COMPARACIÓN DE MÉTODOS ROBUSTOS PARA EL ANÁLISIS CANÓNICO ASIMÉTRICO

María Victoria Fasano[†] y Nadia Laura Kudraszow[†, ‡]

[†]*Departamento de Matemáticas, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, Argentina,* vicky@mate.unlp.edu.ar

[‡]*IMAS, CONICET, Ciudad Universitaria, Pabellón 1, 1428, Buenos Aires, Argentina,* nkudraszow@mate.unlp.edu.ar

Resumen: El análisis de canónico asimétrico o redundancia busca para dos grupos de variables las combinaciones lineales en un grupo que maximicen la varianza explicada del otro por dicha combinación lineal. En este trabajo se propone un método robusto para el análisis de redundancia basado en estimadores para regresión lineal multivariada. Se mostrará el buen desempeño de los métodos propuestos comparado con el método clásico y otros métodos basados en matrices de correlación robustas, mediante un estudio de simulación utilizando muestras con y sin contaminación.

Palabras clave: *Análisis de Redundancia, Regresión Lineal Multivariada, Métodos Robustos*
2000 AMS Subject Classification: 62H20 - 62F35

1. INTRODUCCIÓN

El problema de encontrar relaciones entre grupos de variables es central en el análisis multivariado. Una gran cantidad de métodos fueron sugeridos para lograr este objetivo pero el análisis de correlación canónica es el más utilizado. Clásicamente el análisis canónico se realiza obteniendo las combinaciones lineales de cada grupo de variables que maximizan su correlación restringido a que las varianzas de dichas combinaciones sean iguales a uno. El análisis de correlación canónica es simétrico en las variables: si las intercambiamos, el número de variables canónicas no se modifica, las correlaciones entre las variables canónicas son idénticas y los vectores que definen las variables canónicas se intercambian. Existen situaciones donde esta simetría no es deseable. Puede ocurrir que un grupo, $\mathbf{x}$, sea de variables exógenas que queremos utilizar para prever a las endógenas, $\mathbf{y}$, y queremos un procedimiento que tenga en cuenta esta asimetría, es decir que maximice la explicación de las variables $\mathbf{y}$. El análisis de correlaciones canónicas no resuelve el problema. Podemos tener una alta correlación entre las variables canónicas de $\mathbf{x}$ e $\mathbf{y}$ y una baja correlación entre cada variable del conjunto $\mathbf{y}$ y la variable canónica asociada a $\mathbf{x}$. Para resolverlo Steward y Love [11] propusieron el coeficiente de redundancia. Este coeficiente es una medida de la capacidad predictiva de un grupo de variables respecto al otro. Wollemberg [13] desarrolló el análisis canónico asimétrico o de redundancia como el procedimiento en el que se obtienen combinaciones lineales incorrelacionadas de las variables predictoras, con varianza uno, tales que maximizan el coeficiente de redundancia entre una de las combinaciones lineales antes mencionadas y el otro grupo. El análisis de redundancia (AR) ha sido aplicado en distintas áreas, como ser genética [14], medicina [2], ecología [5] y psicología [1]. Para estimar el coeficiente de redundancia y las combinaciones lineales que lo maximizan se utilizan versiones muestrales de las matrices de correlaciones, las cuales son altamente sensibles a observaciones atípicas, lo que lleva a la necesidad de desarrollar una alternativa robusta. En primer lugar formalizaremos la definición de coeficiente de redundancia y el método de estimación clásico para el análisis de redundancia, esto ayudará a abordar el problema de la estimación robusta en el AR y en particular para introducir el método robusto propuesto basado en regresión robusta multivariada. Finalmente se desarrolló un estudio de simulación con el fin de comparar el desempeño de los estimadores propuestos con otros basados en matrices robustas de covarianza y regresión alternada.

2. COEFICIENTE Y ANÁLISIS DE REDUNDANCIA

Consideremos $\mathbf{y} = (y_1, ..., y_q)'$ el grupo de variables respuesta y $\mathbf{x} = (x_1, ..., x_p)'$ el grupo de variables explicativas. Suponemos, sin pérdida de generalidad, que $\mathbf{y}$ y $\mathbf{x}$ están ambas estandarizadas (tienen media cero y varianza unitaria) y que la matriz de correlaciones conjunta puede escribirse

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_{yy} & \mathbf{R}_{yx} \\ \mathbf{R}_{xy} & \mathbf{R}_{xx} \end{pmatrix},$$

donde $\mathbf{R}_{yy}$ y $\mathbf{R}_{xx}$ son matrices no singulares.

El coeficiente de redundancia de $\mathbf{y}$ dada la variable u está definido como $R(\mathbf{y}\,|u) = \sum_{i=1}^{q} \text{corr}^2(y_i, u)/q$. De manera análoga, se define el coeficiente de redundancia de $\mathbf{x}$ dada la variable v como $R(\mathbf{x}\,|v) = \sum_{i=1}^{p} \text{corr}^2(x_i, v)/p$.

El AR busca las combinaciones lineales $u_1 = \boldsymbol{\alpha}_1'\mathbf{x}$ (primera variable de redundancia), que maximizan el coeficiente de redundancia $R(\mathbf{y}\,|\boldsymbol{\alpha}'\mathbf{x})$, bajo la restricción $\text{var}(\boldsymbol{\alpha}'\mathbf{x}) = 1$. La segunda variable de redundancia, $u_2 = \boldsymbol{\alpha}_2'\mathbf{x}$, se define como la combinación lineal incorrelacionada con u_1, que maximiza el coeficiente de redundancia $R(\mathbf{y}\,|\boldsymbol{\alpha}'\mathbf{x})$ bajo la restricción $\text{var}(\boldsymbol{\alpha}'\mathbf{x}) = 1$. Continuando de esta manera, se pueden calcular un máximo de $r = \text{rango}(\mathbf{R}_{xy})$ variables de redundancia. Debido a como fueron definidos los coeficientes de redundancia y que generalmente p es distinto de q, es claro que al intercambiar $\mathbf{x}$ con $\mathbf{y}$ no se obtienen los mismos máximos para ambos coeficientes. Por lo tanto el AR no es simétrico como el de correlación canónica.

Wollenberg [13] probó que la solución al AR son los vectores $\boldsymbol{\alpha}_i$, normalizados con varianza uno, soluciones de

$$(\mathbf{R}_{xy}\mathbf{R}_{yx} - \lambda\mathbf{R}_{xx})\,\boldsymbol{\alpha} = \mathbf{0}, \tag{1}$$

donde los $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$ asociados verifican $R(\mathbf{y}\,|\boldsymbol{\alpha}_i'\mathbf{x}) = \lambda_i$.

3. Estimación Robusta

El método clásico para estimar las variables de redundancia consiste en estimar la matriz de correlación $\mathbf{R}$ mediante la matriz de correlación muestral y luego calcular los vectores propios de (1), esto hace que las estimaciones del AR sean altamente sensibles a observaciones atípicas. Un enfoque simple para la estimación robusta es robusificar la matriz de correlación y luego aplicar el método tradicional. Así, a partir de una estimación robusta de la matriz de correlación, se calculan los vectores propios de (1) para estimar los $\boldsymbol{\alpha}_i$.

En [9], para la estimación de la matriz de correlación robusta utilizaron un M-estimador como se describe en [7] y el estimador con determinante mínimo de la matriz de covarianza propuesto en [10]. En este trabajo, se propone también utilizar estimadores robustos de posición y escala multivariados del tipo MM y τ (propuestos en [12] y [6] respectivamente).

3.1. Relación con regresión lineal multivariada

En el modelo lineal multivariado con predictores aleatorios, se tiene que el conjunto de variables $\mathbf{y}$ puede explicarse a través del grupo de variables $\mathbf{x}$ mediante:

$$\mathbf{y} = \mathbf{B}_0'\mathbf{x} + \mathbf{u}, \tag{2}$$

donde $\mathbf{B}_0 \in \mathbb{R}^{p\times q}$ es la matriz de los parámetros de regresión y $\mathbf{u}$ es un vector de dimensión q independiente de $\mathbf{x}$.

El estimador clásico de $\mathbf{B}_0$, está dado por

$$\hat{\mathbf{B}}_0 = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} = \mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}$$

debido a que las variables están estandarizadas.

Ahora consideremos la siguiente transformación ortogonal $\mathbf{T} = (\mathbf{R}_{xx}')^{-1/2}$, la cual satisface $\mathbf{T}'\mathbf{R}_{xx}\mathbf{T} = \mathbf{I}_p$. Luego, denotemos $\mathbf{x}^* = \mathbf{T}'\mathbf{x}$ y calculemos el estimador del modelo lineal multivariado de $\mathbf{y}$ dado $\mathbf{x}^*$, al cual llamaremos $\hat{\mathbf{B}}_1$. Observemos que $\hat{\mathbf{B}}_1 = \mathbf{T}'\mathbf{R}_{xy}$.

Muller [8] mostró que la solución de la ecuación (1) es equivalente a resolver el siguiente sistema de vectores propios:

$$\left(\mathbf{T}'\mathbf{R}_{xy}\mathbf{R}_{yx}\mathbf{T} - \lambda\mathbf{I}_p\right)\mathbf{a}_* = \mathbf{0} \tag{3}$$

donde $\mathbf{a}_* = \mathbf{T}^{-1}\boldsymbol{\alpha}$.

La ecuación (3) puede reescribirse como

$$\left(\hat{\mathbf{B}}_1\hat{\mathbf{B}}_1' - \lambda\mathbf{I}_p\right)\mathbf{a}_* = \mathbf{0} \tag{4}$$

Nuestra propuesta robusta para el análisis de redundancia es utilizar un estimador robusto de regresión lineal multivariada para estimar $\mathbf{B}_1$ y luego utilizar una estimación robusta de $\mathbf{T}$ para transformar las direcciones obtenidas al resolver (4) reemplazando $\hat{\mathbf{B}}_1$ por su correspondiente versión robusta.

Los estimadores de regresión lineal multivariada que proponemos utilizar son los MM propuestos en [4] y los τ propuestos en [3].

4. Estudio de Simulación

En esta sección se presentará un estudio de simulación realizado a fin de evaluar el desempeño de los estimadores propuestos con los ya existentes en la literatura bajo muestras sin y con contaminación.

Consideramos $\mathbf{y}_i, \mathbf{x}_i \in \mathbb{R}^4$ que satisfacen el modelo lineal multivariado dado por (2) con

$$B_0 = \begin{pmatrix} 3 & 3 & 3 & 0 \\ 2 & 3 & 2 & 0 \\ 4 & 5 & 5 & 0 \\ 2 & 2 & 1 & 10 \end{pmatrix}.$$

Los errores $\mathbf{u}_i$ y los predictores $\mathbf{x}_i$ son generados a partir de una distribución $\mathcal{N}_4(\mathbf{0}, \mathbf{I})$.

Bajo este modelo los coeficientes de redundancia son $R\left(\mathbf{y}\,|\boldsymbol{\alpha}_1'\mathbf{x}\right) \approx 0{,}744$, $R\left(\mathbf{y}\,|\boldsymbol{\alpha}_2'\mathbf{x}\right) \approx 0{,}220$, $R\left(\mathbf{y}\,|\boldsymbol{\alpha}_3'\mathbf{x}\right) \approx 0{,}005$ y $R\left(\mathbf{y}\,|\boldsymbol{\alpha}_4'\mathbf{x}\right) \approx 0{,}001$. Siendo entonces las dos primeras direcciones, $\boldsymbol{\alpha}_1$ y $\boldsymbol{\alpha}_2$, las más significativas.

Se generaron 200 muestras de tamaño 200. Consideramos muestras no contaminadas y muestras que contienen un $10\,\%$ de datos atípicos idénticos de la forma $(\mathbf{x}_0, \mathbf{y}_0)$ con

$$\mathbf{x}_0 = (-1, 1, 0, 1) \;\; y \;\; \mathbf{y}_0 = m\mathbf{x}_0.$$

Tomamos una malla de valores m, entre -50 y 50.

Para cada réplica j ($j = 1, \ldots, 200$) se obtuvieron las estimaciones de las direcciones $\boldsymbol{\alpha}_i$ y de los coeficientes de redundancia $R\left(\mathbf{y}\,|\boldsymbol{\alpha}_i'\mathbf{x}\right)$ para $i = 1, \ldots, 4$, denotados por $\widehat{\boldsymbol{\alpha}}_i^j$ y $\widehat{R}^j\left(\mathbf{y}\,\middle|\,\widehat{\boldsymbol{\alpha}}_i^{j\prime}\mathbf{x}\right)$, respectivamente.

Como medida de desempeño para cada coeficiente de redundancia, se calculó el error cuadrático medio dado por

$$ECM\left(\widehat{R}\left(\mathbf{y}\,|\widehat{\boldsymbol{\alpha}}_i'\mathbf{x}\right)\right) = \frac{1}{200}\sum_{j=1}^{200}\left(R\left(\mathbf{y}\,|\boldsymbol{\alpha}_i'\mathbf{x}\right) - \widehat{R}^j\left(\mathbf{y}\,\middle|\,\widehat{\boldsymbol{\alpha}}_i^{j\prime}\mathbf{x}\right)\right)^2 \text{ para } i = 1, \ldots, 4.$$

Mientras que, para cada dirección se utilizó como medida de desempeño la media de los ángulos entre la dirección estimada y la verdadera, dada por

$$ADE(\widehat{\boldsymbol{\alpha}}_i) = \frac{1}{200}\sum_{j=1}^{200}\cos^{-1}\left(\frac{\left|\boldsymbol{\alpha}_i'\widehat{\boldsymbol{\alpha}}_i^j\right|}{\|\boldsymbol{\alpha}_i\|\left\|\widehat{\boldsymbol{\alpha}}_i^j\right\|}\right).$$

Se comparó el desempeño de los siguientes estimadores: el clásico (CL), los basados en matrices de correlación robustas (CR-MCD, CR-M, CR-MM, CR-τ), el método robusto RAR propuesto por [9] basado en regresiones alternadas y los métodos propuestos basados en estimadores de regresión robusta (RR-MM y RR-τ).

Para las muestras sin contaminar el menor ADE para todas las direcciones fue alcanzado por los métodos CL y CR-M. Pero estos estimadores, cuando se utilizaron en muestras contaminadas, como era de esperarse, mostraron un crecimiento no acotado del ADE para las dos primeras direcciones cuando el valor absoluto

de m crece. Mientras que RR-MM posee el siguiente menor ADE para todas las direcciones cuando se consideraron muestras sin contaminación y posee en casi todos los valores de m el menor ADE para todas las direcciones cuando se lo aplicó a las muestras contaminadas, siendo estos valores cercanos a los obtenidos para muestras sin contaminación. El desempeño de los restantes métodos robustos es aceptable, excepto el método RAR el cual evidencia los mayores valores de ADE para las dos direcciones más significativas al utilizar muestras sin contaminación y en casi todos los valores de m para muestras contaminadas.

El desempeño de todos los métodos para los coeficientes de redundancia utilizando muestras no contaminadas es aceptable y con valores del ECM muy similares, salvo el método RAR el cual evidencia los mayores valores del ECM para los dos primeros coeficientes. Para muestras contaminadas el mejor desempeño lo presenta el método CR-MCD, el cual posee los menores valores del ECM para todos los m, y es seguido por el método RR-MM con valores del ECM muy cercanos.

AGRADECIMIENTOS

Este trabajo fue parcialmente financiado por los proyectos PIP 112-201101-00339 de CONICET, PICT 2014-0351 de ANPCyT y 20020130100279BA de la Universidad de Buenos Aires.

REFERENCIAS

[1] DESARBO W. S., *Canonical/redundancy factoring analysis.* Psychometrika, 46 (1981), pp. 307-329.

[2] FRIEDMAN B. H. AND THAYER J.E, *Facial muscle activity and eeg recordings: redundancy analysis.* Electroencephalography and Clinical Neurophysiology, 79 (1991), pp. 358-360.

[3] GARCÍA BEN, M. G., MARTINEZ, E., YOHAI, V. J., *Robust estimation for the multivariate linear model based on a τ-scale.* Journal of Multivariate Analysis, 97 (2006), pp. 1600-1622.

[4] KUDRASZOW, N.L. AND MARONNA, R.A., *Estimates of MM type for the multivariate linear model.* Journal of Multivariate Analysis, Vol. 102, 9 (2011), pp. 1280-1292.

[5] LEGENDRE P. AND ANDERSON M. J., *Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments.* Ecological Monographs, 69 (1999), pp. 1-24.

[6] LOPUHAÄ H., *Highly efficient estimators of multivariate location with high breakdown point.* Annals of Statistics, 20 (1992), pp. 398-413.

[7] MARONNA, R.A., *Robust M-Estimators of Multivariate Location and Scatter*, Annals of Statistics, 4 (1976), pp. 51-67.

[8] MULLER K., *Relationships between redundancy analysis, canonical correlation and multivariate regression*. Psychometrika, Vol 46 No 2 (1981), pp. 139-142.

[9] OLIVEIRA, M.R., BRANCO, J.A., CROUX, C. AND FILZMOSER, P.,*Robust Redundancy Analysis by Alternating Regression*. N. Balakrishnan (Ed.), Statistics for industry and technology, Birkhäuser Verlag (2004), pp. 235–246.

[10] ROUSSEEUW, P.J., *Multivariate Estimation With High Breakdown Point*, In W. Grossmann, G. Pflug, I. Vincze, and W. Wertz, editors, Mathematical Statistics and Applications, Vol. B (1985), pp. 283-297.

[11] STEWART, D. K., AND LOVE, W. A., *A general canonical correlation index*, Psychological Bulletin, 70 (1968), pp. 160-163.

[12] TATSUOKA, K.S. AND TYLER, D.E., *On the uniqueness of S-functionals and M-functionals under nonelliptical distributions.* Annals of Statistics, 28 (2000), pp. 1219-1243.

[13] VAN DEN WOLLENBERG, A. L., *Redundancy analysis: an alternative for canonical correlation analysis*, Psychometrika, 42 (1977), pp. 207-219.

[14] VAN EEUWIJK F. A., *Interpreting genotype-by-environment interaction using redundancy analysis*, Theoretical and Applied Genetics, 85 (1992), pp. 89-100.