ANALISIS DE FACTORES

<u>Idea</u>: Modelar la variabilidad (y dependencias entre sus componentes) de $\mathbb{X}$ en términos de un número (usualmente más pequeño que p)[a] de variables <u>no</u> observadas (variables latentes) llamadas "factores".

ejemplo: Supóngase que $\mathbb{X}$ reporta los consumos de p diferentes bienes en un estudio sobre gastos en los hogares durante cierto mes. Estos gastos podrían ser: agua, luz, internet, teléfono, gas. La forma en que las componentes de $\mathbb{X}$ varían (así como la forma en que X_i covaría con X_j ; $i,j = 1,2,\ldots,p$, $i \neq j$), podría quedar

(a) $\mathbb{X}$ vector aleatorio de dimensión p $\mathbb{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$.
Se pretende que el número de factores sea más pequeño que p, es decir "reducir la dimensión".

explicada por dos o tres factores relacionados con la conducta social de los habitantes. Ejemplos de estos factores serían: búsqueda de comodidad, deseo de alcanzar cierta clase social, intención de cubrir necesidades básicas.

En la práctica, estos factores no observados resultan de mayor interés para quienes estudian estos fenómenos que las mismas observaciones $x_{G_1}, \ldots, x_{G_n}$. En general estos factores se interpretan como características comúnes de $x_{G_1}, \ldots, x_{G_n}$, las cuales no son observadas.

Una forma de plantear esta situación en un modelo sería si para cada x_{G_j} tenemos

$$(\mathrm{I}) \ldots \quad X_j = \sum_{l=1}^{K} a_{jl} f_l + \mu_j \ ; \ j = 1, 2, \ldots, P ;$$

f_l ; $l = 1, 2, \ldots, K$, son los factores,

se espera que K sea más pequeño (significativamente) que P

Por ejemplo, en estudios psicológicos, $\mathbb{X}$ podría representar a p resultados correspondientes a una prueba para medir "inteligencia". Un factor común a las observaciones $x_{G_1}, \ldots, x_{G_n}$ un nivel de "intelligencia" general, que se asume para los miembros de la población.

Para un vector aleatorio $\mathbb{X}$ de dimension p, con media $\mathbb{\mu}$ y matriz de covarianzas Σ, un modelo similar a (I) se puede escribir en notación matricial

$$(\mathrm{II}) \quad \ldots \quad \mathbb{X} = Q\mathbb{F} + \mathbb{\mu},$$

donde $\mathbb{F}$ es una matriz (vector de dimensión $k \times 1$), $\mathbb{Q}$ es una matriz de dimensiones $p \times k$ cuyos renglones son los pesos (en la ecuación I) asociados a los componentes de $\mathbb{X}$. Para (II), es usual asumir que: $\mathbb{E}(\mathbb{F}) = \mathbb{0}$ y $VAR(\mathbb{F}) = \mathbb{I}_k$, recordemos que los factores son variables aleatorias <u>no</u> observadas (latentes).

A continuación veremos que si se asume que los últimos $p-k$ valores propios de Σ valen 0, entonces la ecuación (II) nos sirve para expresar un modelo de factores.

Sea $\Sigma = \Gamma \Lambda \Gamma'$ la descomposición de Jordan para Σ y supongamos que $\lambda_{k+1}, \ldots, \lambda_p$ valen 0. En tal caso Σ es singular, pero podemos escribir

$$\Sigma = (\Gamma_1 \; \Gamma_2) \begin{pmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \Gamma_1' \\ \Gamma_2' \end{pmatrix}$$

Γ_1 $p \times k$

Γ_2 $p \times (p-k)$

$\Lambda_1 = \mathrm{DIAG}(\lambda_1, \ldots \lambda_k)$

$$= \sum_{l=1}^{k} \lambda_l \gamma_l \gamma_l',$$

γ_l columna l de Γ_1; $l = 1, 2, \ldots, k$.

Recordemos ahora que las componentes principales para Σ se definen como $Y = \Gamma'(X - \mu)$. Podemos escribir esta relación como

$\Gamma Y = X - \mu$ es decir $X - \mu = \Gamma Y$

$$= \Gamma_1 Y_1 + \Gamma_2 Y_2$$

donde

$$\mathbb{Y} = \begin{pmatrix} \mathbb{Y}_1 \\ \mathbb{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbb{\Gamma}_1' \\ \mathbb{\Gamma}_2' \end{pmatrix} (\mathbb{X} - \mathbb{M})$$

es un vector aleatorio con momentos

$$\mathbb{E}(\mathbb{Y}) = \mathbb{\Gamma}' \, \mathbb{E}(\mathbb{X} - \mathbb{M}) = \mathbb{O}$$

$$\begin{aligned} \mathrm{VAR}(\mathbb{Y}) &= \mathrm{VAR}(\mathbb{\Gamma}'(\mathbb{X} - \mathbb{M})) \\ &= \mathbb{\Gamma}' \, \mathrm{VAR}(\mathbb{X} - \mathbb{M}) \, \mathbb{\Gamma} = \Gamma' \Sigma \Gamma \\ &= \mathbb{\Lambda} = \begin{pmatrix} \mathbb{\Lambda}_1 & \mathbb{O} \\ \mathbb{O} & \mathbb{O} \end{pmatrix}. \end{aligned} \quad \ldots\ldots \; \underline{\text{III}}$$

Notemos que el vector $\mathbb{Y}_2$ tiene una distribucion singular con media $\mathbb{O}$ y matriz de covarianzas $\mathbb{O}$, luego

$$\mathbb{X} - \mathbb{M} = \mathbb{\Gamma}_1 \mathbb{Y}_1$$

$$\begin{aligned} \mathbb{X} &= \mathbb{\Gamma}_1 \mathbb{\Lambda}_1^{1/2} \mathbb{\Lambda}_1^{-1/2} \mathbb{Y}_1 + \mathbb{M} \\ &= \mathbb{Q}\mathbb{F} + \mathbb{M}, \end{aligned}$$

con $\underset{(p \times k)}{\mathbb{Q}} = \mathbb{\Gamma}_1 \underset{(k \times k)}{\mathbb{\Lambda}_1^{1/2}}$ y $\mathbb{F} = \underbrace{\underset{(k \times k)}{\mathbb{\Lambda}_1^{-1/2}} \underset{(k \times 1)}{\mathbb{Y}_1}}$.

notemos que

$$\Sigma = \mathbb{E}\left((\mathbb{X}-\mathbb{M})(\mathbb{X}-\mathbb{M})'\right) = \mathbb{E}\left((\mathbb{Q}\mathbb{F})(\mathbb{Q}\mathbb{F})'\right)$$

$$= \mathbb{Q}\,\mathbb{E}(\mathbb{F}\mathbb{F}')\,\mathbb{Q}' = \mathbb{Q}\,\mathbb{E}\left(\Lambda_1^{-1/2}\mathbb{Y}_1(\mathbb{Y}_1'\Lambda_1^{-1/2})\right)$$

$$= \mathbb{Q}\,\mathbb{E}\left(\Lambda_1^{-1/2}\mathbb{Y}_1\mathbb{Y}_1'\Lambda_1^{-1/2}\right)\mathbb{Q}'$$

$$= \mathbb{Q}\,\Lambda_1^{-1/2}\,\mathbb{E}(\mathbb{Y}_1\mathbb{Y}_1')\,\Lambda_1^{-1/2}\mathbb{Q}'$$

ecuación III pág 5 $\longrightarrow$

$$= \mathbb{Q}\,\Lambda_1^{-1/2}\,\Lambda_1\,\Lambda_1^{-1/2}\,\mathbb{Q}' = \mathbb{Q}\mathbb{Q}'$$

$$= \sum_{j=1}^{k} \lambda_j\, \gamma_j\, \gamma_j'$$

Entonces bajo el supuesto $\lambda_{k+1} = \lambda_{k+2} = \cdots = \lambda_p$, $\mathbb{X}$ se puede expresar como una combinación lineal de k factores ($k<p$) que son no correlacionados (las componentes de $\Lambda_1^{-1/2}\mathbb{Y}_1$).

Problema: En la práctica $\hat{\Sigma}$ no resulta singular, casi siempre.