El Mercado Automovilístico: Un caso para el Análisis Estadístico Multivariante

José Jaén Delgado Grupo 28

1 Introducción

El medio de transporte más utilizado es el coche y por ello un estudio sobre los automóviles puede resultar provechoso para un público considerable: desde los propios usuarios hasta los reguladores políticos. Gracias a que dichos vehículos cuentan con multitud de características mesurables, el Análisis Multivariante resulta una de las técnicas estadísticas más interesantes para obtener información importante.

2 Metodología del Proyecto

Utilizaremos la plataforma computacional **MATLAB** al objeto de continuar nuestro estudio sobre la base de datos de automóviles. El foco de atención pasa a la aplicación de métodos estadísticos de visualización y reducción de dimensionalidad.

El código para realizar las operaciones se incluirá en un apéndice al final del trabajo. Salvo que se especifique lo contrario, la autoría de las diferentes aplicaciones no predifinidas en **MATLAB** de las técnicas estadísticas que emplearemos se atribuye a Baíllo & Grané (2007).

3 Descripción de los Datos

Recordamos la composición y origen de la base de información relevante para el proyecto.

Los datos se han obtenido de una de las páginas web más conocidas dentro de la comunidad Data Science: ***UCI ML Repository***, de la Universidad de California.

La base de datos con la que trabajaremos cuenta con un total de **199** coches y originalmente fue diseñada con el propósito de desarrollar modelos de ***Machine Learning*** para predecir el precio de los automóviles.

Contamos con un total de **11 variables** o rasgos explicativos, de las cuales **9** son **cuantitativas** y el resto categóricas. Dentro de estas últimas, trabajeremos con **una variable binaria** y **una multiestado**.

A continuación se presenta una lista detallada de cada variable:

A) Variables Cuantitativas:

- `distancia_ejes`: Distancia entre ruedas delanteras y traseras en centímetros
- `largo`: Medido en centímetros
- `ancho`: Medido en centímetros
- `altura`: Medido en centímetros
- `peso`: En libras
- `motor`: Dimensión del motor
- `caballos`: Potencia del coche en caballos de vapor
- `max_revoluciones`: Máximo de revoluciones por minuto
- `precio`: Cuantía de venta en dólares

B) Variable Binaria:

- `combustible`: Toma valor 1 si es gasolina, en caso de diésel es 0

C) Variable Multiestado:

- `rueda_motriz`: Tracción trasera (valor 1), Tracción delantera (valor 2), 4x4 (valor 3)

Nótese que se ha realizado una selección concreta de las variables del conjunto de datos original, pues en realidad se contaba con un total de 26. Las transformaciones no lineales realizadas previamente no serán tomadas en cuenta, si bien `peso` se expresará en kilogramos, `max_revoluciones` y `precio` serán divididas entre 10 y 100 respectivamente. Consecuentemente trabajaremos con cientos de dólares como unidad de medida en el caso de los precios y una cantidad de revoluciones por minuto reducida

4 Reducción de Dimensionalidad

Antes de dar paso a la aplicación de las técnicas estadísticas procedemos a exponer brevemente el objetivo que se persigue con las mismas, explicando su utilidad.

Actualmente, el ***Big Data*** o volumen masivo de datos es uno de los temas dominantes en el mundo de la Analítica. Esto se explica en buena parte debido a los grandes avances de las tecnologías de ***Cloud Computing*** y la aparición de aplicaciones capaces de procesar bases de datos complejas y de considerable tamaño. Por ello, se precisan de métodos que faciliten el análisis de dicha información.

Los algoritmos de ***Machine Learning*** y ***Deep Learning*** cuando son aplicados a grandes bases de datos presentan usos de memoria y espacios de tiempo de entrenamiento distantes de ser calificados de intrascendentes. Igualmente, cualquier modelo que utilice todas las variables explicativas para tareas de predicción o clasificación puede 'sobreajustarse' a los datos disponibles (*overfitting*). El fin de los modelos no es tanto explicar de forma exhaustiva una base de datos específica, sino generalizar de forma correcta a nuevas observaciones.

Consecuentemente, contar con muchas variables explicativas da lugar a la 'maldición de la dimensionalidad' o ***curse of dimensionality***: si bien es deseable que el investigador tenga disponible información suficiente, también es cierto que la complejidad computacional y la precisión de los modelos se resienten.

De forma meramente informativa, entre las técnicas de reducción de la dimensionalidad detacan la ***Regularización L1*** o ***Regresión LASSO*** propuesta por Tibshirani (1996), la cual asigna valores nulos a coeficientes de las variables más prescindibles. En la misma línea se encuentra el uso de ***Autoencoders***, idea originalmente desarrollada por Ballard (1987), consistente en Redes Neuronales Artificiales que a partir de los propios inputs descubren estructuras dentro de los datos que permiten un *feature learning* eficiente. Se obtiene una codificación de los datos que posteriormente se valida con las variables originales.

En este proyecto ponemos en práctica las ideas expresadas, concretamente mediante el ***Análisis de Componentes Principales*** (PCA por sus siglas en inglés) y el ***Multidimensional Scaling*** (MDS). Conviene señalar como limitación que tanto el número de observaciones como el de variables, 199 y 11 respectivamente, no presentan un serio problema en términos de logística de modelización o análisis. No obstante, como ejemplo ilustrativo de la utilidad de ambas técnicas y de la posibilidad de descubrir información valiosa resultan del todo convenientes.

5 Análisis de Componentes Principales

La siguiente sección está dedicada al empleo de PCA en nuestro conjunto de datos, técnica desarrollada por Pearson (1921) y Hotelling (1933). El objetivo principal será describir la información contenida en la **matriz centrada** de **datos cuantitativos** mediante un conjunto de variables menor que el de variables originales. Dicho fin se conseguirá aprovechando la correlación existente entre las columnas de la matriz previamente mencionada.

Por el Teorema de la Dimensión, si el rango de la matriz de covarianzas, r, definido como $r = rang(\mathbf{S})$ es menor o igual que el número de variables (p) la diferencia entre éste y el rango representa las p - r variables que son combinaciones lineales de otras (y por tanto las podremos descartar sin pérdida de generalidad). De forma análoga se puede emplear la matriz de correlaciones $\mathbf{R}$ para llegar a la misma conclusión. La elección de $\mathbf{S}$ o $\mathbf{R}$ como base para realizar PCA dependerá de la varianza de las variables cuantitativas, optando por $\mathbf{R}$ si los valores difieren sustancialmente.

En nuestro caso los valores de las varianzas son parecidos únicamente para las variables medidas en centímetros, por lo que las conclusiones principales las obtendremos en base a $\mathbf{R}$. Modificamos la función `comp2()` de Baíllo & Grané (2007) para dividir los elementos de los Componentes Principales por su desviación típica.

Dado que la relación entre las variables es tan importante para el uso correcto de PCA, calculamos una medida escalar de interdependencia lineal, η^2 o coeficiente de intensidad de relaciones lineales entre los datos cuantitativos, definido como:

$$\eta^2 = 1 - \det(\mathbf{R})$$

Para nuestros datos η^2 es muy cercano a 1 (ausencia de incorrelación) por lo que no es esperable que muchos Componentes Principales sean necesarios para describir toda la información importante de la matriz.

Definimos los Componentes Principales $\tilde{\mathbf{Y}}$ como una combinación obtenida a partir de las variables originales:

$$\tilde{\mathbf{Y}} = \mathbf{X}\mathbf{D}_s^{-1}\tilde{\mathbf{T}}, \quad \mathbf{D}_s^{-1} \coloneqq \text{diag}(\sqrt{S_{11}}, \ldots, \sqrt{S_{pp}})$$

Donde $\tilde{\mathbf{T}}$ son los autovectores de la matriz de correlaciones $\mathbf{R}$ de los datos cuantitativos. Aprovechamos que $\mathbf{cR} \geq 0$ ($\forall \mathbf{c} \neq 0$) y simétrica por lo que su descomposición espectral es:

$$\mathbf{R} = \tilde{\mathbf{T}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{T}}'$$

Para determinar el número de Componentes Principales finales utilizaremos el Criterio de Kaisser modificado o corrección de Jollife. No incluimos los componentes cuyos autovalores sean menores que 0.7, ya que se ha comprobado que cuando $p \leq 20$ (como en nuestro caso) el criterio de Kaisser tiende a incluir pocos componentes.

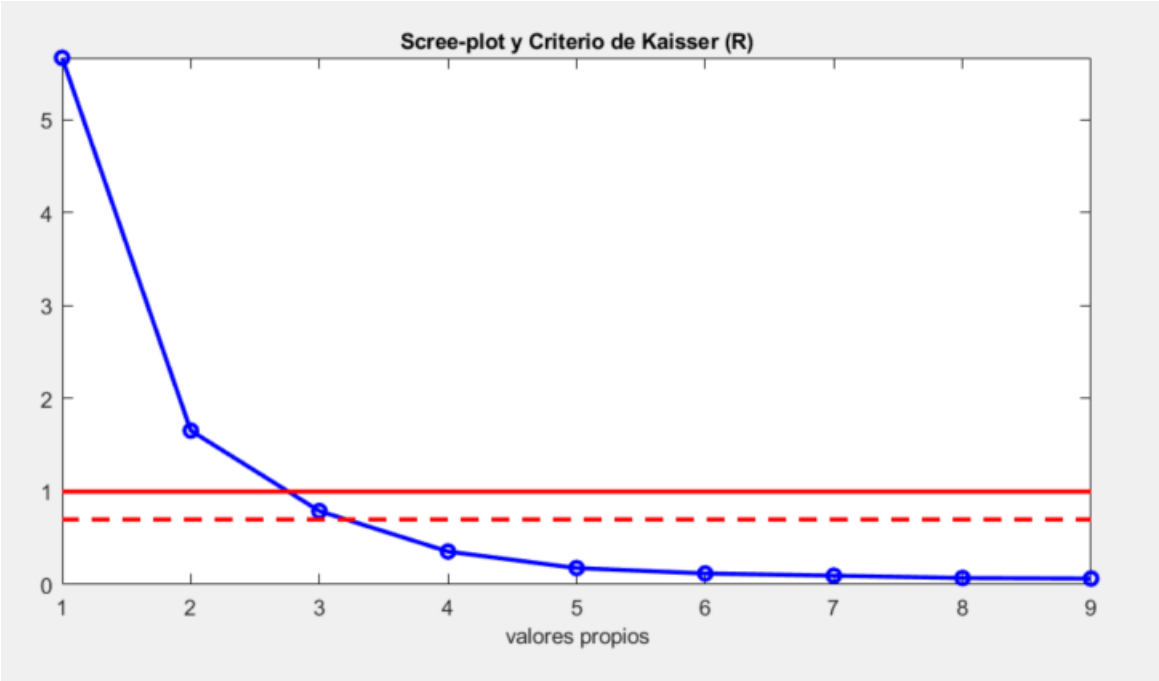

Figure 1: Selección PCA con Criterio Kaisser Modificado

Apreciamos el efecto de la corrección de Jollife (línea discontinua roja) al incluir un tercer componente principal que ha sido excluido por el criterio de Kaisser (línea continua roja). Por tanto, interpretaremos los resultados relevantes empleando tres Componentes Principales.

Obtenemos los valores de cada componente a continuación:

$$
\begin{aligned}
Y_1 &= 0.0584X_1 + 0.0637X_2 + 0.0633X_3 + 0.0259X_4 \\
&\quad + 0.0669X_5 + 0.0607X_6 + 0.0520X_7 - 0.0209X_8 + 0.0605X_9 \\
Y_2 &= -0.0252X_1 - 0.0124X_2 - 0.0024X_3 - 0.0480X_4 \\
&\quad + 0.0032X_5 + 0.0185X_6 + 0.0365X_7 + 0.0372X_8 + 0.0199X_9 \\
Y_3 &= 0.0840X_1 + 0.0799X_2 + 0.0318X_3 + 0.1943X_4 \\
&\quad - 0.0124X_5 - 0.1272X_6 + 0.0124X_7 + 0.3958X_8 - 0.0140X_9
\end{aligned}
$$

La representación gráfica de dichos componentes quedaría de la siguiente forma:

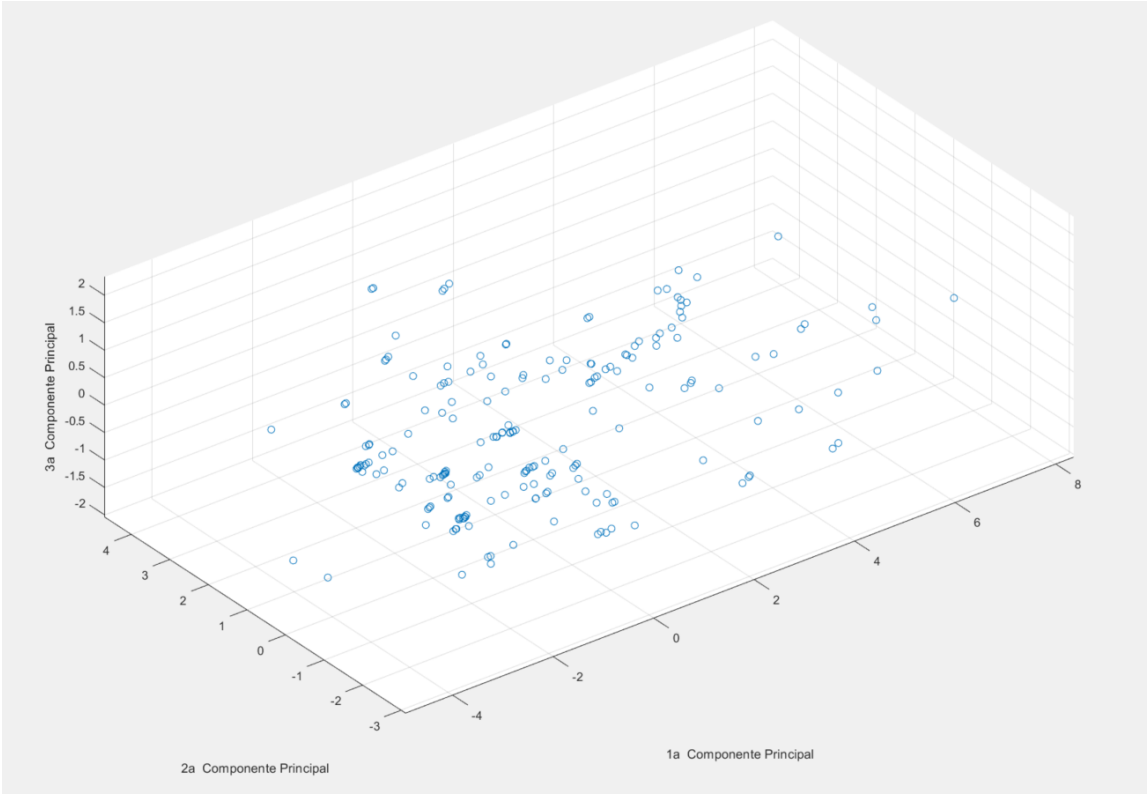

Figure 2: PCA a partir de $\mathbf{R}$ (90.16 %)

Los componentes Y_1, Y_2, Y_3 explican el 90.16% de la variabilidad de la matriz de datos, por lo que la modificación de Jollife coincide con el valor usual del criterio de Porcentaje Explicado (al menos 90%), escenario que no se daría si únicamente explotáramos la información propuesta por los dos primeros componentes (81.37%) y mucho menos por el primero solamente (62.96%).

Asimismo, una rápida revisión de la Figura 1 revela que los descensos de pendiente son poco significativos a partir del tercer autovalor (sería posible considerar los dos primeros autovalores aunque es una perspectiva muy agresiva bajo nuestro punto de vista). Consecuentemente, dado que el criterio de Porcentaje Explicado y el Scree test de Cattell (visualización de pendientes) coinciden, nuestra decisión de incluir tres Componentes Principales a partir del criterio de Kaisser modificado se ve reforzada.

Procediendo a la interpretación de los Componentes Principales cabe apuntar que no nos consideramos expertos en la materia objeto de estudio y por tanto las asociaciones que presentamos pueden no corresponderse fielmente con la realidad técnica de los

automóviles. Pese a que la subjetividad es ubicua en la Estadística (sirva como ejemplo la selección de un a priori adecuado en los Métodos Bayesianos), es deseable contrastar información con profesionales del área del conocimiento que se está analizando. En nuestro caso hemos consultado funtes de información especializadas en coches.

Es reseñable que en Y_1 todos los coeficientes son positivos (a excepción de X_8 o `max_revoluciones`), por lo que podría considerarse un buen proxy del tamaño del coche. Siendo cierto que los índices de tamaño en PCA requieren signos positivos en todos los coeficientes, hemos de señalar que en el proyecto anterior `max_revoluciones` era una de las variables que presentaba correlación más débil con el resto, y de hecho en valor absoluto es la que menos contribuye a Y_1 (`largo`, `ancho`, `peso` las que más).

Para Y_2 ocurre un caso diferente, los cuatro primeros coeficientes son negativos, coincidentes con las dimensiones del coche medidas en centímetros. La aseguradora británica Zuto (2020) realizó un estudio concluyendo que los coches modernos pesan un 70% más que los antiguos dentro de la misma gama. Además, los avances tecnológicos han permitido mejorar considerablemente la potencia de los automóviles, incrementando los caballos de vapor, las revoluciones máximas y la dimensión del motor. Sobre este último punto, la compañía LeasePlan (2017) apunta que los coches eléctricos presentan un mayor motor, dado que para producir un mayor amparaje se requieren bobinas más grandes. Por tanto, Y_2 puede interpretarse como un índice de modernidad del coche, siendo `altura`, `caballos` y `max_revoluciones` las variables más determinantes (valor absoluto mayor).

Finalmente, en el caso de Y_3, nuestro último Componente Principal, la tarea interpretativa se vuelve más intricada. Por simplicidad obviaremos las variables con menor valor absoluto. Nos centramos pues en `max_revoluciones`, `altura`, `motor` y `distancia_ejes`. Únicamente `motor` tiene un valor negativo, por lo que Y_3 es candidato a discernir entre los coches de grandes dimensiones aquellos que mejor optimizan el tamaño del motor. Se penaliza asimismo a coches de pequeña dimensión, y aún más a éstos mismos que no aprovechen los recursos para fabricar un motor potente que no ocupe mucho espacio.

De esta manera, mientras un coche registre datos cuyos valores absolutos sean de considerable cuantía, lo interpretaremos como un automóvil de gran 'tamaño' (proxy), y por tanto obtendrá grandes resultados positivos en el eje '1ª PCA' de nuestro *gráfico tridimensional*. Si el vehículo en cuestión sobresale en cuanto a su peso y no destaca tanto por su altura o anchura sino por las características técnicas, es indicativo de que

es un modelo moderno o con tecnología sofisticada (hacia los puntos máximos del eje '2ª PCA'). Por último, el eje '3ª PCA' muestra en sus valores superiores aquellos coches de importante altura, revoluciones máximas y distancia entre los ejes que maximizan la potencia dado un menor motor. En los valores intermedios de dicho eje podremos encontrar coches más pequeños pero eficientes en la creación del motor.

En cuanto a las correlaciones, un breve repaso de las mismas nos puede ayudar a identificar la procedencia de los diferentes componentes principales.

Table 1: Correlaciones con Componentes Principales

Variables	Y_1	Y_2	Y_3
`distancia_ejes`	0.8441	-0.3997	0.1574
`largo`	0.9207	-0.1974	0.1496
`ancho`	0.9151	-0.0376	0.05947
`altura`	0.3749	-0.7629	0.3639
`peso`	0.9669	0.0502	-0.0231
`motor`	0.8772	0.2946	-0.2383
`caballos`	0.7521	0.5793	0.0232
`max_revoluciones`	-0.3028	0.5916	0.7413
`precio`	0.8756	0.3154	-0.0262

El primer componente principal guarda una gran correlación con el peso, largo, ancho, dimensión del motor, precio y distancia entre los ejes, siendo asimismo fuerte con los caballos de vapor. Tiene sentido pues interpretar Y_1 como una aproximación al tamaño del coche, puntualizando una vez más que la correlación con `max_revoluciones` es débil.

Por su parte, Y_2 está correlada fuertemente con la altura, seguido de los caballos, las revoluciones máximas y el peso, rasgos distintivos de la modernidad del coche como habíamos apuntado. Los valores de las correlaciones empiezan a resentirse y son significativamente menores que Y_1, dado que la variabilidad explicada de los datos cuantitativos es menor.

Y_3 presenta correlaciones menores, explicando en menor medida la información que tenemos. Destacan las revoluciones máximas seguidas por la altura, el motor y distancia entre ejes. La interpretación de este componente es compleja y consideramos que puede servir más bien para identificar la sofisticación ingenieril del coche, en tanto en cuanto premia a aquellos vehículos con menor motor.

6 Multidimensional Scaling

Como uno de los inconvenientes de PCA identificamos el abandono de las variables cualitativas para su aplicación. Es posible superar este problema con MDS, pues en vez de emplear una matriz de datos cuantitativos como input, se utiliza una **matriz de cuadrados de distancias**. Este método fue propuesto por Young & Householder (1938), redescubierto por Torgerson (1952) y Gower (1966).

El objetivo será obtener unos ejes o **Coordenadas Principales** $Y_1, \dots Y_m$ a partir de una matriz de distancias adecuada, donde sea posible realizar una representación euclídea que coincida con las distancias calculadas. Las principales dificultades asociadas con la aplicación de MDS son la interpretación de las Coordenadas Principales y la complejitud computacional.

Algorithm 1: Algoritmo de obtención representación MDS

Input: Matriz de datos mixtos $\mathbf{X}$ con variables cuantitativas $\mathbf{X}_1$ y cualitativas $\mathbf{X}_2$

Output: Coordenadas Principales $Y_1, Y_2, \dots, Y_m$

1. Ordenar la matriz de datos: $\mathbf{X} = [\mathbf{X}_1\, \mathbf{X}_2]$

2. Calcular:

$$\mathbf{M} = (\mathbf{x}_{i1} - \mathbf{x}_{j1})' \, \mathbf{S}^{-1} \, (\mathbf{x}_{i1} - \mathbf{x}_{j1}) \qquad \mathbf{C} = \alpha \oslash p$$

$$\mathbf{M}^{(2)} = \mathbf{M}^{\circ 1/2} \circ \mathbf{M}^{\circ 1/2} \qquad \mathbf{C}^{(2)} = 2(\mathbf{1}\mathbf{1}' - \mathbf{C})$$

$$\mathbf{M}^* = \mathbf{M}^{(2)} / \texttt{vgeom}(\mathbf{M}^{(2)}) \qquad \mathbf{C}^* = \mathbf{C}^{(2)} / \texttt{vgeom}(\mathbf{C}^{(2)})$$

$$\mathbf{D}^{(2)} = \mathbf{M}^* + \mathbf{C}^*$$

3. Construir:

$$\textbf{i)}\ \mathbf{G} = -\frac{1}{2}\mathbf{H}\mathbf{D}^{(2)}\mathbf{H} \ \ (\text{si } \mathbf{c}\mathbf{G}\mathbf{c}' \geq 0 \ \ \forall \mathbf{c} \neq \mathbf{0})$$

$$\textbf{ii)}\ \tilde{\delta}_{ij}^2 = \begin{cases} \delta_{ij}^2 + c & \text{si} \quad i \neq j \\ 0 & \text{si} \quad i = j \end{cases} \quad \text{y volver a } \textbf{i)} \text{ (en caso contrario)}$$

4. Diagonalizar: $\mathbf{G} = \mathbf{U}\Lambda\mathbf{U}'$

5. Obtener: $\mathbf{Y} = \mathbf{U}\Lambda^{1/2}$

En el algoritmo previamente expuesto puede apreciarse cómo modificamos la Distancia de Gower al sumar matrices de cuadrados de distancias con una misma variabilidad geométrica. La primera de éstas, $\mathbf{M}^*$ mide la disparidad entre las variables cuantitativas, mientras que la segunda, $\mathbf{C}^*$, realiza lo mismo con las cualitativas. Nótese que hemos procedido a agrupar la variable binaria y multiestado de nuestro conjunto de datos dado que de otra forma se daba demasiado peso a éstas, llegando a perturbar considerablemente la visualización de la configuración MDS. Esto se debe a que la Distancia de Gower 'penaliza' las variables cuantitativas, siendo necesario para corregirlo un método de estimación robusta. En lo atinente a la distancia de las variables cuantitativas, se ha optado por el empleo de la Distancia de Mahalanobis pues es resiliente a los cambios de escala y no ignora la introducción de variables redundantes. Para las variables cualitativas comenzamos calculando la similaridad, dividiendo el número de atributos coincidentes entre observaciones con el total de variables explicativas (*matching coefficients*), y posteriormente se transforma a distancia estadística.

Hemos calculado asimismo la Distancia de Gower sin nuevas implementaciones al objeto de comparar ambas. Para ello es necesario imponer la condición de misma variabilidad geométrica, como propusieron Cuadras & Fortiana (1995), ya que si no cualquier ejercicio de paralelismo no resultaría del todo correcto.

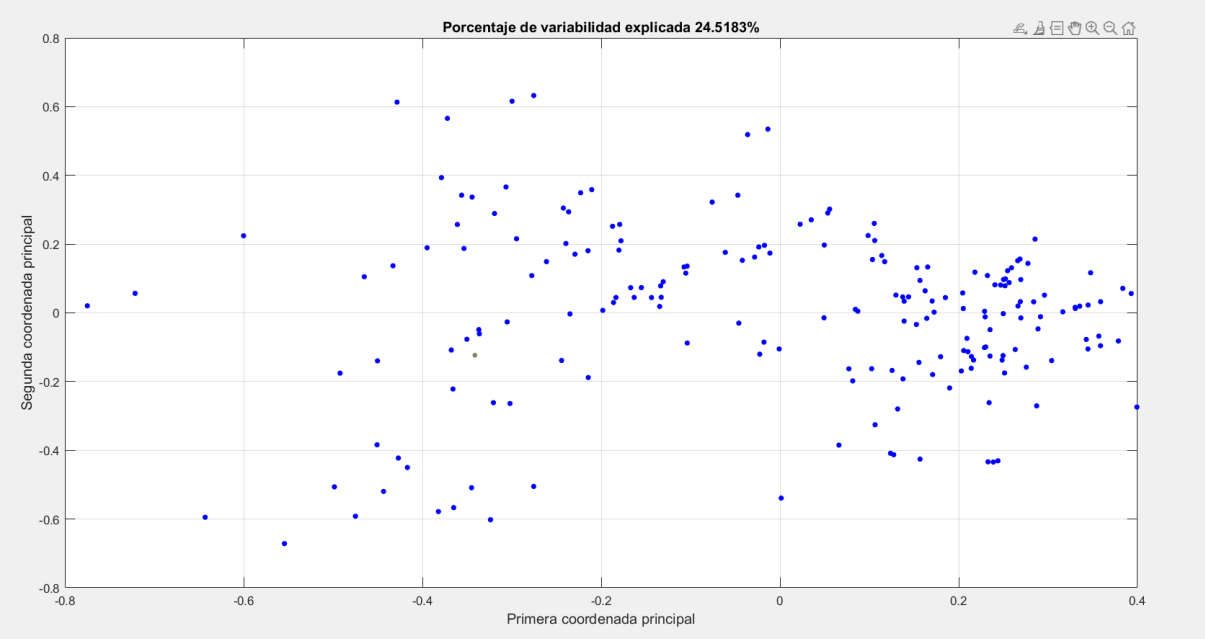

Figure 3: MDS a partir de Distancia de Gower propia (24.52 %)

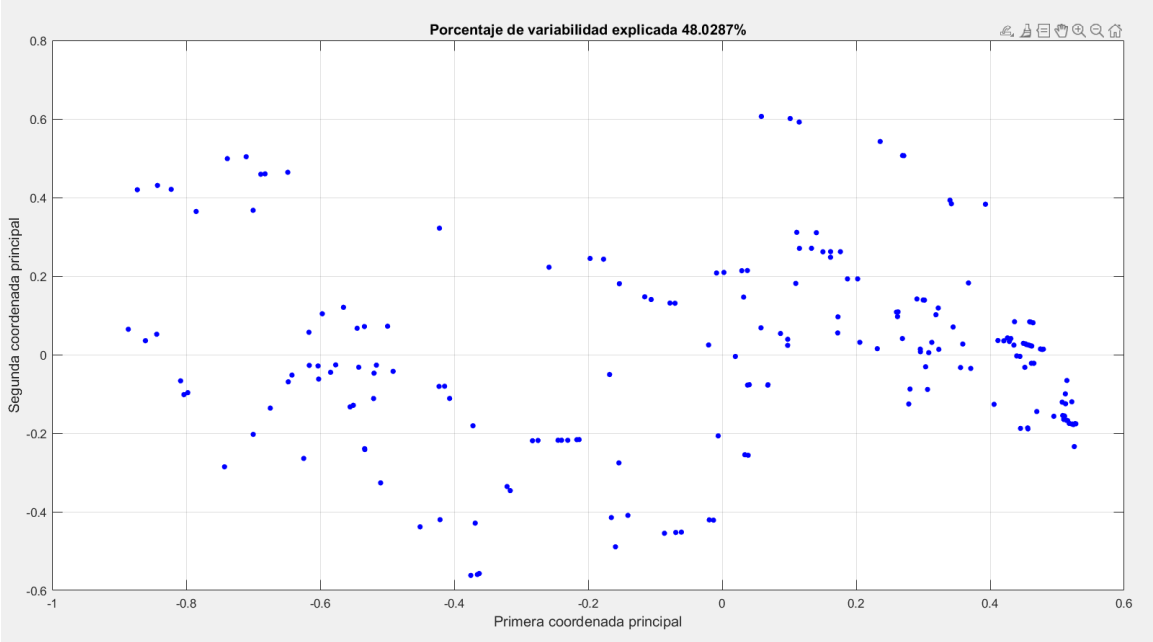

Figure 4: MDS a partir de Distancia de Gower (48.03 %)

La variabilidad explicada de la Distancia de Gower sigue una evolución no lineal, aumentando cada vez a menor ritmo. Aportamos gráfico de la Distancia que hemos propuesto. Se aprecia una tendencia de incremento de variabilidad explicada en una razón de 10 puntos porcentuales por cada autovalor, si bien decae a partir del noveno.

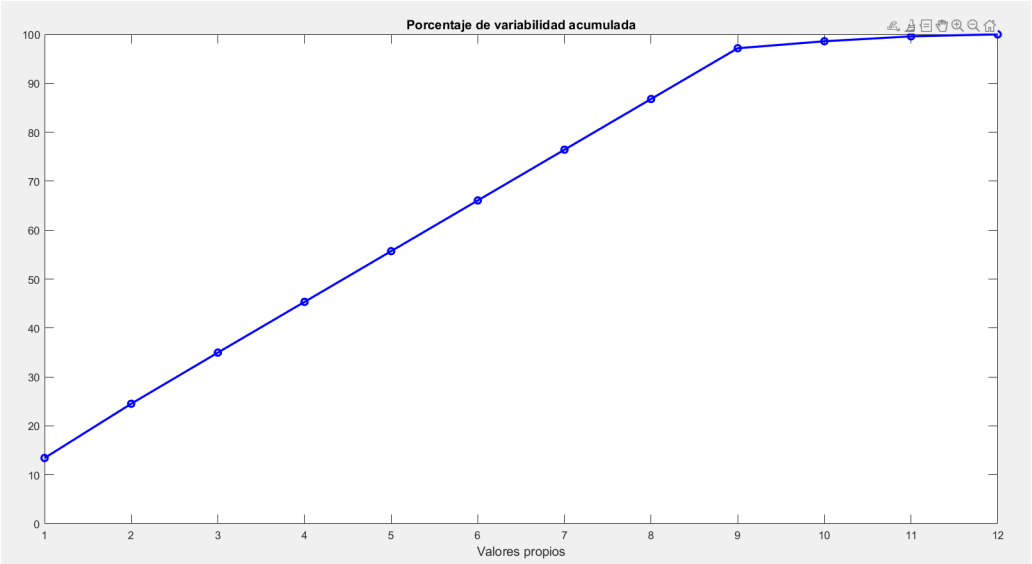

Figure 5: Porcentaje de variabilidad explicado

Empleando nuestra distancia observamos una mayor cantidad de puntos concentrados en los valores positivos del eje de abscisas (en adelante eje Y_1). La diferencia podría encontrarse en que la Distancia de Mahalanobis considera ciertos efectos que la Distancia Manhattan (utilizada por la Distancia de Gower) no podía superar.

Aunque en el caso de nuestra propia distancia la variabilidad explicada por las dos primeras Coordenadas Principales es considerablemente menor (24.52% contra un 48.03%), no deberíamos guiarnos por este criterio para decantarnos por una distancia u otra, pues se ignoran aspectos como una correcta definición de distancias, etc.

Resulta lógico que en un contexto de reducción de la dimensionalidad se descarte el criterio del 90% de variabilidad explicada como *ratio decidenci* del número de Coordenadas Principales (para nuestra distancia requeriríamos nueve), pues la práctica habitual es emplear dos o tres. En nuestro caso escogemos Y_1, Y_2, que explican respectivamente el 13.43% y 11.01% de la variabilidad de los datos. Excluimos el resto de Coordenadas Principales dado que la correlación que exhiben con los rasgos explicativos de nuestra matriz de datos es débil en prácticamente todos los casos (aportamos una prueba para Y_3 en la tabla).

Para la interpretación de las Coordenadas Principales nos serviremos de la relación entre éstas y las variables originales. Tenemos en cuenta a la hora de calcular correlaciones que existen diferentes tipos de variables. Así, emplearemos la correlación de Pearson para las variables cuantitativas, la V de Cramer para las binarias y la correlación de Spearman para el resto. La tabla con los valores correspondientes se muestra a continuación, así como la visualización en forma de mapa de calor.

Table 2: Correlaciones con Coordenadas Principales

Variables	Y_1	Y_2	Y_3
`distancia_ejes`	-0.6479	-0.2717	0.3626
`largo`	-0.6990	-0.1061	0.2298
`ancho`	-0.6521	-0.1696	0.3639
`altura`	-0.1269	-0.4378	-0.2436
`peso`	-0.8783	-0.0677	0.0569
`motor`	-0.6960	0.1272	0.2097
`caballos`	-0.7003	0.5172	-0.0019
`max_revoluciones`	0.2009	0.7209	-0.0396
`precio`	-0.8000	0.0980	0.1787
`combustible`	0.2350	0.5770	0.1500
`rueda_motriz`	0.7481	-0.1853	-0.0872

Nótese cómo la tercera Coordenada Principal apenas está correlacionada con las variables originales, de ahí que no se incluya en nuestro análisis para evitar cualquier interpretación desacertada o confusa. Procedemos con Y_1, Y_2 en adelante, constando que en lo sucesivo Y_2 denotará el eje de ordenadas para los comentarios sobre gráficos.

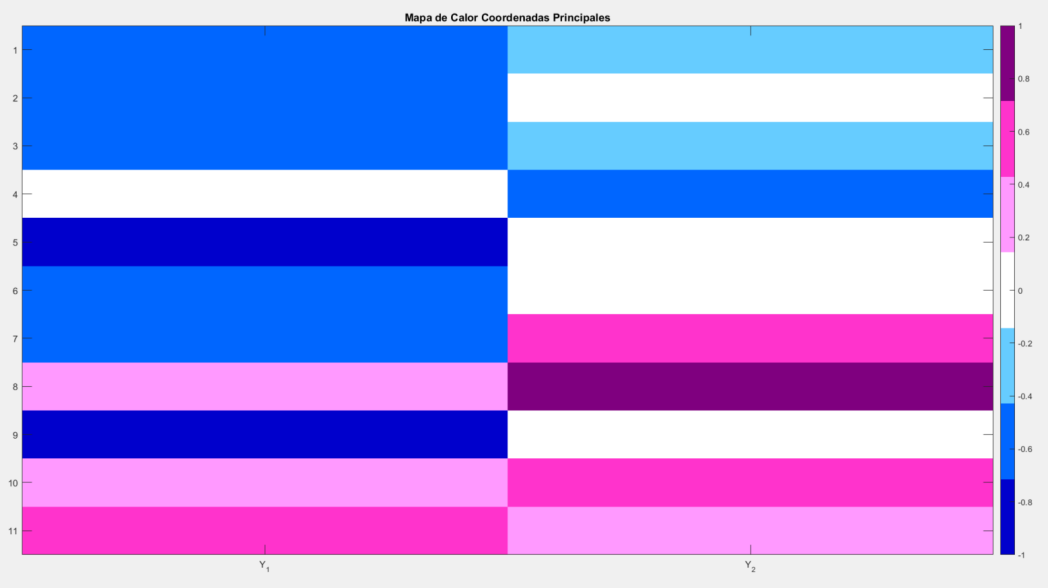

Figure 6: Correlación Coordenadas Principales y Variables

Las variables más correladas con Y_1 son (en orden): `peso`, `precio`, `rueda_motriz`, `caballos`, `largo`, `motor`, `ancho`, `distancia_ejes`, no jugando un papel relevante el resto (correlación por debajo de 0.3 en valor absoluto). El signo de la correlación es negativo en todos los casos salvo cuando se trata de la rueda motriz. En consecuencia, nos encontraremos a la izquierda del eje Y_1 del *gráfico de representación MDS* cuando el automóvil presente grandes dimensiones (coche pesado, ancho, con gran altura y distancia entre ejes), sea más caro y potente (muchos caballos de vapor y motor considerable). Podrán apreciarse más hacia la derecha los vehículos 4x4 y con tracción delantera. Por tanto es un índice de pequeñez y poca intensidad del coche, lo cual está asociado al precio.

En el caso de Y_2 las correlaciones son débiles menos para el caso de `max_revoluciones`. Le siguen `combustible`, `caballos` y `altura` (la única negativa de las mencionadas). En lo más alto del *gráfico de representación MDS* se situarán los coches de gran potencia (caballos y revoluciones altas) pero que no sobresalen en cuanto a la altura, premiando ligeramente los que usen gasolina. Podría entenderse como un índice de adecuación de los vehículos a la serie deportiva

Nótese que este tipo de coches presenta una gran potencia medida en caballos de vapor y revoluciones, así como una altura baja-media para incremental el aerodinamismo del vehículo (rasgos que identificamos en Y_2).

Ponemos en práctica la interpretación de las Coordenadas Principales que hemos sugerido de forma visual: presentamos una serie de gráficos donde se diferencian las variables de diversas formas (estudio de perfiles).

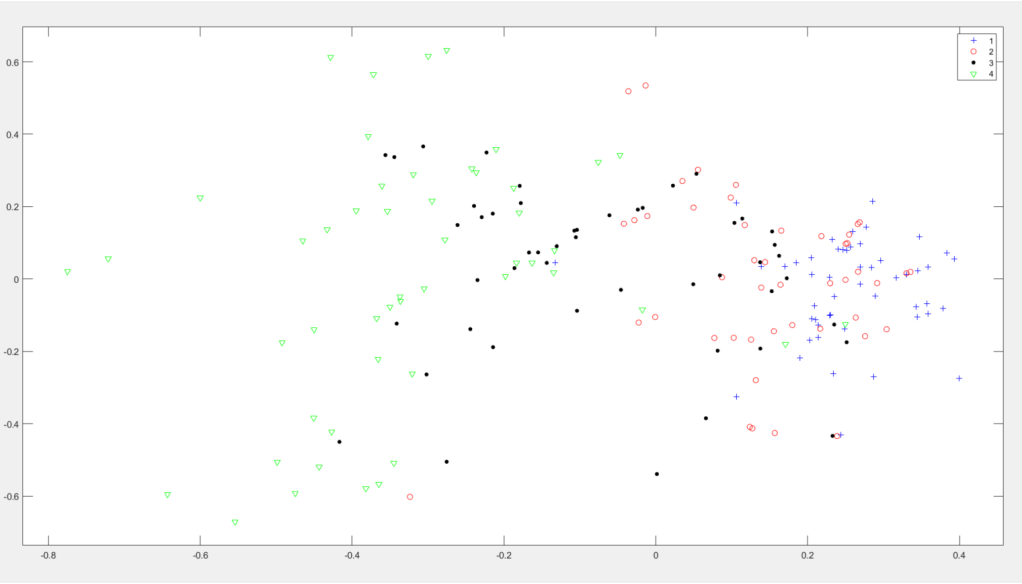

Figure 7: Precio por cuartiles

El gráfico de arriba agrupa los vehículos según su precio por cuartiles. Al objeto de que resulte más intuitivo, considérese el primer cuartil como modelos de gama baja de coches, el segundo cuartil como gama media-baja, tercer cuartil equivale a gama media-alta y último cuartil la gama alta. Recordemos a su vez que en el eje de abscisas, Y_1 o índice de pequeñez y poca intensidad, encontraremos automóviles de gran tamaño y vigor escorados a la izquierda, mientras que los coches más pequeños y de tecnología modesta estarán a la derecha. No resulta sorprendente que los coches de gama baja y media-baja (colores azul y rojo) presenten valores de Y_1 positivos. Asimismo, la gama alta (color verde) está significativamente presente en el extremo izquierdo del gráfico (valores negativos). Por último, la gama media-alta es más difícil de clasificar, mostrándose presente tanto en el lado negativo como positivo de los ejes (punto intermedio). En cuanto al eje Y_2, como ya comentamos, las observaciones con valores negativos están asociadas a vehículos distantes de encajar en el prototipo deportivo y que por ello son altos y poco potentes. De forma contraria, aquellos puntos

que observamos en el gráfico situados más hacia arriba, serán coches aerodinámicos y de gran intensidad. No es de extrañar que los automóviles con valores positivos considerables en el eje Y_2 asimismo se encuentran en el lado negativo de Y_1.

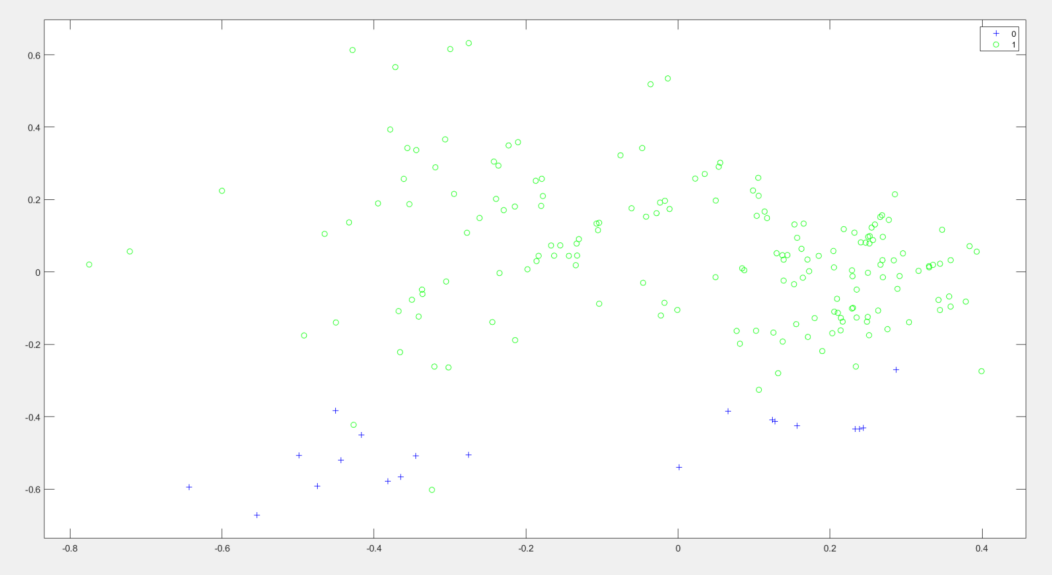

Figure 8: Representación MDS según combustible

Avanzando hacia el el análisis de las variables cualitativas, comenzamos por la parte binaria. En la Figura 8 se observa la representación MDS distinguiendo dos valores de `combustible`: gasolina (verde) y diésel (azul). La ligera mayoría de los automóviles diésel se caracterizan por su potencia y dimensiones, pues se localizan en la parte izquierda del eje Y_1, siendo de los valores más pequeños (indicativo de la capacidad del coche). Asimismo, los coches diésel se encuentran por debajo de los de gasolina en el eje Y_2, pues ya avanzamos que esta Coordenada Principal premiaba el uso de gasolina al presentar una asociación positiva. Cabe destacar que los automvviles con gasolina, al haber tantos y ser diversos, se encuentran dispersos a lo largo del eje Y_1, por lo que existen diferentes dimensiones y potencia. Es apreciable cómo el aerodinamismo es mayor en los coches de gasolina, consistente con el dato de que la mayoría de coches deportivos al tener tanta potencia, necesitan de una combustión directa y rápida, difícilmente compatible con la densidad del diésel (Auto10, 2018).

En cuanto a la variable multiestado, para los valores tracción trasera (coloz azul), tracción delantera (verde) y 4x4 (rojo) de `rueda_motriz`, observamos cómo los coches de tracción trasera son los más potentes y con mayor dimensión, seguidos de los 4x4 y por último los de las ruedas delanteras. Sobre el índice aerodinámico/deportivo o

Y_2 los todoterrenos no sobresalen como es esperable, los coches de tracción trasera son variados, así como los de tracción delantera. Señalamos como punto de mejora para éstos últimos la intensidad del motor y las revoluciones, existiendo espacio para mejorar en los aspectos ingenieriles y tecnológicos.

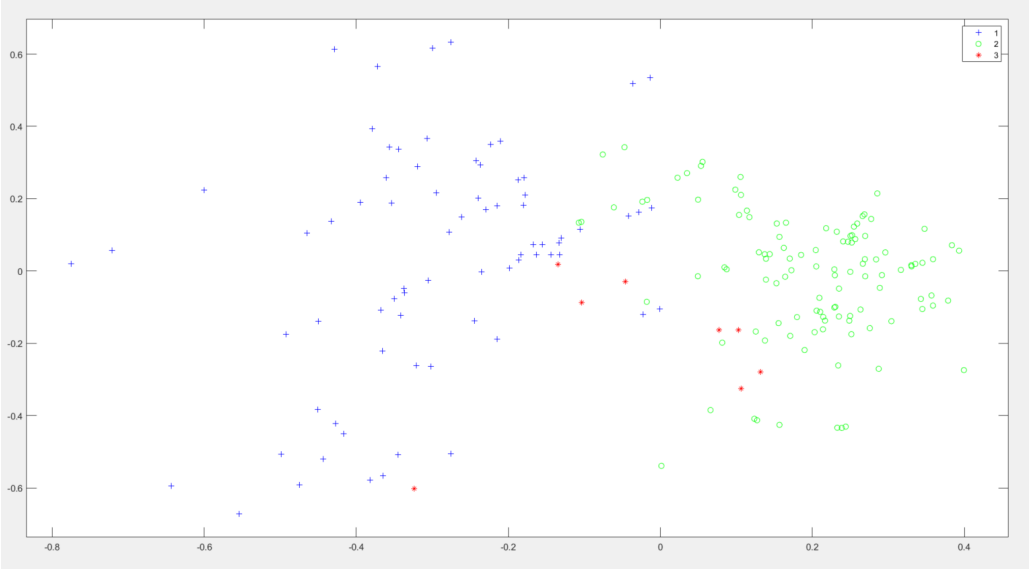

Figure 9: Representación MDS según tracción

Finalmente comentamos varios perfiles provenientes de dos variables cuantitativas, siendo la primera el peso del coche (Figura 9). Debido a que los cuatro grupos creados en el gráfico se identifican con los cuartiles, interprétese el color azul o primer cuartil como los coches ligeros, el colo rojo o segundo cuartil como automóviles de peso bajo-medio, el color negro o tercer cuartil como peso medio-alto y el color verde como vehículos pesados. El eje Y_1 presenta los coches pesados concentrados en el lado más negativo y los ligeros en el extremo derecho o valores más positivos. Resulta lógico que sea así pues Y_1 es un índice de pequeñez y el peso está positivamente asociado con la altura y ancho de los vehículos. En cuanto a los pesos intermedios, si bien es cierto que las observaciones atinentes al tercer cuartil están más proximas a los coches pesados, existe cierta dispersión que mezcla estos dos grupos. Concentrándonos en Y_2, resaltan los pesos medios como los más dinámicos, e incluso algunos coches pesados consiguen puntuar sorprendentemente bien en el índice de aerodinamismo (en parte se explica porque precisan de una mayor potencia para poder circular correctamente).

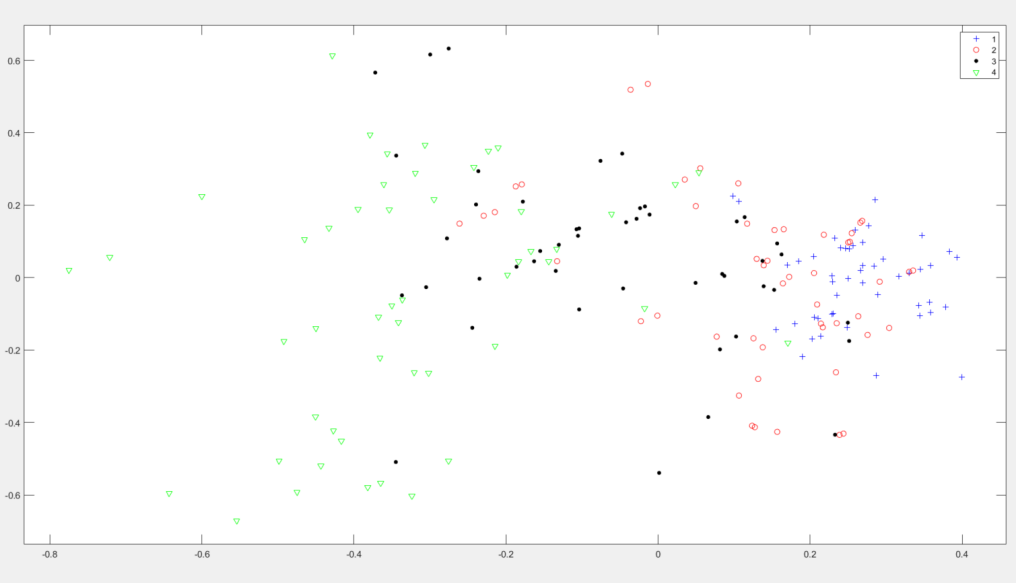

Figure 10: Representación MDS según peso

Concluyendo el proyecto con la variable motor: el color azul serían los motores más modestos, el rojo los mecanismos de automoción medios-pequeños, el color negro los motores medios-grandes y el color verde los de considerable dimensión. Observamos cómo los coches con motor grande lideran el índice de aerodinamismo (mayores valores en Y_2) ya que los deportivos cuentan con motores increíblemente potentes, requiriendo un mayor tamaño. Esto se traduce en muchos caballos de vapor y por ello encontramos estos mismos coches en los valores negativos del eje Y_1. Justamente lo contrario sucede con los automóviles con menor motor, y en cuanto a los que identificamos como intermedios experimentan una mezcla al encontrarse juntos.

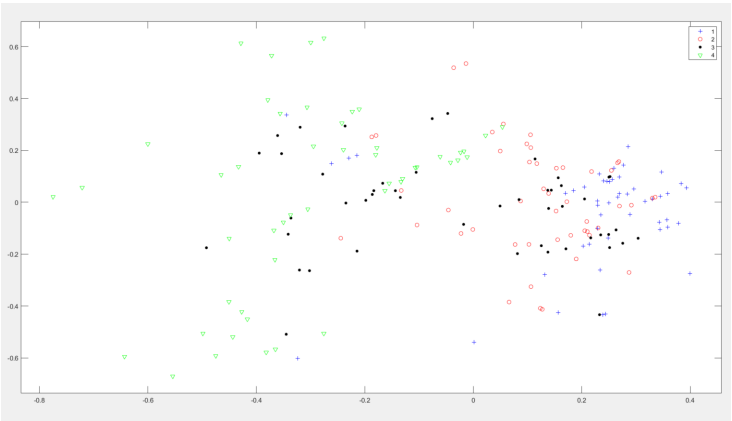

Figure 11: Representación MDS según motor

7 Conclusión

Mediante la aplicación de técnicas de reducción de la dimensionalidad hemos sido capaces de crear diferentes grupos o perfiles que resumen la información contenida en nuestra matriz de datos. Mientras PCA sólo es aplicable a las variables cuantitativas, MDS puede extenderse a todas. Concretamente hemos identificado que los índices más útiles han sido los de tamaño, modernidad y optimización de recursos (PCA) así como las dimensiones y fuerza del motor y aerodinamismo de los vehículos (MDS).

El principal objetivo ha sido mostrar el potencial de dichos métodos estadísticos, haciendo constar que su utilidad resaltaría aún más en contextos de bases de datos complejos. No obstante, ha de tenerse en cuenta que la interpretación de los resultados de PCA y MDS no es sencilla, recomendando en todo caso consultar con expertos cualquier duda o dificultad que se pueda encontrar para superar dicho obstáculo.

8 Apéndice

El código de MATLAB puede encontrarse en el siguiente repositorio de ***GitHub***, incluyendo las funciones que hemos modificado: `PCA()`, `correlacions`().

Las funciones no modificadas y que no son propias de MATLAB se atribuyen a Baíllo & Grané (2007).

9 Referencias

Auto 10 (2018). *Diez pros y contras de los Coche Gasolina.* ***Enlace al artículo***.

Baíllo, A. & Grané, A. (2007).*100 Problemas Resueltos de Estadística Multivariante.* Delta Publicaciones

Ballard, D.H. (1987). Modular Learning in Neural Networks. *Association for the Advancement of Artificial Intelligence, 1987 Conference.*

Cuadras, C. M. & Fortiana, J. (1995). A Continuous Metric Scaling Solution for a Random Variable. *Journal of Multivariate Analysis*

Gower, J.C. (1966). Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis. *Biometrika*

LeasePlan (2017). *Coche de Combustión vs Coche Eléctrico: ¿Cuál Gana?* ***Enlace al artículo***.

Motorpasion (2012). *El motor de combustión es el más eficiente hoy: FALSO* ***Enlace al artículo***.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*

Torgerson, W.S (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*

Young, G. & Householder, A. S. (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika*

Zuto (2020). *The Car Size Evolution.* Zuto Car Finance