\
p(x_i|f_i) &= N(\Lambda f_i, \Psi) \\
p(y_i|f_i) &= N(\beta' f_i, \sigma^2)
\end{aligned}
$$

so

$$
(x_i, y_i) \sim N(0, WW' + \Omega) \qquad W = \begin{pmatrix} \Lambda \\ \beta' \end{pmatrix} \text{ and } \Omega = \begin{pmatrix} \Psi & 0 \\ 0 & \sigma^2 \end{pmatrix}
$$

It can be shown that

$$
y_i|x_i \sim N(x_i'\theta, \tau^2)
$$

where $\theta = V_x^{-1}\Lambda\beta$, $\tau^2 = \sigma^2 + \beta'(I_m - \Lambda' V_x^{-1}\Lambda)\beta$, and $V_x = \Lambda\Lambda' + \Psi$.

[7] West (2003) called this Bayesian factor regression. When $\Psi = \psi I_m$, Yu *et al.* (2006) call this supervised PCA. See Chapter 12 of Murphy (2012) *Machine Learning: A Probabilistic Perspective*.

Partial least squares

The technique of partial least squares (PLS) (Gustafsson 2001; Sun et al. 2009) is an asymmetric or more "discriminative" form of supervised PCA. The key idea is to allow some of the (co)variance in the input features to be explained by its own subspace, f_i^x, and to let the rest of the subspace, f_i^s, be shared between input and output:

$$\begin{aligned} p(f_i) &= N(f_i^s; 0, I_{r_s}) N(f_i^x; 0, I_{r_x}) \\ p(x_i|f_i) &= N(\Lambda_s f_i^s + \Lambda_x f_i^x, \Psi) \\ p(y_i|f_i) &= N(\beta_s' f_i^s, \sigma^2) \end{aligned}$$

so

$$(x_i, y_i) \sim N(0, WW' + \Omega) \qquad W = \begin{pmatrix} \Lambda_s & \Lambda_x \\ \beta_s' & 0 \end{pmatrix} \text{ and } \Omega = \begin{pmatrix} \Psi & 0 \\ 0 & \sigma^2 \end{pmatrix}$$

Again,

$$y_i|x_i \sim N(x_i'\theta, \tau^2)$$

where $\theta = V_x^{-1}\Lambda_s\beta_s$, $\tau^2 = \sigma^2 + \beta_s'(I_m - \Lambda_s' V_x^{-1}\Lambda_s)\beta_s$, and $V_x = \Lambda\Lambda' + \Psi$.

We should choose $r = r_s + r_x$ large enough so that the shared subspace does not capture covariate-specific variation.

Canonical correlation analysis

Canonical correlation analysis (CCA) is like a symmetric unsupervised version of PLS: it allows each view to have its own "private" subspace, but there is also a shared subspace.

$$
\begin{aligned}
p(f_i) &= N(f_i^s; 0, I_{r_s}) N(f_i^x; 0, I_{r_x}) N(f_i^y; 0, I_{r_y}) \\
p(x_i|f_i) &= N(\Lambda_s f_i^s + \Lambda_x f_i^x, \Psi) \\
p(y_i|f_i) &= N(\beta_s' f_i^s + \beta_y' f_i^y, \sigma^2)
\end{aligned}
$$

so

$$(x_i, y_i) \sim N(0, WW' + \Omega)$$

where

$$W = \begin{pmatrix} \Lambda_s & \Lambda_x & 0 \\ \beta_s' & 0 & \beta_y' \end{pmatrix} \quad \text{and} \quad \Omega = \begin{pmatrix} \Psi & 0 \\ 0 & \sigma^2 \end{pmatrix}$$

Again,

$$y_i | x_i \sim N(x_i'\theta, \tau^2)$$

where $\theta = V_x^{-1}\Lambda_s\beta_s$, $\tau^2 = \sigma^2 + \beta_y'\beta_y + \beta_s'(I_m - \Lambda_s' V_x^{-1}\Lambda_s)\beta_s$, and $V_x = \Lambda_s\Lambda_s' + \Lambda_x\Lambda_x' + \Psi$.

A few more references

Gustafsson (2001). A probabilistic derivation of the partial least-squares algorithm. *Journal of Chemical Information and Modeling*, 41, 288-294.

West (2003) Bayesian Factor Regression Models in the "Large p, Small n" Paradigm. *Bayesian Statistics 7*.

Yu, Yu, Tresp and Wu (2006) Supervised probabilistic principal component analysis. In *Proc. of the Int'l Conf. on Knowledge Discovery and Data Mining.*

Sun, Ji, Yu and Ye (2009) On the equivalence between canonical correlation analysis and orthonormalized partial least squares. In *Intl. Joint Conf. on AI.*

Hahn, He and Lopes (2017) Bayesian Factor Model Shrinkage for Linear IV Regression With Many Instruments. *Journal of Business and Economic Statistics*.

Sparse factor augmented VAR[8]

Beyeler and Kaufmann (2017) Factor augmented VAR revisited - A sparse dynamic factor model approach

The framework proposed in BBE05 collects N non-trending observed variables in a $N \times 1$ vector X_t, where $t = 1, \ldots, T$. These variables are assumed to contain information on some pervasive k, $k << N$, economic factors f_t^* which are not directly observable to the econometrician but are relevant determinants of some m observed series Y_t. The FAVAR representation for $[f_t^{*\prime} Y_t^{\prime}]$ writes

$$\begin{aligned} \begin{bmatrix} X_t \\ Y_t \end{bmatrix} &= \begin{bmatrix} \lambda^{*f} & \lambda^{*Y} \\ 0 & I_m \end{bmatrix} \begin{bmatrix} f_t^* \\ Y_t \end{bmatrix} + \begin{bmatrix} \xi_t \\ 0 \end{bmatrix} \\ \Phi^*(L) \begin{bmatrix} f_t^* \\ Y_t \end{bmatrix} &= \begin{bmatrix} \eta_t^{*f} \\ \eta_t^Y \end{bmatrix} \quad \eta_t^* \sim N(0, \Sigma^*) \\ \Psi(L)\xi_t &= \varepsilon_t, \quad \varepsilon_t \sim N(0, \Omega) \end{aligned} \tag{1}$$

where λ^{*f} and λ^{*Y} are the factor loading matrices with dimension $N \times k$ and $N \times m$, respectively, and I_m represents the identity matrix of dimension m. A AR process of order p characterizes the process of $[f_t^{*\prime} Y_t^{\prime}]$. We assume that the common co-movement in X_t is fully explained by f_t^* and Y_t. Therefore, common and idiosyncratic shocks are uncorrelated, i.e. $E(\eta_t^* \varepsilon_t^{\prime}) = 0$, and idiosyncratic components ξ_t follow series-specific independent VAR processes, i.e. $\Psi(L)$ and Ω are, respectively, diagonal processes and diagonal with elements $\{\Psi(L), \Omega\} = \{\psi_i(L), \omega_i^2 | \psi_i(L) = 1 - \psi_{i1}L - \cdots - \psi_{iq}L^q, i = 1, \ldots, N\}$.

[8] Bernanke, Boivin and Eliasz (2005) Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive(FAVAR) Approach *The Quarterly Journal of Economics*, 120, 387-422.

Sparsity

The sparse factor loading matrices λ^{*f} and λ^{*Y} will be estimated freely, i.e. without imposing identification restrictions, see also section 2.3. To induce sparsity, we work with a hierarchical point mass-normal mixture prior distribution on the factor loadings λ^{*}_{ij}, $i = 1, \ldots, N$, $j = 1, \ldots, k + m$ (see e.g. West 2003, Carvalho et al. 2008)

$$\pi(\lambda^{*}_{ij}|\beta_{ij}, \tau_j) = (1 - \beta_{ij})\delta_0(\lambda^{*}_{ij}) + \beta_{ij} N(0, \tau_j) \tag{2}$$

$$\pi(\beta_{ij}|\rho_j) = (1 - \rho_j)\delta_0(\beta_{ij}) + \rho_j B(ab, a(1 - b)) \tag{3}$$

$$\pi(\rho_j) = B(r_0 s_0, r_0(1 - s_0)) \tag{4}$$

where δ_0 is a Dirac delta function that assigns all probability mass to zero and $B(uv, u(1 - v))$ denotes a beta distribution with mean v and precision u. For τ_j, we assume an inverse Gamma prior distribution $IG(g_0, G_0)$. The factor-independent

Application to the US economy

We apply our methodology to a large panel of series for the US economy to illustrate estimation and identification of the sparse FAVAR.

We find evidence for a high degree of sparsity and indeed, given the structure of estimated zero loadings, we achieve model identification.

In addition to one observed factor, i.e. the federal funds rate (FFR), we estimate seven unobserved factors.

The variance share explained by the common component amounts to 52 percent.

FRED-QD database: Federal Reserve Bank of St. Louis.

It consists of 253 macroeconomic time series for the US economy which are regularly updated and reported at a quarterly frequency starting in 1959Q1.

The FRED-QD database has been constructed along the lines of the data set used in Stock and Watson (2012) [9]

In addition, we include the utilization adjusted total factor productivity (TFP) series from Fernald (2012)[10]

Final set: 224 times series from 1965Q1 to 2015Q2.

[9] Stock and Watson (2012) Disentangling the channels of the 2007-2009 recession. NBER Working Paper Series 18094.

[10] Fernald (2012) A quarterly, utilization-adjusted series on total factor productivity. Federal Reserve Bank of San Francisco Working Paper Series.

Sparse loadings

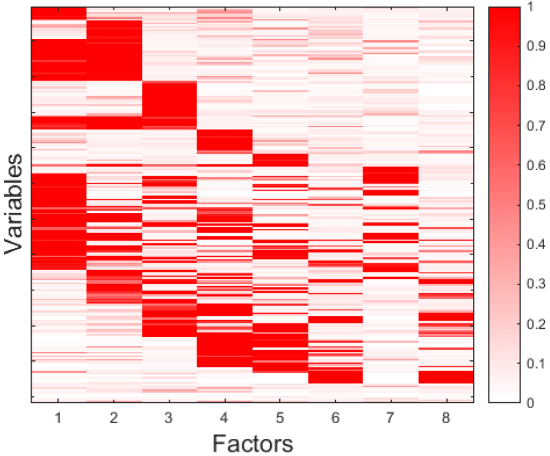


Figure 1: Posterior probabilities of non-zero factor loading.

Impulse response functions

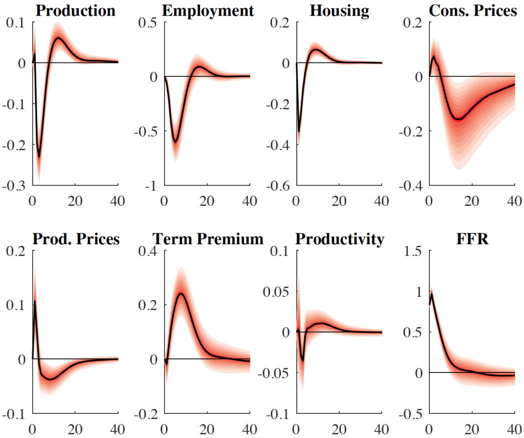


Figure 4: Impulse responses of the factors to an unanticipated change in the FFR.