XLVIII Coloquio Argentino de Estadística

VI JORNADA DE EDUCACIÓN ESTADÍSTICA "MARTHA DE ALIAGA"

27 al 30 oct 2020

Poster:

Estimación robusta multivariada en presencia de datos faltantes

Martín Marfia, Nadia L. Kudraszow

FACULTAD DE CIENCIAS ECONÓMICAS

UNC Universidad Nacional de Córdoba

SAE

Estimación robusta multivariada en presencia de datos faltantes

Martín Marfia† & Nadia L. Kudraszow‡
†*Depto. de Cs. Básicas - Facultad de Ingeniería - UNLP* , ‡*CMaLP - UNLP - CONICET*

1. Resumen

Nuestro objetivo es proponer una generalización del estimador de tipo MM [1] para el modelo de posición y escala multivariado que sea capaz de enfrentar los dos problemas más comunes en la calidad de un conjunto de datos: los datos atípicos y la presencia de datos faltantes. Para ello, nuestro enfoque es considerar como estimador de escala inicial la escala de las distancias de Mahalanobis parciales del S-estimador generalizado [2], y usarlo como punto de partida para calcular un M-estimador cuya función rho tiene un parámetro para controlar la eficiencia.

Palabras clave: datos faltantes, estimación robusta.

2. Introduccíon

Para $1 \leq i \leq n$, sean $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$, vectores aleatorios de dimensión p (iid) y $\mathbf{u}_i = (u_{i1}, \ldots, u_{ip})^\top$, vectores (iid) e independientes de los $\mathbf{x}_i$ compuestos por unos y ceros: x_{ij} fue observada cuando $u_{ij} = 1$. Sea $\mathbf{x}^{(\mathbf{u})}$ la parte de $\mathbf{x}$ que fué observada y sea $p(\mathbf{u}) = \sum_{j=1}^{p} u_j$.

Llamamos $\mathbf{\Sigma^{(u)}}$ a la submatriz de $\mathbf{\Sigma}$ correspondiente a las entradas no nulas de $\mathbf{u}$ y $\mathbf{m}^{(\mathbf{u})} \in \mathbb{R}^{p(u)}$ al correspondiente subvector de $\mathbf{m} \in \mathbb{R}^p$. Además, sea $\mathbf{\Sigma}^{*(\mathbf{u})} = \mathbf{\Sigma}^{(\mathbf{u})}/|\mathbf{\Sigma}^{(\mathbf{u})}|^{1/p(\mathbf{u})}$ y $d(\mathbf{x}, \mathbf{m}, \mathbf{\Sigma}) = (\mathbf{x} - \mathbf{m})^\top \mathbf{\Sigma}^{-1}(\mathbf{x} - \mathbf{m})$ la distancia de Mahalanobis al cuadrado entre $\mathbf{x}$ y $\mathbf{m}$. Finalmente, $\rho_1 = \rho(t/c)$ donde ρ es una $\rho-$función y c es una constante elegida para controlar la eficiencia.

3. Estimador MM

Proponemos usar como estimador inicial los estimadores de posición y escala *GSE*, $\left(\hat{\mathbf{m}}_{\mathbf{GS}}, \hat{\mathbf{\Sigma}}_{\mathbf{GS}}\right)$, de [2] y luego consideramos la solución de

$$\left(\hat{\mathbf{m}}_{\mathrm{R}}, \hat{\mathbf{\Gamma}}_{\mathrm{R}}\right) = \arg \min_{\mathbf{m}, |\mathbf{\Sigma}|=1} T_{\mathrm{R}}\left(\mathbf{m}, \mathbf{\Sigma}\right),$$

$$\hat{\mathbf{\Sigma}}_{\mathrm{R}} = \hat{\sigma}_{\mathbf{GS}} \hat{\mathbf{\Gamma}}_{\mathrm{R}},$$

donde $\hat{\sigma}_{\mathbf{GS}} = \left|\hat{\mathbf{\Sigma}}_{\mathbf{GS}}\right|^{1/p}$ y $T_R\left(\mathbf{m}, \mathbf{\Sigma}\right) =$

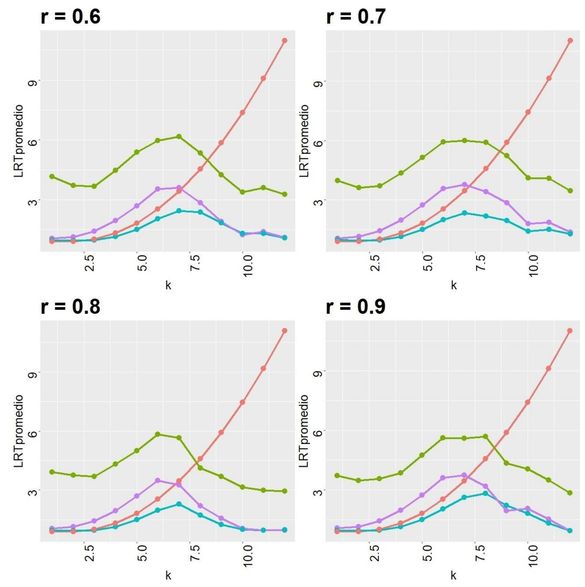
$$\frac{1}{\sum_{j=1}^{n} c_{p(\mathbf{u}_i)}} \sum_{i=1}^{n} c_{p(\mathbf{u}_i)} \rho_1 \left(\frac{d\left(\mathbf{x}_i^{(\mathbf{u}_i)}, \mathbf{m}^{(\mathbf{u}_i)}, \mathbf{\Sigma}^{*(\mathbf{u}_i)}\right)}{\hat{\sigma}_{\mathbf{u}_i} c_{p(\mathbf{u}_i)}} \right)$$

donde $\hat{\sigma}_{\mathbf{u}_i} = \left|\hat{\mathbf{\Sigma}}_{\mathbf{GS}}^{(\mathbf{u}_i)}\right|^{1/p(\mathbf{u}_i)}$ para $i = 1, \ldots, n$.

4. Estudio de simulación

Generamos muestras de tamaño $n = 100$ de una $N_p(\mathbf{0}, \mathbf{\Sigma})$. Como el estimador es escala equivariante pero no afín equivariante asumimos $\mathbf{\Sigma}_{ii} = 1$ y $\mathbf{\Sigma}_{ij} = r$ para $i \neq j$ y tomamos $r = 0.6, 0.7, 0.8, 0.9$. Introdujimos 10% de contaminación puntual a k distancias de Mahalanobis del $\mathbf{0}$ $(k = 1, \ldots, 12)$ en la dirección del autovector de $\mathbf{\Sigma}$ con autovalor más chico. Se ha observado empíricamente que esta es la posición menos favorable para ubicar datos atípicos . El porcentaje de datos faltantes se fijó en 10% (otras proporciones dieron patrones similares). El número de réplicas fué $N = 1000$. La performance de cada $\hat{\Sigma}_n$ se midió usando el promedio sobre las réplicas del $LRT(\Sigma, \Sigma_0) = tr(\Sigma\Sigma_0^{-1}) - \log\det(\Sigma\Sigma_0^{-1}) - p$. Se calcularon además las eficiencias relativas muestrales con respecto al estimador EM basado en el promedio de las distancias LRT cuando hay 10% de datos faltantes y no hay datos atípicos.

5. Resultados de la simulación



A la izquierda se muestran las distancias LRT promedio en función del k, de los siguientes estimadores :

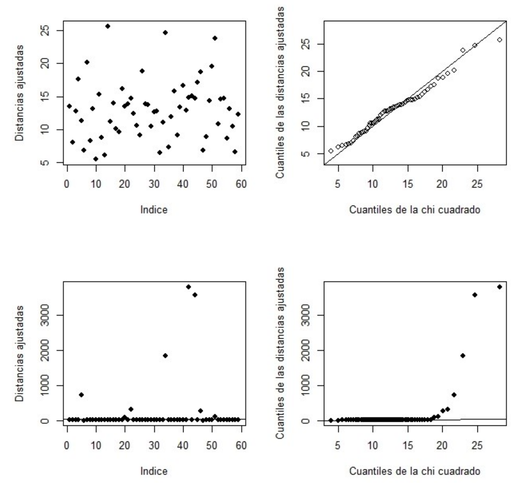
1. EM, el estimador gaussiano.
2. EMVE, el S estimador extendido para datos faltantes.
3. GSE, el S estimador generalizado con EMVE como estimador inicial.
4. GMM, nuestro estimador propuesto, usando una ρ función bicuadrada.

Notamos un buen desempeño de nuestra propuesta tanto en los distintos escenarios bajo contaminación como en la eficiencia relativa al EM sin contaminación (ver Tabla 1).

Tabla 1:

r	GSE	GMME	EMVE	EM
0.6	0,88	0,90	0,23	1
0.7	0,87	0,91	0,23	1
0.8	0,87	0,92	0,24	1
0.9	0,87	0,91	0,25	1

6. Ejemplo con datos reales



Se eliminaron al azar, con probabilidad 0.2, entradas del data set *wine* (del paquete RobStatTM de R) que contiene para 59 vinos cultivados en la misma región de Italia, las cantidades de 13 componentes. En el gráfico se comparan las distancias de Mahalanobis ajustadas $d^* = (\chi^2_{13})^{-1}(\chi^2_q(d))$ donde d es la distancia parcial de una fila con q entradas observadas de 13, obtenidas con el estimador gaussiano EM (arriba) y nuestro estimador propuesto (abajo), y χ^2_p es la función de distribución χ^2 con p grados de libertad. Observamos que el estimador gaussiano no detecta datos atípicos, mientras que nuestra propuesta detecta 8. Se consideraron como datos atípicos las observaciones con distancia ajustada mayor a $(\chi^2_{13})^{-1}(0.999)$.

Referencias

[1] D. E. Tatsuoka, K. S. y Tyler. The uniqueness of s and m-functionals under non-elliptical distributions. *The Annals of Statistics, 28, 1219-1243.*, 2000.

[2] Yohai V. y Zamar R. Danilov M. Robust estimation of multivariate location and scatter in the presence of misssing data. *Journal of the American Statistical Association,107:499, 1178-1186.*, 2014.

7. Busquedas a futuro

Queremos estudiar las propiedades asintóticas (consistencia y normalidad asintótica) del estimador propuesto.

Este método puede utilizarse para el caso en que la contaminación no es por observaciones sino por celda (cell-wise) eliminando las celdas contaminadas identificadas mediante un método de detección y luego aplicando el MM-estimador para datos faltantes.

Contacto

Mail: martin.marfia@ing.unlp.edu.ar