**Ciccioli, Patricia; Bussi, Javier**

*Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística*

# COMPONENTES PRINCIPALES ESFÉRICAS Y MATRIZ DE COVARIANCIA DE DETERMINANTE MÍNIMO: UNA APLICACIÓN SOBRE INDICADORES DE CARENCIAS CRÍTICAS.

## Resumen

El Análisis de Componentes Principales (ACP) es una técnica muy utilizada dentro de los métodos estadísticos multivariados. El objetivo de este método es representar un conjunto de n observaciones con p variables a través de un número menor de variables construidas como combinaciones lineales de las originales y conservar la mayor variabilidad posible de los datos. En este trabajo se presentan dos métodos robustos, el correspondiente a Matriz de Covariancia de Determinante Mínimo (Minimum Covariance Determinant, MCD en sus siglas en inglés), y el método de Componentes Principales Esféricas (Spherical Principal Components, SPC en sus siglas en inglés). El objetivo de este trabajo es comparar estos dos métodos con el método clásico en una aplicación sobre indicadores socio-económicos de carencias críticas para ciudades y comunas de más de 2000 habitantes de la provincia de Santa Fe, los cuales provienen del Censo Nacional de Población, Hogares y Viviendas 2010. Para poder resumir las diferencias existentes entre las ciudades y comunas es necesario retener más componentes principales en los métodos robustos (MCD y SPC) que en el método clásico. La razón es que el método clásico utiliza medidas de variabilidad que son influenciadas por los valores extremos mientras que las técnicas robustas MCD y SPC utilizan una medida de dispersión sólida que está libre de este inconveniente.

## Abstract

Principal Components Analysis (PCA) is widely used in multivariate statistical analysis. The objective of this method is to represent a set of n observations with p variables through a smaller number of variables that are linear combinations of the original ones, keeping as much as possible the original variability in the data. Two robust methods are presented in this work: the Minimum Covariance Determinant method (MCD) and the Spherical Principal Components method (SPC). The objective of this work is to compare these two methods with the classic PCA when applied to data related to indicators of critical needs in cities with more than 2000 inhabitants in the province of Santa Fe. The data comes from the National Census from 2010. In order to summarize the differences among the cities it is necessary to consider a greater number of principal components in the robust methods than in the classic method. The reason is that the latter uses variability measures that are influenced by outliers while the robust methods use a solid measure that is free from this problem.

**Palabras claves: ANÁLISIS DE COMPONENTES PRINCIPALES, COMPONENTES PRINCIPALES ESFÉRICAS, MATRIZ DE COVARIANCIA DE DETERMINANTE MÍNIMO**

## 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 un conjunto de $n$ observaciones con $p$ variables a través de un número menor de variables construidas como combinaciones lineales de las originales y conservar la mayor variabilidad posible de los datos. El mismo permite una representación gráfica de los datos en un espacio de dimensión inferior, produciendo como resultado final que muchas características de los datos puedan ser resumidas en unos pocos aspectos que permitan mostrar las diferencias existentes entre las observaciones.

El enfoque clásico de ACP mide la variabilidad a través de la matriz de covariancia o correlación muestral. La presencia de observaciones atípicas puede distorsionar las estimaciones de estas matrices y en consecuencia alterar los resultados. Con el fin de solucionar este problema, se han propuesto métodos robustos para componentes principales, que se dividen según su enfoque en los métodos que consideran la estimación robusta de la matriz de covariancia y los que obtienen directamente la estimación de las componentes principales de manera robusta.

En este trabajo se presentan dos métodos, uno de cada enfoque, en primer lugar, el correspondiente a *Matriz de Covariancia de Determinante Mínimo* (Minimum Covariance Determinant, *MCD* en sus siglas en inglés), y en segundo lugar, el método de *Componentes Principales Esféricas* (Spherical Principal Components, *SPC* en sus siglas en inglés). El objetivo de este trabajo es comparar estos dos métodos con el método clásico en una aplicación sobre indicadores de carencias críticas. El conjunto de datos está formado por indicadores socio-económicos de carencias críticas para ciudades y comunas de más de 2000 habitantes de la provincia de Santa Fe, los cuales provienen del Censo Nacional de Población, Hogares y Viviendas 2010.

## METODOLOGÍA

### Análisis de Componentes Principales (ACP) Clásico

El Análisis de Componentes Principales (ACP) ha sido extensamente detallado en la literatura estadística (Johnson & Wichern, 2007; Peña, 2002). Es una técnica de análisis de datos multivariados que permite reducir la dimensión de un conjunto de datos. Al aplicarla a un grupo de $n$ observaciones en el que originalmente se miden $p$ variables, esta técnica genera $q$ $(q < p)$ nuevas variables no observables que son combinaciones lineales de las $p$ variables originales, manteniendo la mayor variabilidad posible de los datos. Así, el ACP permite una representación gráfica de los datos en un espacio de dimensión inferior y produce como resultado final que muchas características de los datos puedan ser resumidas en unos pocos aspectos que muestren las diferencias existentes entre las observaciones. El método utiliza como medida de la variabilidad de los datos la matriz de covariancia, pero cuando las variables originales están observadas en diferentes unidades de medida o los valores de las variables presentan dispersiones disímiles, entonces se recomienda trabajar con la matriz de correlación. El método fue propuesto originalmente por Karl Pearson en el año 1901, y posteriormente desarrollado en 1933 por Harold Hotelling. El número $m$ de componentes principales a seleccionar, será tal que, el porcentaje de variación parcial

recogida por las $m$ primeras componentes principales, sea un porcentaje aceptable (70% o 80%) de la variancia original. De esta manera, estas componentes pueden representar a las variables originales, con mínima pérdida de información.

**Matriz de Covariancia de Determinante Mínimo (MCD)**

Este método permite estimar la matriz de covariancia de manera robusta, la cual posteriormente puede ser utilizada para el cálculo de las componentes principales. El estimador MCD fue propuesto por Rousseeuw y Van Driessen en 1999, y ha ganado mucha popularidad por poseer propiedades asintóticas que hacen posible la comparación con otros estimadores robustos con buenas propiedades. Además, cuenta con la ventaja de ser un método excelente para detectar datos atípicos multivariados.

El estimador MCD para un conjunto de datos $x_1, x_2, \dots, x_n$ en $\mathbb{R}^p$ es definido por el siguiente subconjunto $\{x_{i1}, x_{i2}, \dots, x_{ih}\}$ de $h$ observaciones cuya matriz de covariancia posee el menor determinante a través de todos los posibles subconjuntos de tamaño $h$. La estimación MCD de localización $\widehat{\boldsymbol{\mu}}_{\boldsymbol{MCD}}$ y escala $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{MCD}}$ son respectivamente, la media aritmética y un múltiplo de la matriz de covariancia de la muestra de ese subconjunto.

$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{MCD}} = \frac{1}{h}\sum_{j=1}^{h} x_{ij}$$

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{MCD}} = C_{fcc}C_{fcmp}\frac{1}{h-1}\sum_{j=1}^{h}(x_{ij} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{MCD}})(x_{ij} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{MCD}})^T$$

donde:

$C_{fcc}$ (factor de corrección de consistencia) se selecciona de modo que el estimador tenga ciertas propiedades deseables.

$$C_{fcc} = \frac{h/n}{P(\chi^2_{p+2} < \chi^2_{p+1-(h/n)})}$$

$C_{fcmp}$ (factor de corrección para muestras pequeñas) se incluye para que el estimador sea insesgado en muestras pequeñas.

Una recomendable elección para $h$ es $h = \frac{n+p+1}{2}$ porque luego el punto de ruptura para el MCD resulta ser máximo. Si $h = n$, el estimador MCD de localización y escala se reduce a la media y matriz de covariancia de todos los datos.

El cálculo del estimador MCD está lejos de ser trivial. El algoritmo busca de forma exhaustiva todos los subconjuntos de tamaño $h$ de $n$ para encontrar el subconjunto con el

determinante más pequeño de la matriz de covariancia, pero esto es posible solamente para conjuntos de datos muy pequeños.

Para solucionar este inconveniente, existe un algoritmo muy rápido, debido a Rousseeuw y Van Driessen, que es el que se utiliza en la práctica. El algoritmo se basa en C-pasos donde C significa concentración, ya que se busca una matriz de covariancia más concentrada, con determinante más pequeño.

El algoritmo "paso-C" inicia con el cálculo de los estimadores MCD de un primer subconjunto de tamaño $h$, al que denominamos $h_1$, es decir, $(\widehat{\boldsymbol{\mu}}_{\boldsymbol{h_1}}, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{h_1}})$; en el siguiente paso se calculan los estimadores MCD para el subconjunto $h_2$, $(\widehat{\boldsymbol{\mu}}_{\boldsymbol{h_2}}, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{h_2}})$, con posible menor determinante de la matriz de covariancia, es decir, $det(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{h_2}}) \leq det\,(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{h_1}})$, mediante el cálculo de las distancias relativas:

$$d_i = \sqrt{(x_i - \widehat{\boldsymbol{\mu}}_{\boldsymbol{h_1}})^T \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{h_1}}{}^{-1} (x_i - \widehat{\boldsymbol{\mu}}_{\boldsymbol{h_1}})}$$

Luego se calcula $(\widehat{\boldsymbol{\mu}}_{\boldsymbol{h_2}}, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{h_2}})$, para el subconjunto de tamaño $h$ que tiene menor distancia. Rousseeuw y Van Driessen han demostrado que el proceso de iteración propuesta por el "paso-C" converge en un número finito de pasos a un mínimo (local).

Dado que no existe garantía de que el mínimo global de la MCD será alcanzado, la iteración se debe comenzar muchas veces de diferentes subconjuntos iniciales, para obtener una solución aproximada. El procedimiento es muy rápido para pequeños conjuntos de datos, pero para que sea realmente rápido en grandes conjuntos de datos, varias mejoras pueden ser incorporadas a través de distintos procedimientos: *Subconjuntos Iniciales, Reducción de C-pasos, Partición y Anidación*.

El estimador MCD no es muy eficiente en modelos normales, especialmente si $h$ se selecciona de modo que se obtenga el máximo punto de ruptura. Para superar la baja eficiencia del estimador MCD, se puede utilizar una versión reponderada. Para este fin a cada observación $x_i$ se le asigna el peso $w_i$, definido como:

$$\begin{cases} w_i = 1 \quad si \left( (x_i - \hat{\mu}_{MCD})^T \hat{\Sigma}_{MCD}{}^{-1} (x_i - \hat{\mu}_{MCD}) \right) \leq \chi^2_{p;0,975} \\ w_i = 0 \quad en\ otro\ caso \end{cases}$$

Luego la versión reponderada es calculada como:

$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{R}} = \frac{1}{v} \sum_{i=1}^{n} w_i x_i$$

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{R}} = C_{r.fcc} C_{r.fcmp} \frac{1}{v-1} \sum_{i=1}^{n} w_i (x_i - \widehat{\boldsymbol{\mu}}_{\boldsymbol{R}})(x_i - \widehat{\boldsymbol{\mu}}_{\boldsymbol{R}})^T$$

Donde $\nu$ es la suma de los pesos $\nu$=$\sum_{i=1}^{n} w_i$ y nuevamente el factor de multiplicación $C_{r.fcc}$ y $C_{r.fcmp}$ son seleccionados de manera que el estimador sea consistente e insesgado para muestras pequeñas. Las estimaciones reponderadas ($\widehat{\boldsymbol{\mu}}_{\boldsymbol{R}}$, $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{R}}$) que tienen el mismo punto de ruptura que el estimador inicial, tienen mejor eficiencia y es utilizada por defecto.

A continuación, se describe la metodología para la obtención de componentes principales, utilizando el método robusto MCD:

1. Se estima el vector de medias $\widehat{\boldsymbol{\mu}}_{\mathbf{MCD}}$ y la matriz de covariancia muestral $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{MCD}}$.
2. Se estiman las matrices de autovalores $\widehat{\boldsymbol{\Lambda}}_{\boldsymbol{MCD}}$ y autovectores $\widehat{\boldsymbol{B}}_{\boldsymbol{MCD}}$ de la matriz de covariancia muestral $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{MCD}}$. Cabe destacar que los autovalores estimados $\hat{\lambda}_1, \dots, \hat{\lambda}_p$ son también las variancias de las componentes principales robustas.
3. Se definen y calculan las componentes principales de la misma manera que en el ACP clásico.
4. Se calculan los porcentajes de variancias explicadas para cada una de las componentes principales de manera habitual.
5. Se selecciona el número de componentes principales por medio de alguno de los métodos de selección dispuestos a tal fin.

**Componentes Principales Esféricas (SPC)**

El método SPC es un método simple pero efectivo, propuesto por Locantore, Marron, Simpson, Tripoli, Zhang and Cohen (1999).

Obtiene directamente la estimación de las componentes principales en forma robusta, sin la necesidad de estimar en primer lugar la matriz de covariancia. La idea de este método es realizar Análisis de Componentes Principales clásica, proyectada sobre una esfera unidad.

Sea $X$ una variable aleatoria con distribución elíptica (normal multivariada), y sea $Y = (X - \boldsymbol{\mu})/\|X - \boldsymbol{\mu}\|$ la normalización de $X$ a la superficie de la esfera unidad, centrada en $\boldsymbol{\mu}$. Boente y Fraiman (1999) demostraron que los autovectores $\boldsymbol{t_1}, \dots, \boldsymbol{t_p}$ de la matriz de covariancia de $Y$ coinciden con los de $\boldsymbol{\Sigma}$ (matriz de covariancia de $X$). Ellos mostraron además que si $\sigma(.)$ es cualquiera estadística de dispersión entonces los valores $\sigma(x^T \boldsymbol{t_j})^2$ son proporcionales a los autovalores de $\boldsymbol{\Sigma}$.

Este resultado es la base para un simple enfoque robusto de componentes principales, llamado Componentes Principales Esféricas. Sea $\widehat{\boldsymbol{\mu}}$ un estimador multivariado robusto de localización, se calcula:

$$Y = \begin{cases} (x_i - \widehat{\boldsymbol{\mu}})/\|x_i - \widehat{\boldsymbol{\mu}}\| & si \quad x_i \neq \widehat{\boldsymbol{\mu}} \\ 0 & en\ otro\ caso \end{cases}$$

Sea $\widehat{\boldsymbol{V}}$ la matriz de covariancia de los $Y_i's$ con sus correspondientes autovectores $\boldsymbol{b_j}$ $(j = 1, \dots, p)$. Ahora se calcula $\widehat{\lambda_j} = \hat{\sigma}(x^T \boldsymbol{b_j})^2$ donde $\hat{\sigma}$ es un estimador robusto de dispersión (como por ejemplo el MAD). Llamamos $\hat{\lambda}_{(j)}$ al valor ordenado de los autovalores estimados, $\hat{\lambda}_{(1)} \geq \cdots \geq \hat{\lambda}_{(p)}$ y $\boldsymbol{b}_{(j)}$ el correspondiente autovector. Luego las primeras $q$ direcciones principales son dadas por los $\boldsymbol{b}_{(j)}'\boldsymbol{s}$ con $j = 1, \dots, q$.

Para que el método robusto de componentes principales esféricas sea invariante bajo transformaciones ortogonales de los datos, no es necesario que $\widehat{\boldsymbol{\mu}}$ sea afín equivariante es decir, no es necesario que los resultados permanezcan invariantes ante cualquier transformación lineal no singular, pero si ortogonal equivariante, es decir: $\widehat{\boldsymbol{\mu}}(TX) = T\widehat{\boldsymbol{\mu}}(X)$ para todo $T$ ortogonal. La elección más simple para $\widehat{\boldsymbol{\mu}}_{\boldsymbol{SPC}}$ es la "Mediana en el espacio":

$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{SPC}} = arg\min_{\mu} \sum_{i=1}^{n} \|x_i - \boldsymbol{\mu}\|$$

Esta estimación tiene un punto de ruptura igual a 0,5.

Para este método no es necesaria una estimación robusta para la matriz de covariancia, ya que se estiman directamente de manera robusta los autovalores y autovectores, obteniéndose finalmente las componentes principales esféricas.

Este procedimiento es determinístico y muy rápido, y puede ser calculado con datos colineales sin ajustes especiales. A pesar de su simplicidad el método SPC funciona muy bien.

**MATERIALES**

Los datos correspondientes a indicadores socio-económicos de carencias críticas que se utilizan para el análisis comparativo en este trabajo provienen del Censo Nacional de población, hogares y viviendas 2010 con la colaboración del Instituto Provincial de Estadística y censos (IPEC). Estos datos se encuentran disponibles en la página del instituto.

La base de datos está constituida por 10 indicadores de carencias críticas (variables socio-económicas) que fueron medidos para 158 ciudades y comunas de más de 2.000 habitantes de la provincia de Santa Fe. Este conjunto de datos presenta observaciones extremas o atípicas.

Las variables socio-económicas son:

$X_1$: Porcentajes de jefes de hogar sin asistencia escolar

$X_2$: Porcentajes de jefes de hogar con educación primaria incompleta

$X_3$: Porcentajes de jefas de hogar sin asistencia escolar

$X_4$: Porcentaje de hogares con hacinamiento por cuarto

$X_5$: Porcentaje de hogares sin caño de agua dentro de la vivienda

$X_6$: Porcentaje de hogares en vivienda sin retrete con descarga de agua

$X_7$: Porcentaje de hogares en vivienda con piso de tierra

$X_8$: Porcentaje de población de 6 a 12 años que no asiste a establecimiento educacional

$X_9$: Porcentaje de población de 14 a 19 años que asiste a nivel de instrucción primario

$X_{10}$: Porcentaje de población de 15 a 19 años que no estudia ni trabaja

Una localidad es considerada comuna si tiene entre 2.000 y 10000 habitantes y ciudad si supera los 10.000 habitantes. A continuación, se presentan algunas definiciones y aclaraciones referidas a las variables anteriormente definidas.

Hacinamiento: Corresponde a los hogares que presentan más de tres personas por cuarto.

Jefe o jefa de hogar: Es la persona considerada como tal por los demás miembros del hogar. En cada hogar hay un solo jefe o jefa, por lo tanto, hay tantos jefes y jefas como hogares.

Cañería de agua dentro de la vivienda: Sistema de suministro de agua conectado a una red de tuberías por medio de la cual se distribuye el agua en su interior.

## RESULTADOS

La Tabla Nº1 muestra algunas estadísticas básicas de las variables que fueron registradas. Se puede observar que las ciudades y comunas de Santa Fe tienen altos promedios de: Jefes de hogar con educación primaria incompleta (18,37%), hogares sin caño de agua dentro de la vivienda (12,56%), hogares en vivienda sin retrete con descarga de agua (9,71%) y población de 14 a 19 años que asiste a nivel de instrucción primario (8,88%). Las variables antes mencionadas y hogares en vivienda con piso de tierra, tienen además desvíos y rangos altos; lo cual puede deberse a la presencia de una gran cantidad de valores extremos. También se puede visualizar que hay ciudades o comunas que no tienen jefas de hogar sin asistencia escolar, viviendas con piso de tierra y población de 15 a 19 años que no estudia ni trabaja.

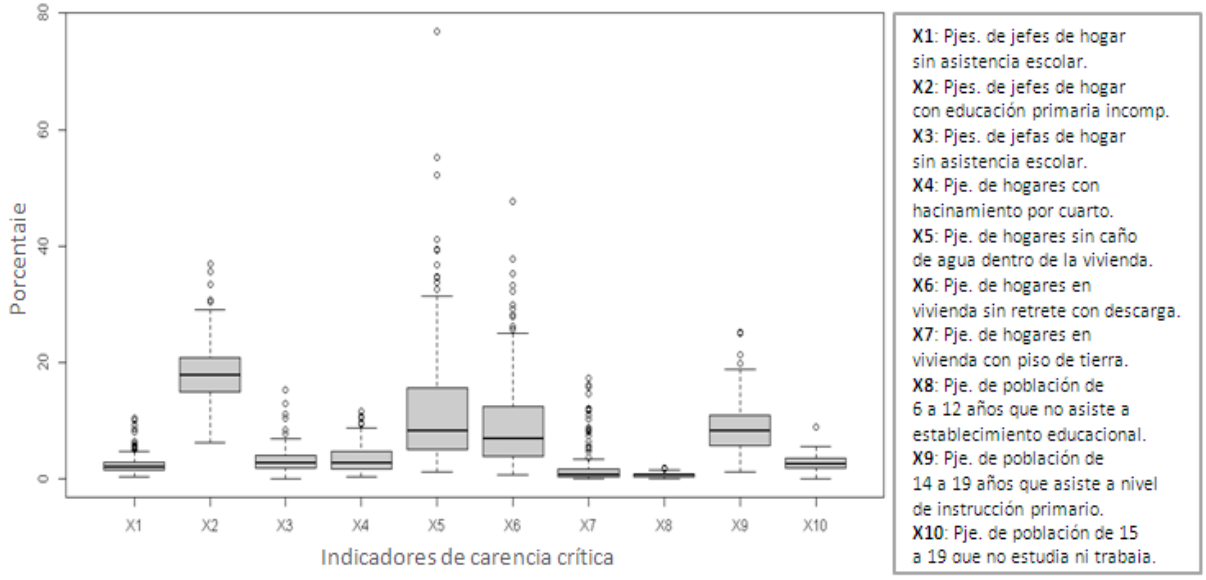
En el Gráfico Nº1 se muestran los gráficos de cajas y bigotes (boxplots). Las variables: porcentaje de jefes de hogar sin asistencia escolar ($X_1$), porcentaje de jefas de hogar sin asistencia escolar ($X_3$), porcentaje de hogares con hacinamiento por cuarto ($X_4$), porcentaje de hogares en vivienda con piso de tierra ($X_7$), porcentaje de población de 6 a 12 años que no asiste a establecimiento educacional ($X_8$) y porcentaje de población de 15 a 19 años que no estudia ni trabaja ($X_{10}$) toman valores bajos con poca variabilidad. Las variables: porcentaje de jefes de hogar con educación primaria incompleta ($X_2$), porcentaje de hogares sin caño de agua dentro de la vivienda ($X_5$), porcentaje de hogares en vivienda sin retrete con descarga de agua ($X_6$) y porcentaje de población de 14 a 19 años que asiste a nivel de instrucción primario ($X_9$), toman valores mayores y más dispersos; las variables: porcentaje de jefes de hogar con educación primaria incompleta ($X_2$) y porcentaje de población de 14 a 19 años que asiste a nivel de instrucción primario ($X_9$) parecerían presentar una distribución simétrica, mientras que porcentaje de hogares sin caño de agua dentro de la vivienda ($X_5$) y porcentaje de hogares en vivienda sin retrete con descarga de agua ($X_6$) presentan una leve asimetría hacia la derecha. Todas las variables presentan valores atípicos (outliers) de

manera individual, pero las variables: porcentaje de jefes de hogar sin asistencia escolar ($X_1$), porcentaje de jefas de hogar sin asistencia escolar ($X_3$), porcentaje de hogares sin caño de agua dentro de la vivienda ($X_5$), porcentaje de hogares en vivienda sin retrete con descarga de agua ($X_6$) y porcentaje de hogares en vivienda con piso de tierra ($X_7$) son las que presentan mayores cantidades.

Tabla Nº1: Estadísticas básicas

| | Jefes sin asistencia escolar ($X_1$) | Jefes con primaria incompl ($X_2$) | Jefas sin asistencia escolar ($X_3$) | Hogares con hacina-miento ($X_4$) | Hogares sin caño de agua ($X_5$) | Hogares sin retrete con descarga ($X_6$) | Hogares con piso de tierra ($X_7$) | Población de 6 a 12 años que no asiste a establec. educativo ($X_8$) | Población de 14 a 19 años que asiste a educ. primaria ($X_9$) | Población de 15 a 19 años que no estudia ni trabaja ($X_{10}$) |
|---|---|---|---|---|---|---|---|---|---|---|
| **Mínimo** | 0,37 | 6,30 | 0,00 | 0,43 | 1,23 | 0,78 | 0,00 | 0,00 | 1,16 | 0,00 |
| **Máximo** | 10,39 | 37,02 | 15,39 | 11,60 | 76,95 | 47,66 | 17,37 | 1,84 | 25,14 | 9,00 |
| **Media** | 2,59 | 18,37 | 3,36 | 3,43 | 12,56 | 9,71 | 2,10 | 0,72 | 8,88 | 2,70 |
| **Mediana** | 2,07 | 17,79 | 2,77 | 2,74 | 8,26 | 6,91 | 0,79 | 0,69 | 8,18 | 2,54 |
| **Desvío** | 1,72 | 5,32 | 2,28 | 2,39 | 11,60 | 8,44 | 3,46 | 0,38 | 4,39 | 1,26 |
| **R.I** | 1,33 | 6,01 | 2,04 | 2,86 | 10,43 | 8,58 | 1,31 | 0,51 | 5,27 | 1,68 |
| **Rango** | 10,02 | 30,72 | 15,39 | 11,17 | 75,72 | 46,88 | 17,37 | 1,84 | 23,98 | 9,00 |

Gráfico Nº1: Gráficos de cajas y bigotes (boxplots)




A partir de las proporciones de variancia explicada se puede observar que existe una gran diferencia entre el ACP clásico y los métodos de ACP robustos (Tabla Nº2). Si se

consideran las 3 componentes principales que explican la mayor proporción, es notorio que en el caso clásico solamente la primera explica el 87% de la variación total de las variables consideradas. Si se utilizara este método, posiblemente se seleccionaría una única componente principal reduciendo los datos a una sola dimensión. Pero si se considera el método MCD, la primera componente principal explica el 48% de la variación total, casi 40 puntos menos que en el caso clásico. Las primeras dos componentes explican el 78%, y si se deseara al menos alcanzar el mismo porcentaje que explica la primera componente del método clásico, sería necesario seleccionar 3 componentes principales, logrando una proporción explicada del 92%, solo 5 puntos por encima. Esto implicaría reducir el número de variables a 3 dimensiones. De todas formas, en el análisis se tuvieron en cuenta únicamente las dos primeras componentes principales ya que el porcentaje de variancia explicada por las mismas es aceptable para diferenciar las localidades de la provincia de Santa Fe.Si se considera el método SPC, el porcentaje de variancia de la primera componente es 73% y es necesario elegir dos componentes para alcanzar el porcentaje de variancia explicada por la primera componente principal del método clásico.

Tabla Nº2: Variancia explicada por las primeras 3 componentes principales según el método considerado (ACP Clásico/MCD/SPC).

| | | CP1 | CP2 | CP3 |
|---|---|---|---|---|
| **ACP Clásico** | **Proporción de variancia explicada** | 0,87 | 0,05 | 0,04 |
| | **Proporción acumulada** | 0,87 | 0,92 | 0,96 |
| **Matriz de Covariancia de Determinante Mínimo (MCD)** | **Proporción de variancia explicada** | 0,48 | 0,30 | 0,14 |
| | **Proporción acumulada** | 0,48 | 0,78 | 0,92 |
| **Componentes Principales Esféricas (SPC)** | **Proporción de variancia explicada** | 0,73 | 0,14 | 0,08 |
| | **Proporción acumulada** | 0,73 | 0,87 | 0,95 |

Al analizar los resultados referidos a las cargas de las componentes principales, se puede ver que los resultados difieren según el método considerado (Tabla Nº 3). A partir del ACP clásico basado en la matriz de covariancia, se puede observar que las variables que más aportan a la conformación de la primera componente principal (CP1) en sentido positivo, son: porcentaje de hogares sin caño de agua dentro de la vivienda ($X_5$) y porcentaje de hogares en vivienda sin retrete con descarga de agua ($X_6$), ambas relacionadas con las condiciones de infraestructura de la vivienda. Por lo tanto, esta componente principal va a diferenciar a las ciudades y comunas de la provincia de Santa Fe según aquellas que tengan altos niveles de carencia asociado a las condiciones de infraestructura de la vivienda y aquellas que tengan bajos niveles de carencia.

Si se considera el método MCD (Tabla Nº 3), se observa que las variables que más aportan a la conformación de la primera componente principal (CP1) en sentido positivo, son: porcentaje de hogares sin caño de agua dentro de la vivienda ($X_5$) y porcentaje de hogares en vivienda sin retrete con descarga de agua ($X_6$), ambas relacionadas con las condiciones de infraestructura de la vivienda y además porcentaje de población de 14 a 19 años que asiste a nivel de instrucción primario ($X_9$). Por lo tanto, esta componente principal distingue a las localidades de la provincia de Santa Fe según tengan altos o bajos niveles de carencia asociado a las condiciones de infraestructura de la vivienda y considera también a la educación en jóvenes. Luego, la variable que más influye en la conformación de la

segunda componente principal (CP2) en sentido positivo, es porcentaje de jefes de hogar con educación primaria incompleta ($X_2$). Por lo tanto, la segunda componente, diferencia a las localidades de Santa Fe según el grado de carencia asociado a la educación de los jefes de hogares.

Tabla Nº3: Cargas de las componentes principales seleccionadas de acuerdo a la proporción de variancia explicada según el método considerado (ACP Clásico/MCD/SPC).

| | **ACP Clásico** | **MCD** | | **SPC** | |
|---|---|---|---|---|---|
| | **CP1** | **CP1** | **CP2** | **CP1** | **CP2** |
| **Jefes sin asistencia escolar ($X_1$)** | 0,10 | 0,05 | 0,07 | 0,08 | 0,04 |
| **Jefes con primaria incompleta ($X_2$)** | 0,26 | 0,09 | **0,95** | 0,29 | **0,87** |
| **Jefas sin asistencia escolar ($X_3$)** | 0,12 | 0,05 | 0,11 | 0,10 | 0,08 |
| **Hogares con hacinamiento ($X_4$)** | 0,13 | 0,17 | -0,08 | 0,15 | -0,15 |
| **Hogares sin caño de agua ($X_5$)** | **0,73** | **0,63** | 0,11 | **0,70** | -0,01 |
| **Hogares sin retrete con descarga ($X_6$)** | **0,53** | **0,59** | -0,11 | **0,55** | -0,26 |
| **Hogares con piso de tierra ($X_7$)** | 0,18 | 0,08 | -0,01 | 0,13 | -0,03 |
| **Población de 6 a 12 que no asiste a establec. educativo ($X_8$)** | 0,01 | 0,02 | -0,02 | 0,01 | -0,02 |
| **Población de 14 a 19 que asiste a educ. primaria ($X_9$)** | 0,20 | **0,45** | -0,20 | 0,26 | **-0,36** |
| **Población de 15 a 19 que no estudia ni trabaja ($X_{10}$)** | 0,01 | 0,08 | -0,09 | 0,01 | -0,12 |

En el caso del análisis a través de método SPC (Tabla Nº 3)**,** se visualiza que las variables que más influyen sobre la primera componente principal (CP1) en sentido positivo, son: porcentaje de hogares sin caño de agua dentro de la vivienda ($X_5$) y porcentaje de hogares en vivienda sin retrete con descarga de agua ($X_6$). Por lo tanto, esta componente principal distingue a las ciudades y comunas de la provincia de Santa Fe según tengan altos o bajos niveles de carencia asociado a las condiciones de infraestructura de la vivienda. Las variables que más aportan a la conformación de la segunda componente principal (CP2) son: porcentaje de jefes de hogar con educación primaria incompleta ($X_2$), en sentido positivo y porcentaje de población de 14 a 19 años que asiste al nivel de instrucción primario ($X_9$), en sentido negativo. Por lo tanto, esta componente principal distingue a las localidades de la provincia de Santa Fe, según tengan alto grado de carencia asociado a la educación en adultos y bajo porcentaje de adolescentes que reciben educación primaria o bajo grado de carencia relacionado a la educación en adultos y alto porcentaje de adolescentes en escuela primaria.

Se observa que en general el método clásico de componentes principales asocia el significado de carencia socio-económico con las condiciones de infraestructura de la vivienda, mientras que, en los métodos robustos además de considerarse las condiciones de infraestructura de la vivienda también se toma en cuenta el grado de educación de jóvenes y adultos. Esto puede deberse a que, por ejemplo, la variable porcentaje de población de 14 a 19 años que asiste a nivel de instrucción primario ($X_9$) tiene pocos outliers y por lo tanto la medida de variabilidad del método clásico no la toma en cuenta, mientras que las variables ($X_5$) y ($X_6$) tienen gran variabilidad y una gran presencia de outliers que toman valores muy altos y dispersos, haciendo que presenten altas cargas en la primera componente principal del método clásico.

En el caso robusto se observa que las interpretaciones varían según el método. Pero si se consideran las 2 primeras componentes principales, en ambos métodos se asignan las cargas más importantes a las mismas variables (Tabla Nº 3).

### CONCLUSIONES

En este trabajo se presentan dos técnicas robustas para el análisis de componentes principales: Matriz de Covariancia de Determinante Mínimo (MCD) y Componentes Principales Esféricas (SPC) y se las compara con el Análisis de Componentes Principales (ACP) clásico en una aplicación sobre indicadores de carencias críticas.

Para poder resumir las diferencias sociales y económicas existentes entre las ciudades y comunas de Santa Fe es necesario retener más componentes principales en los métodos robustos (MCD y SPC) que en el método clásico. Se observa que para el método clásico solo una componente principal es suficiente mientras que, para los métodos robustos MCD y SPC se necesitan al menos dos componentes principales para poder resumir las diferencias presentes en los datos.

Puede notarse que las variables que más aportan en la conformación de la primera componente principal en el método clásico son aquellas que contienen una mayor variabilidad en los datos con una gran cantidad de outliers dispersos, los cuales toman valores altos. De esta manera, se puede observar que el método clásico está influenciado por valores extremos, dando resultados e interpretaciones que pueden estar alejados del comportamiento del conjunto central de datos que representa la gran mayoría de ellos.

## DISCUSIÓN

Se puede observar que en las componentes principales clásicas se concentra más la variabilidad explicada en la primera componente a diferencia de lo que ocurre en las componentes principales robustas. La razón es que las dos estimaciones utilizan diferentes medidas de variabilidad; el método clásico utiliza medidas de variabilidad que son influenciadas por los valores extremos, y así grandes valores atípicos en la dirección de los primeros ejes principales inflará las variancias correspondientes y por lo tanto, aumentará su proporción de variabilidad explicada. Por otro lado, las técnicas robustas MCD y SPC utilizan una medida de dispersión sólida que está libre de este inconveniente, dando una medida más exacta de la variabilidad no explicada por la mayor parte de los datos. Las componentes principales robustas no se ven afectadas por estos valores extremos y, por lo tanto, muestran resultados, que al no verse distorsionados, representan las características principales de la mayoría de las observaciones de manera más apropiada.

Si bien los métodos robustos no obtienen exactamente los mismos resultados, reducen la dimensión de los datos a la misma cantidad de componentes principales. Dado que su

aplicación es sencilla, es conveniente utilizar estos métodos cuando se presentan observaciones atípicas.

## REFERENCIAS BIBLIOGRÁFICAS

**Allasia, M.B.** (2013). *Métodos Estadísticos Robustos en el Contexto de Aplicaciones de Calidad*. Tesina de Licenciatura. Universidad Nacional de Rosario, Argentina.

**Boente, G., Fraiman, R.** (1999). *Discussion on robust principal analysis for functional data by N. Locantore, J. Marron, D. Simpson, N. Tripoli, J. Zhang and K. Cohen.*

**Bussi, J.; Ciccioli, P.** (2015). *Una revisión de los distintos métodos robustos para el análisis de componentes principales*. Vigésimas Jornadas "Investigaciones en la Facultad" de Ciencias Económicas y Estadística.

**Ciccioli, P.** (2016). *Métodos Estadísticos Robustos para el Análisis de Componentes Principales: Estudio sobre indicadores de carencias críticas.* Tesina de Licenciatura, Facultad de Ciencias Económicas y Estadística, Universidad Nacional de Rosario.

**Hualpa Benavente, F.P.** (2012). *Componentes principales mediante el método robusto MCD: Matriz de covarianzas de determinante mínimo.* Tesina de Licenciatura. Universidad Nacional Mayor de San Marcos, Perú.

**Huber, P.J.; Ronchetti, E.,** (2009), *Robust Statistics. Second edition*, New York: John Wiley & Sons, Inc.

**Hubert, M., Rousseeuw, P. J.** (2005). ROBPCA: *A New Approach to Robust Principal Component Analysis. Technometrics*.

**Johnson, R.A.,Wichern, D.W** (2007), *Applied Multivariate Statistical Analysis*. *Sixth Edition*, Prentice Hall.

**Jureckova, J., Picek, J.** (2006). *Robust Statistical Methods with R.* USA: Chapman & Hall/CRC.

**Locantore, J. Marron, D. Simpson, N. Tripoli, J. Zhang and K. Cohen.** (1999). *Robust principal components analysis for functional data.*

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

**Maronna, R.A , Zamar, R** (2002). *Robust estimation of location and dispersión for hight dimensional data sets. Technometrics.*

**Montaño, N., Zurita, G.** (2009). *Estimadores Robustos para el vector de medias y la matriz de varianzas y covarianzas de vectores aleatorios multivariados*. *Rev.Tecnológica ESPOL*.

**Peña, D.** (2002), *Análisis de datos multivariantes*. *Alianza Editora, Madrid, España.*

**Rousseeuw, P. J, Driessen, K. V.** (1999). *A fast algorithm for the minimum covariance determinant estimator. Technometrics.*

**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.

**SAS Institute Inc.** (2008). *SAS/STAT 9.0 User`s Guide*. Cary, NC:SAS Institute Inc.

**Todorov, V., Filzmoser, P.** (2009). *An object-oriented framework for robust multivariate analysis. Journal of Statistical Software*.