Universidad de SanAndrés

21 de noviembre 2018

El dinero no es todo...

Componentes Principales Esparzas en la EPH

Florencia Hnilo

Motivación

- Diferenciar grupos sociales relevantes para análisis de pobreza y desigualdad
- Facilitar la interpretación
- Disminución de costos

Alternativas y desventajas

Trade-off fidelidad estadística vs practicidad

- PCA: combinaciones lineales de todas las variables originales, sobreajuste si $n < p$
- Restricción de coeficientes: Jolliffe (2002), Hausman (1982), Vines (2000), Cadima y Jolliffe (1995)

LASSO+PCA=SPCA

- Preserva la propiedad de reducción de dimensiones de PCA
- Realiza selección de variables como LASSO
- Cumple propiedades de Zou, Hastie & Tibshirani (2006):
  1. Sin restricción de penalidad, igual a PCA
  2. Eficiente tanto para n grande como para p grande
  3. No ignora variables importantes

Sea $X_1, ..., X_n \in \mathbb{R}^p$ una muestra aleatoria, primero calcular los PC y para cada α_k resolver:

$$\beta_k = \arg\min_{\beta} \sum_{j=1}^{n} (\alpha_k' X_j - \beta' X_j)^2 + \lambda_1 \sum_{i=1}^{d} |\beta[i]| + \lambda_2 \sum_{i=1}^{d} |\beta[i]|^2$$

Vector β_k normalizado $\Rightarrow$ pesos de la k-ésima componente principal esparza

- A mayor λ_1, más esparzo
- λ_2 promueve *grouping effect*.
- Desventaja: sensible a observaciones atípicas

Aplicación a la EPH

- Período: 2004-2014
- 59 variables, 2.432.617 de observaciones
- Muchos *missing values*, 2 alternativas
- Paquetes: **elasticnet**, *mixOmics* y *sparsepca*
- 2 componentes principales (más de 98% de la variabilidad explicada)

Resultados

	Alternativa 1				Alternativa 2			
	SPCA		SPCA sin codusu		SPCA		SPCA sin codusu	
Variable	PC1	PC2	PC1	PC2	PC1	PC2	PC1	PC2
codusu	0.999	-0.003	0	0	0.999	-0.034	0	0
pondera	0	0	0.0002	0	0	0	0.0002	0
itf	0.035	0.028	-0.030	0.941	0.033	0.964	-0.964	0.266
ipcf	0.013	0.015	-0.016	0.336	0.010	0.264	-0.266	-0.964
pp04d_cod	-0.003	-0.999	0.999	0.033	-	-	-	-
v8_m	0	0	0	0.003	0	0.002	-0.001	-0.009
ingreso_alquiler_monto	0	0	0	0	0	0.002	-0.001	-0.010
% varianza explicada	92.31%	7.29%	93.13%	6.37%	99.59%	0.38%	95.27%	3.15%

Fuente: Elaboración propia.

Conclusiones

- Elección de *codusu* sospechosa
- Las otras variables tienen sentido: ingreso, ocupación, alquileres...
- El dinero no es todo... ¡pero cómo ayuda!

¡Muchas gracias!