EST-46114 Métodos Multivariados y Datos Categóricos

Sesion 09 - Analisis de Factores - Parte 2/3

Juan Carlos Martinez-Ovando

Maestria en Ciencia de Datos

Objetivos

- Revisaremos procedimientos para su implementacion.

Parametros

Los parametros del modelo resultan en:

- λ, matriz de cargas de dimension $(p \times k)$

–aunque por la restriccion de identificabilidad, puede reducirse el numero de parametros efectivos–

- Σ, matriz de dimension $(p \times p)$ con p parametros efectivos

Recordemos que la coleccion $(\boldsymbol{f}_i)_{i=1}^n$ son variables latentes.

Verosimilitud

La **verosimilitud extendida** para los parametros y variables latentes queda definida como

$$lik\left(\lambda, \Sigma, (\boldsymbol{f}_i)_{i=1}^{n} | \text{datos}\right) \propto \prod_{i=1}^{n} \mathrm{N}(\boldsymbol{x}_i | \lambda \boldsymbol{f}_i, \Sigma) \mathrm{N}(\boldsymbol{f}_i | \mathbf{0}, \mathbf{1}),$$

con datos $= (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$.

> *La verosimilitud extendida debe analizarse empleando metodos numericos tanto en el paradigma bayesiano como frecuentista de inferencia.*

Aprendizaje frecuentista

- El reto, en este caso, consiste en maximizar *lik* $(\lambda, \Sigma, (\boldsymbol{f}_i)_{i=1}^n | \text{datos})$ respecto a las variables latentes $(\boldsymbol{f}_i)_{i=1}^n$!!!
- La libreria `psych` provee rutinas para realizar el analisis frecuentista. La funcion base en esta libreria es `factor.pa()`.

Aprendizaje bayesiano 1/

En el caso bayesiano, se requiere dotar al modelo con distribuciones iniciales sobre los **parametros** λ y Σ, mas no sobre $(\boldsymbol{f}_i)_{i=1}^n$ pues, por ser latentes, el modelo ya incorpora una distribucion para estas.

Especificacion

Respecto a λ, podemos adoptar

$$\lambda_{jl} \sim \mathrm{N}(\lambda_{jl}|0, C_0), \text{ para } j \neq l,$$

$$\lambda_{jj} \sim \mathrm{N}(\lambda_{jj}|0, C_0)\mathbb{I}(\lambda_{jj} > 0), \text{ para } j = 1, \ldots, k.$$

Respecto a Σ, se puede adoptar

$$\sigma_j^2 \sim \mathrm{GaInv}(\sigma_j^2|v_0/2, v_0 s_0^2/2), \text{ para } j = 1, \ldots, k,$$

donde s_0^2 es la moda marginal de la varianza idiosincratica, y v_0 son grados de libertad fijos –ambos hiperparametros conocidos–.

Aprendizaje bayesiano 2/

- La distribucion final para parametros y variables latentes no es calculable de manera analitica cerrada.
- Eviten usar distribuciones iniciales vagas, en la medida de lo posible.
- La distribucion final de los parametros y variables latentes se puede aproximar empleando metodos numericos.

En `R` pueden usar confiadamente la libreria `MCMCpack`, que incluye una rutina que implementa el agoritmo *Markov chain Monte Carlo (MCMC)* para el modelo de factores, en la funcion `MCMCfactanal`.

Datos simulados 1/

Simulamos datos con la siguiente estructura

$$\lambda = \begin{pmatrix} 0.99 & 0.00 \\ 0.00 & 0.99 \\ 0.90 & 0.00 \\ 0.00 & 0.90 \\ 0.50 & 0.50 \end{pmatrix}$$

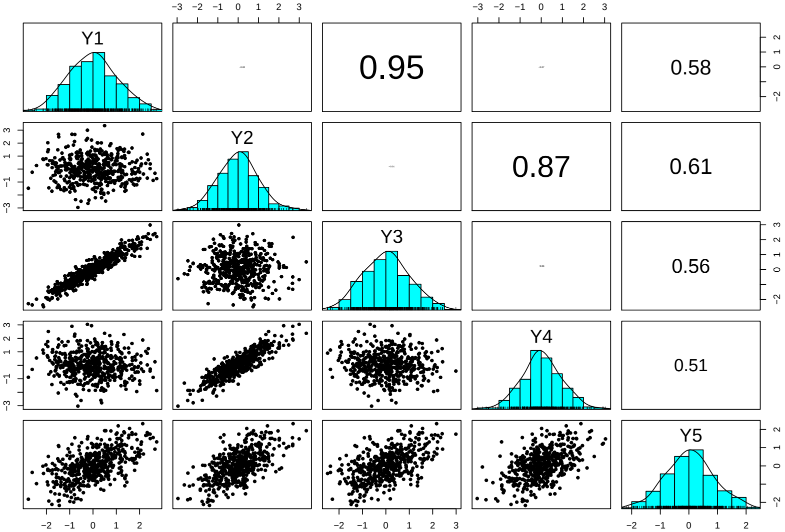
y

$$\Sigma = \mathrm{diag}\begin{pmatrix} 0.01 \\ 0.05 \\ 0.10 \\ 0.15 \\ 0.20 \end{pmatrix},$$

i.e.

$$p = 5,\ \text{y}\ k = 2.$$

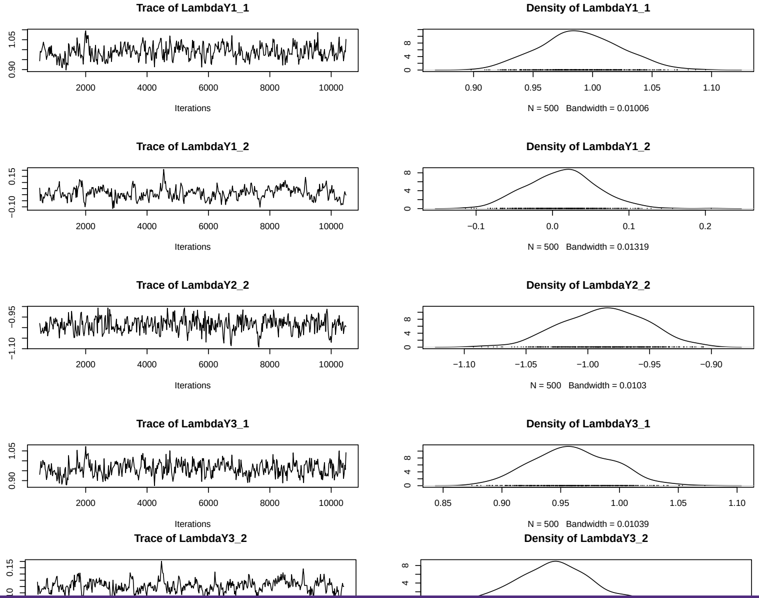
Datos simulados 2/



Datos simulados 3/

```
M.sim <- 10000
M.burn <- 500
posterior.data.sim <- MCMCfactanal(~Y1+Y2+Y3+Y4+Y5,
                          factors=k,
                          lambda.constraints=list(Y2=c(1,0)),
                          verbose=0, store.scores=FALSE,
                          a0=1, b0=0.15,
                          data=data.sim,
                          burnin=M.burn, mcmc=M.sim, thin=20)
```

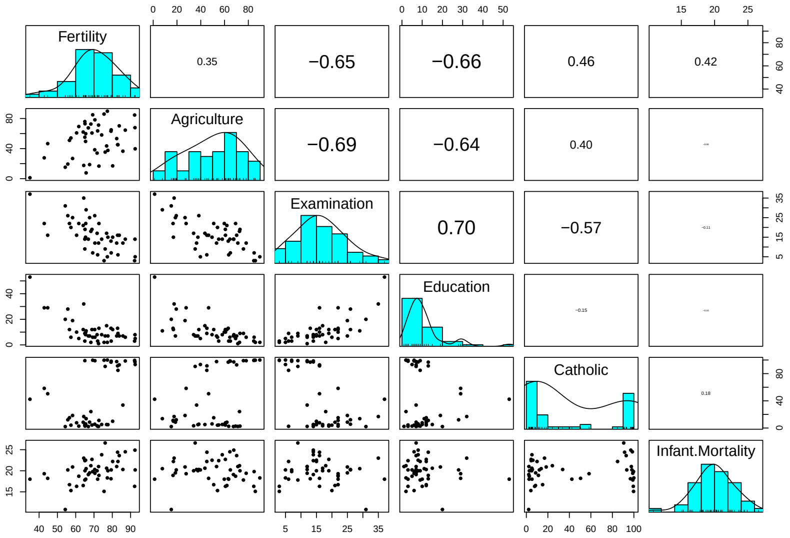
Datos simulados 3/



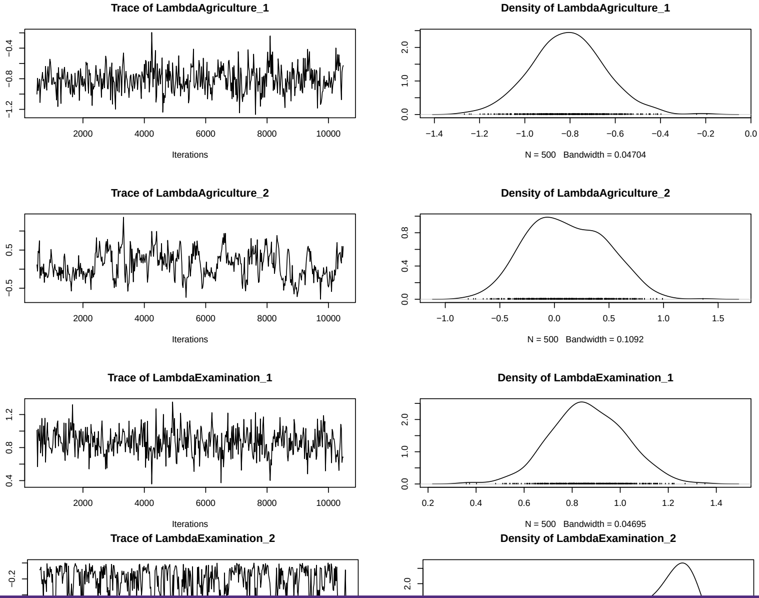
Datos simulados 4/

```
##
## Iterations = 501:10481
## Thinning interval = 20
## Number of chains = 1
## Sample size per chain = 500
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                  Mean       SD  Naive SE Time-series SE
## LambdaY1_1  0.989539 0.033698 0.0015070      0.0025961
## LambdaY1_2  0.012320 0.045911 0.0020532      0.0048523
## LambdaY2_2 -0.986894 0.033689 0.0015066      0.0021965
## LambdaY3_1  0.960562 0.033960 0.0015187      0.0026702
## LambdaY3_2  0.004417 0.045732 0.0020452      0.0047901
## LambdaY4_1 -0.055377 0.021446 0.0009591      0.0009591
## LambdaY4_2 -0.888135 0.036156 0.0016170      0.0020452
## LambdaY5_1  0.595442 0.030928 0.0013831      0.0017652
```

Ilustracion swiss 1/



Ilustracion swiss 2/



Ilustracion swiss 3/

```
##
## Iterations = 501:10481
## Thinning interval = 20
## Number of chains = 1
## Sample size per chain = 500
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                                Mean     SD Naive SE Time-series SE
## LambdaAgriculture_1        -0.80854 0.1614 0.007217       0.011013
## LambdaAgriculture_2         0.11370 0.3571 0.015969       0.045083
## LambdaExamination_1         0.86601 0.1535 0.006866       0.009017
## LambdaExamination_2        -0.23517 0.1927 0.008619       0.018360
## LambdaEducation_1           0.85468 0.1449 0.006480       0.008594
## LambdaInfant.Mortality_2    0.09013 0.6147 0.027491       0.094162
## PsiAgriculture              0.30978 0.1611 0.007206       0.011086
## PsiExamination              0.22834 0.1237 0.005531       0.008688
```

Siguiente sesion

- Estudio del analisis de factores latentes dinamicos/factores latentes en volatilidad.
- Inferencia sobre el numero de factores.
- Comparacion inferencial entre PCA y analisis de factores.
- Prediccion con factores latentes.

Lectura complementaria

- Stock & Watson (2002) *Forecasting using principal components from a large number of predictors.* `est46114_s08_suplemento1.pdf`
- West (2003) *Bayesian factor regression models in the "large p, small n" paradigm.* `est46114_s08_suplemento2.pdf`
- Pitt & Shephard (1999) *Time varying covariances: A factor stochastic volatility approach.* `est46114_s08_suplemento3.pdf`

Table of Contents