Bussi, Javier; Ciccioli, Patricia

Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística

UNA REVISIÓN DE LOS DISTINTOS MÉTODOS ROBUSTOS PARA EL ANALSIS DE COMPONENTES PRINCIPALES

Resumen

En este trabajo se presenta una revisión de algunos de los métodos robustos más difundidos desarrollados hasta la actualidad para el análisis de componentes principales (ACP). Se analizan sus características, sus coincidencias y diferencias. Se presentan además las funciones que se encuentran programadas en el paquete estadístico R de algunos de estos métodos.

Abstract

This work presents a revision of some of the most renowned robust methods for Principal Component Analysis (PCA) developed up to date. Their characteristics are analyzed, together with their similarities and differences. The functions programmed in the R statistical package of some of these methods are also presented.

Palabras claves: ANÁLISIS DE COMPONENTES PRINCIPALES, MÉTODOS ROBUSTOS, PAQUETE ESTADÍSTICO R

INTRODUCCIÓN

El análisis de componentes principales (ACP) es una técnica muy utilizada dentro de los métodos estadísticos multivariados. El objetivo de este método es representar adecuadamente un conjunto de n observacionescon p variables a través de un número menor de variables construidas como combinaciones lineales de las originales. La técnica se basa en el cálculo de autovalores y autovectores de la matriz de covariancias o de correlaciones de las variables originales. La presencia de valores atípicos en los datos puede distorsionar la matriz de covariancias muestrales. Por este motivo se han propuesto diversas formas de tratar esta dificultad a partir de técnicas robustas. En este trabajo se consideran algunas de las técnicas más difundidas desarrolladas hasta la actualidad. Se presenta una primera gran división entre estas técnicas robustas, en primer lugar las que consideran la estimación robusta de la matriz de variancias y covariancias y en segundo lugar las técnicas que trabajan con la estimación de la componente principal directamente, obtenida de una manera robusta. En la sección 2 se presenta brevemente el ACP clásico. En la sección 3 se presentan algunos métodos basados en la estimación robusta de la matriz de variancias y covariancias. En la sección 4 se presentan algunos métodos que obtienen directamente la estimación robusta de la componente principal. En la sección 5 se mencionan algunos métodos presentados que se han desarrollado a través de funciones en el paquete estadístico R y en la sección 6 se presentan los comentarios finales.

2. ACP CLÁSICO

El análisis de componentes principales (ACP) tiene por objetivo representar apropiadamente la información provista por un grupo de n observaciones donde se consideran p variables, reduciendo el número de estas pero resignando una baja cantidad de información. Esta representación se realiza a través de la creación de r nuevas variables no observables que resultan ser combinaciones lineales de las p variables originales (r<p).

Por ejemplo, en casos donde se presentan variables con alta asociación, se puede reducir el número de las mismas, seleccionando sólo algunas de ellas pero que expliquen un alto porcentaje de la variabilidad total original. Es posible representar las observaciones en un espacio de menor dimensión (r) pudiendo identificar variables latentes que expliquen la variabilidad de los datos y al mismo tiempo generando variables no correlacionadas que facilitan la interpretación de los resultados obtenidos.

Suponiendo que se cuenta con n elementos de una población en los que se miden p-variables, los cuales son las observaciones de un vector aleatorio $\boldsymbol{x}$ p-dimensional con media $\boldsymbol{\mu}$ con una matriz de covariancias $\boldsymbol{\Sigma}$ de dimensión $p \times p$. La primera componente principal es la combinación lineal de las variables originales que resulta de la proyección que tiene mayor variabilidad de manera tal que:

$$Var(\boldsymbol{b}_1'\boldsymbol{x}) = max \quad \text{dado que} \quad //\boldsymbol{b}_1// = 1 \tag{1}$$

siendo $\boldsymbol{x}'\boldsymbol{b}_1$ la forma que toma la combinación lineal. La segunda componente principal cumple (1) y además que $\boldsymbol{b}_2'\boldsymbol{b}_1 = 0$. De manera similar se computan las componentes principales subsiguientes. El número total de componentes es *p*, los autovalores de la matriz $\boldsymbol{\Sigma}$ son $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ y sus respectivos autovectores son $\boldsymbol{b}_1, \boldsymbol{b}_2, ..., \boldsymbol{b}_p$. La variancia de cada componente principal está dada por:

$$Var(\boldsymbol{b}_j'\boldsymbol{x}) = \lambda_j$$

El número q de componentes puede ser seleccionado a través del criterio de porcentaje de variancia explicada:

$$\frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

El ACP ha sido extensamente tratado en la bibliografía estadística, una detallada explicación del mismo puede ser consultada en Peña (2002) y Johnson y Wichern (2007).

3. MÉTODOS BASADOS EN LA ESTIMACIÓN ROBUSTA DE LA MATRIZ DE VARIANCIAS Y COVARIANCIAS

Se presentan los métodos de Stahel y Donoho, los M-estimadores, los S-estimadores y los métodos MCD y MVE.

3.1. Método de Stahel y Donoho

El método que resulta ser más razonable y directo es reemplazar la matriz de variancias y covariancias por alguna estimación robusta. En este método, simplemente se reemplaza la matriz utilizando la estimación de Stahel y Donoho. Esta estimación consiste en un cálculo ponderado de la matriz que queda definido como:

$$\hat{\Sigma} = \frac{1}{\sum_{i=1}^{n} W_{i2}} \sum_{i=1}^{n} W_{i2} \ (X_i - \hat{\mu})(X_i - \hat{\mu})'$$

Se define:

$$t(x,a) = \frac{x'a - \hat{\mu}(a'x)}{\hat{\sigma}(a'x)}$$

donde $a \in \mathbb{R}^p$ con $\|a\| = 1$, es una dirección. Se denomina $a'X = \{a'X_1, \ldots, a'X_n\}$ la proyección del conjunto de datos X a lo largo de a. El grado de extremidad (*outlyingness*) de x queda definido por: $t(x) = \max_a t(x,a)$, siendo t invariante y siendo entonces los estimadores equivariantes. Las funciones de ponderación deberán cumplir las siguientes condiciones: $tW_1(t)$ y $t^2W_2(t)$ son acotadas para $t \geq 0$ .Se puede demostrar que bajo esta condición el punto de ruptura es BP=1/2. Se define la función de pesos Huber.

Según Maronna, (2006), resulta $\hat{\mu} = \frac{1}{\sum_{i=1}^n W_{i1}} \sum_{i=1}^n W_{i1}\ x_i$

La estimación alcanza el máximo punto de quiebre si $\hat{\mu}$ es la mediana de la muestra y la escala es: $\hat{\sigma}(z) = \frac{1}{2}(\tilde{z}_k + \tilde{z}_{k+1})$.donde $\tilde{z}_i$ denota los valores ordenados de $|\tilde{z}_i - \mathrm{Med(z)}|$ y k=[(n+p)/2]. La elección de la función de peso es importante para combinar robustez y eficiencia. Además de la función de peso Huber también define las funciones de peso: Maronna y Yohai (1995), Maronna y Zammar (2002) y Zuo et al. (2004).

Según Filtmozer-Todorov (2009), el peso Wi de cada observación es inversamente proporcional para el “outlier” de la observación. Ambos sostienen que el cálculo exacto del estimador no es posible y estos últimos sostienen que se puede encontrar una solución aproximada por submuestreo de un gran número de direcciones y el cálculo de las medidas de extremidad.

3.2. M-estimadores

Este método se basa en el principio de Máxima verosimilitud para obtener los M-estimadores del vector de medias y de la matriz de covariancias, en el supuesto que la función de distribución de $\vec{X}$ tiene una estructura elipsoidal.

Los M estimadores se definen,en general como solución de:

$$\sum_{i=1}^n W_1(d_i)\ (x_i - \mu) = 0$$

$$\frac{1}{n}\sum_{i=1}^{n} W_2(d_i)\,(x_i-\mu)(x_i-\mu)' = \hat{\hat{\Sigma}}$$

donde las funciones W_1 y W_2 no son necesariamente iguales.

Los M-estimadores pueden ser vistos como vector de medias ponderados y como matriz de covariancias ponderadas, en la que el peso de cada observación dependerá de alguna distancia para el estimador de posición.

Se define: $\vec{\mu} = \frac{\sum_{i=1}^{n} W_1(d_i)\vec{x_i}}{\sum_{i=1}^{n} W_1(d_i)}$

Para obtener unicidad en las soluciones de las ecuaciones anteriores, es necesario requerir que $dW_2(d)$ sea una función no decreciente.

Se considera una M-estimación de localización y dispersión monótona si $dW_2(d\ )$ es no

decreciente, y redescendiente en caso contrario.

3.3 S-estimadores

Con el fin de disminuir los residuos, es necesario definir una estimación de la matriz de covariancias y por ende obtener un criterio para hacer pequeñas las distancias d_i desde las observaciones a las estimación robusta de la media en base a una medida robusta de la variación.

Los S-estimadores propuestos por Davies (1987), sugieren un M-estimador de escala $\hat{\sigma}$ que satisface:

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{d_i}{\hat{\sigma}}\right) = \delta$$

donde ρ es una función suave acotada . Bajos ciertas condiciones el S-estimador puede considerarse un M-Estimador. Los S- estimadores tienen una gran conexión con los M-estimadores y la solución que se obtiene es también una solución para una ecuación definiendo un M-estimador como una media ponderada y una matriz de covariancias ponderada (Filtmozer-Todorov , 2009).

3.4 Estimador MCD

Otra posibilidad de obtener un estimador robusto es computar un estimador truncado $\hat{\sigma}$ en vez de un M-estimador. Se consideran solo las primeras d_i más pequeñas. La estimaciones obtenidas en este caso reciben el nombre de Mínimo Determinante de la Covariancia (MCD en inglés). El estimador MCD está definido por el siguiente subconjunto $\{x_{i1},..x_{ih}\}$ de h

observaciones cuya matriz de covariancias, tenga el menor determinante a través de todos los posibles subconjuntos de tamaño h. Existe un algoritmo muy rápido debido a Rousseeuw y Van Driessen. Se ha demostrado que el proceso de iteración propuesta, converge en un número finito de pasos a un mínimo (local). El estimador MCD de localización y escala resultan ser la media aritmética y un múltiplo de la matriz de covariancias de la muestra de ese subconjunto h de observaciones. Una recomendable elección para h es [(n+p+1)/2] porque logra el punto de ruptura máximo.

3.5 Estimador MVE

Esta estimación busca el elipsoide de volumen mínimo (Minimun Volume Ellipsoid, MVE en sus siglas en inglés) que contiene al menos la mitad de los puntos del conjunto de datos. Su nombre se debe a que trata de hallar entre todas las elipsoides que contengan al menos la mitad de los puntos, aquella que tenga menor volumen, es decir, un determinante menor de la matriz de covariancias. La tasa de consistencia del MVE es baja y la estimación de localización se da como el centro de este elipsoide y la estimación de covariancia es proporcionada por su forma. El MVE fue el estimador más popular por su alto punto de ruptura, pero más tarde fue sustituido por el MCD, principalmente debido a la disponibilidad de un algoritmo eficiente para su cálculo.

4. MÉTODOS BASADOS EN LA ESTIMACIÓN DE CADA COMPONENTE PRINCIPAL DE MANERA ROBUSTA

En esta sección se presentan los métodos que computan directamente una estimación robusta de los autovectores y autovalores. A continuación se detallan el método de Componentes Principales Esféricas, los estimadores de Proyección y el método de Hubert.

4.1 Componente Principales Esféricas

En el caso de las Componentes Principales Esféricas (Spherical Principal Components, SPC según sus siglas en inglés), si las observaciones tienen una distribución elíptica (normal multivariada) y Var(x) existe, resulta ser un múltiplo de Σ .Si y es la normalización de x a la superficie de la esfera unidad, centrado en μ , Boente y Fraiman (1999) mostraron que los autovectores $t_1,\ldots,t_p$ (pero no los autovalores) de la matriz de covariancia de y coinciden con los de Σ . Se demostró además que si $\sigma(.)$ es alguna estadística de dispersión entonces los valores $\sigma(x^{'}t_j)^2$ son proporcionales a los autovalores de Σ . Este es el principio básico subyacente para las SPC.

Se calculan los valores de las observaciones y en base a algún estimador robusto de localización, se computan los autovectores de su matriz de covariancias estimada. Luego, a través de algún estimador robusto de dispersión se computan los autovalores estimados y las primeras q componentes principales y sus direcciones están dadas por los q primeros autovectores computados.

4.2 Estimadores de Proyección

Si $\hat{\Sigma}$ es la estimación de la matriz de covariancias de x y $\hat{\sigma}(.)$ es el desvío estandar resulta:

$$\hat{\sigma}(a'x)^2 = a'\hat{\Sigma}a \quad \forall \; a \in \mathbf{R}^p \quad (*)$$

Sería deseable tener un estimador robusto $\hat{\Sigma}$ que verifique esta relación cuando $\hat{\sigma}$ es una medida robusta de dispersión como el MAD. Se demuestra que SD es la única medida de dispersión que satisface (*) y por lo tanto el objetivo es inalcanzable. Para superar esta dificultad, fue propuesto el P-estimador de dispersión como la mejor aproximación para (*). Específicamente un P-estimador de dispersión es una matriz $\hat{\Sigma}$ que satisface:

$$\max_{a \neq 0} \left| \log\left(\frac{\hat{\sigma}(a'x)^2}{a'\hat{\Sigma}a}\right) \right| = min$$

Una idea similar para el estimador de localización fue propuesto por Tyler (1994). Los estimadores de proyección fueron introducidos inicialmente por Li y Chen (1985) utilizando un M-estimador de escala S_n como un Índice de Proyección (IP).

4.2 Método de Hubert

El método propuesto por Hubert et al (2005) trata de combinar las ventajas de ambos enfoques, el ACP basado en una matriz de covariancia robusta y ACP basado en los estimadores de proyección. El algoritmo utilizado (ROBPCA) encuentra un centro robusto m y una matriz de cargas P. Sus columnas, que son ortogonales, definen un nuevo sistema de coordenadas. Los puntajes definidos en una matriz T son las coordenadas de las observaciones centradas con respecto a las cargas. El algoritmo conduce a una estimación robusta de la matriz de covariancias, generalmente singular, a través de tres pasos principales. Los pasos del algortimo se presentan en Filtmozer y Todorov (2009) y una descripción más detallada del mismo se desarrolla en Hubert et al (2008).

5. MÉTODOS PRESENTES EN EL PAQUETE ESTADÍSTICO R

Para el enfoque que computa inicialmente una estimación robusta de la matriz de covariancias, el paquete **stats** en el R base contiene la función princomp() que realiza el ACP sobre una matriz de datos numérica. Esta función tiene un parámetro covmat que puede tomar una matriz de covariancias o una lista de covariancias como la que provee cov.wt. En el caso que se suministre esta lista, la misma es utilizada en lugar de la matriz de covariancias y permite obtener un ACP robusto suministrando la matriz de covariancias provista por cov.mve (correspondiente al estimador MVE) o cov.mcd (correspondiente al

estimador MCD) del paquete **MASS**. Si bien estas funciones están presentes en el paquete base, la función PcaCov calcula Componentes Principales Robustas reemplazando la matriz de covariancias por alguna de sus estimaciones robustas (Stahel y Donoho, M-estimador, S-estimador,MCD, MVE), siendo el parámetro cov.control cualquier objeto derivado de la clase base CovControl. Si el parámetro es omitido, por defecto se utiliza el método MCD. Cualquier nuevo estimador que se adapte a los conceptos de este marco de programación puede ser utilizado como parámetro de entrada en la función PcaCov(). El aporte de esta función radica en la unificación de interfaces que se nivelan y además porque permite implementar un conjunto de nuevos procedimientos robustos multivariados.

En el caso donde se estiman los autovectores y autovalores de la matriz de covariancias en forma robusta, existe el paquete pcaPP que presenta dos algoritmos para el método de proyección expresados en las funciones PCAproj() y PCAgrid(). Una gran ventaja de los métodos de proyección es que calculan los autovalores en forma consecutiva y en los casos de datos con muchas variables, si el interés está centrado en las primeras dos o tres componentes principales, esto reduce considerablemente los tiempos computacionales. Para el caso del método SPC la función correspondiente es PcaLoncatore() y en el caso del método de Hubert la función resulta ser PcaHubert().

COMENTARIOS FINALES

Este trabajo presenta una revisión de los distintos métodos robustos desarrollados hasta la actualidad para el análisis de componentes principales (ACP) presente en la bibliografía. Esta revisión no es completa ni extremadamente detallada pero permite reunir en un solo texto de manera ordenada y simple los métodos más difundidos que han sido desarrollados hasta la actualidad. Se analizan algunas de sus características y coincidencias presentándose además las diferencias presentes entre los distintos métodos y en algunos casos, distintas versiones, según el autor, de un mismo método. Se presentan además las adaptaciones de algunos métodos que se encuentran programados en el paquete estadístico R hasta la fecha, lo que permite al lector encontrar en forma directa y concisa la función en R que corresponde al método robusto que se desee utilizar. Esta contribución no es menor, dado la gran variedad de información existente respecto al ACP robusto y la ausencia de un único texto que reúna esta información.

REFERENCIAS BIBLIOGRÁFICAS

Bonifazi, F. (2015). *Estimación robusta de la función de autocorrelación*. Tesina de grado de la Licenciatura en Estadística, Facultad de Ciencias Económicas y Estadística, UNR.

Bonifazi, F.; Méndez, F. (2014). Estimación robusta de la función de autocorrelación. *Decimoctavas Jornadas "Investigaciones en la Facultad" de Ciencias Económicas y Estadística.*

Box, G. E. P.; Jenkins, D. A. (1976). *Time Series Analysis and Control*, 2nd edition. Holden-Day.

Box, G. E. P.; Pierce, D. A. (1970). Distribution of residual autocorrelations in autorregresive-integrated moving average time series models. *J. American Statistica*

Association, 65, 1509-1526.

Buishand, T.A.; Beersma, J.J. (1993). Jackknife Tests for Differences in Autocorrelation between Climate Time Series. *Journal of Climate,*1, 6, 2490-2495.

Chambers, M. J. (2013), Jackknife Estimation of Stationary Autoregressive Models. *Journal of Econometrics*, 171, Issue 1, January, 142–157

Chan, W.; Wei, W. (1992), A comparison of some estimators of time series autocorrelations. *Computational Statistics & Data Analysis*, 14, 146-163.

Dürre, A.; Fried, R.; Liboschik, T. (2014). *Robust estimation of (partial) autocorrelation*. Discussion Paper.

Efron, B., Tibshirani, R. (1993)**.** *An Introduction to the Bootstrap***.** New York: Chapman & Hall.

Efron, B. (1979). Bootstrap Methods: another look at the jackknife. *Annals of Statistics*, 7, 1-26.

Ma, Y.; Genton, M. (2000), Highly robust estimation of the autocovariance function. *Journal of Time Series Analysis*, 21, Nº6, 663-684.

Maronna, R.A.; Martin, R.D.; Yohai, V.J. (2006), *Robust Statistics: Theory and Methods*. John Wiley and Sons.

Quenouille, M.H. (1949). Approximate tests of correlation in time series. Mathematical

Proceedings of the Cambridge Philosophical Society, 45, 483-484

Salibian-Barrera, M. (2000). *Contributions to the theory of robust inference*. Ph.D. thesis, Dept. Statist., Univ. British Columbia,Vancouver.

Salibian-Barrera, M., Van Aelst S., Willems G. (2006). PCA Based on Multivariate MM-Estimators with Fast and Robust Bootstrap. *Journal of the American Statistical Association,* 101, 1198-1211.

Salibian-Barrera, M., Zamar, R. H. (2002). Bootstrapping robust estimates of regression.*The Annals of Statistics*, 30, 556-582.

Van Aelst, S., Willems, G. (2005).Multivariate regression S-estimators for robust estimation and inference.*StatisticaSinica*, 15, 981-1001.

Van Aelst, S., Willems, G. (2013). Fast and robust bootstrap for multivariate inference: The R package FRB. *Journal of Statistical Software*, 53 (3),1–32.URL: http://www.jstatsoft. org/v53/i03/.