UNIVERSIDAD NACIONAL DE COLOMBIA

Estudio de las técnicas de reducción de dimensión basadas en componentes principales: Análisis de componentes principales no lineales

Trabajo presentado como requisito para optar por el título de:
Magíster en ciencias - Matemática aplicada

Presentado por:
Juan David Giraldo Otálvaro

Directora:
PhD. Nubia Esteban Duarte
Profesora Departamento de Matemáticas y Estadística
Universidad Nacional de Colombia

Co-Directora:
PhD. Aymara Martínez Aragón
Profesora Departamento de Matemáticas y Estadística
Universidad Nacional de Colombia

Universidad Nacional de Colombia
Facultad de Ciencias Exactas y Naturales, Departamento de Matemáticas
Manizales, Colombia
2021

Agradecimientos

Después de un largo proceso de aprendizaje, disciplina y compromiso he llegado a este punto clave en el camino que me he propuesto para crecer no solo como profesional sino también como persona. Camino en el cual mis padres y hermanas siempre han estado a mi lado brindándome apoyo y cariño en todo momento, al igual que mis mejores amigos y primo quienes siempre han estado prestos a brindarme su mano y me han permitido estructurar prioridades a la hora de enfocar el camino correcto entre la academia y diferentes actividades por las que siento pasión. De igual forma, debo agradecer a mis tutoras y compañeros no solo por el conocimiento brindado sino también por tornarse en una segunda familia incondicional. Por último, agradezco a mi alma mater por brindarme la oportunidad de ejercer el hermoso arte de la docencia y a todos aquellos estudiantes con los que pude compartir no solo conocimiento sino también invaluables recuerdos y experiencias.

Resumen

Estudio de las técnicas de reducción de dimensión basadas en componentes principales: Análisis de componentes principales no lineales.

En la estadística multivariada un gran desafío en el manejo correcto de grandes cantidades de datos es el análisis de variables de carácter cuantitativo y cualitativo al mismo tiempo, es decir, análisis de datos mixtos. En lo relacionado al tratamiento de datos solamente cuantitativos existen varias técnicas que ayudan en la reducción de la dimensión, en donde el Análisis de Componentes Principales (PCA) es la metodología de mayor relevancia. Para el análisis de datos mixtos, la técnica de Análisis de Componentes Principales proporciona una base fundamental para otras técnicas multivariadas como lo es el Análisis de Componentes Principales No Lineales (NLPCA), la cual no está muy bien documentada y tal vez aplicada sin la rigurosidad que la teoría requiere. Por otro lado, su uso no ha sido extendido a la metodología de las cartas de control como herramienta que apoya la gestión de calidad desde un punto de vista analítico.

Por lo anterior, en este trabajo se describe de forma teórica la metodología de Análisis de Componentes Principales y se formaliza una técnica que permita el procesamiento de datos mixtos con el fin de facilitar la reducción de dimensión bajo el marco del PCA seleccionando la técnica de Análisis de Componentes Principales No Lineales (NLPCA), la cual incluye en su procesamiento la cuantificación óptima de datos cualitativos de manera no lineal con el fin de encontrar las mejores relaciones entre las variables. Se propone adaptar las cartas de control desarrolladas para variables múltiples y componentes obtenidas a partir del PCA, a las técnicas NLPCA obteniendo herramientas novedosas de gran interés para la interpretación de datos. Las metodologías descritas son aplicadas a un conjunto de datos reales pertenecientes al Proyecto "Corazones de Baependi" (Processo Fapesp 2007/58150-7) del Laboratorio de Genética y Cardiología Molecular

(Incor/USP).

Palabras Clave: Componentes Principales, Componentes Principales no Lineales, Escalamiento Óptimo, Análisis de Homogeneidad, Mínimos Cuadrados Alternantes.

Abstract

Study of dimension reduction techniques based on Principal Components: Non-linear Principal Components Analysis

In multivariate statistics, a great challenge in the correct handling of large amounts of data is the analysis of variables of a quantitative and qualitative nature at the same time, that is, analysis of mixed data. Regarding the treatment of only quantitative data, there are several techniques that help in dimensional reduction, where the Principal Component Analysis (PCA) is the most relevant methodology. For the analysis of mixed data, the Principal Component Analysis technique provides a fundamental basis for other multivariate techniques such as Nonlinear Principal Component Analysis (NLPCA), which is not very well documented and perhaps applied without rigor. that the theory requires. On the other hand, its use has not been extended to the control chart methodology as a tool that supports quality management from an analytical point of view.

Due to the above, in this work the Principal Component Analysis methodology is described theoretically and a technique is formalized that allows the processing of mixed data in order to facilitate the reduction of dimensions under the framework of the PCA by selecting the technique Non-linear Principal Components Analysis (NLPCA), which includes in its processing the optimal quantification of qualitative data in a non-linear way in order to find the best relationships between the variables.

It is proposed to adapt the control charts developed for multiple variables and components obtained from the PCA, to the NLPCA techniques, obtaining novel tools of great interest for data interpretation.

The methodologies described are applied to a set of real data belonging to the Project "Hearts of Baependi "(Processo Fapesp 2007 / 58150-7) of the Molecular Genetics and Cardiology Laboratory (Incor / USP).

Key words: Principal Components Analyisis, Nonlinear Principal Components, Optimal Scaling, Homogeinity Analysis, Alternating Least Squares.

Índice general

Índice de tablas

Índice de figuras

Capítulo 1

Introducción

1.1. Introducción

En muchas situaciones, un analista tiene como objetivo entender el comportamiento de bases de datos de altísima dimensión a través de una base de datos de baja dimensión sin pérdida considerable de información, o en otros casos simplemente necesita retirar la variabilidad que pueda ser considerada como ruido de los datos. Para ello, entre las diversas herramientas que se tienen a disposición en el análisis multivariado, las técnicas de reducción de dimensión son las apropiadas para cumplir esta tarea, por esta razón, se han constituido como un área primordial en el análisis multivariado (Bishop, 2006; Johnson & Wichern, 2007; Diaz & Morales, 2012).

En este contexto, una herramienta muy conocida y utilizada para diversos análisis es el análisis de componentes principales (PCA - Principal Component Analysis), que al igual que otras técnicas como el escalamiento multidimensional (MDS - Multidimensional Scaling), proyectan los datos en un espacio de menor dimensión. Sin embargo, estas técnicas únicamente pueden identificar las relaciones de los datos de una forma lineal y trabajar con datos netamente cuantitativos en el caso del PCA, o cualitativos si se trata del MDS no métrico (Lee & Verleysen, 2007). De forma que si existen relaciones no lineales entre los datos, la representación obtenida de ellos no será la más adecuada y en el caso de incluir variables cualitativas éstas no podrán ser analizadas junto a las demás variables. Teniendo presente las limitaciones mencionadas de las herramientas clásicas de reducción de dimensión, se decide explorar el desarrollo y comportamiento de una herramienta alternativa que permita tanto el procesamiento de datos de carácter mixto, es decir, el

análisis conjunto de variables cuantitativas y cualitativas, y que, además, cuente con la capacidad de analizar de una forma no solamente lineal las relaciones existentes entre las variables, a saber el análisis de componentes principales no lineales o NLPCA (Nonlinear Principal Components Analysis), que será el foco principal del presente trabajo.

El análisis de componentes principales no lineales, NLPCA es una metodología estadística multivariada basada en la filosofía del PCA que permite un tratamiento robusto de datos cualitativos, combina en su proceso escalamiento óptimo (OS - Optimal Scaling) y un algoritmo de mínimos cuadrados alternantes (ALS - Alternating Least Squares) para escalar las variables cualitativas, tanto nominales como ordinales, a medida que se encuentra la proyección de los datos en un nuevo subespacio de componentes (Gifi, 1990; Michailidis & De Leeuw, 1998; Mair, 2018).

1.2. Estado del arte

En los últimos años el análisis de datos multivariados ha tomado gran fuerza en todos los ámbitos profesionales y económicos de la sociedad. Esto viene de la mano de un gran desarrollo tecnológico que ha permitido la captación de grandes volúmenes de datos de los cuales es necesario obtener información de forma rápida, eficiente y confiable. Corrientes teóricas recientes se han llamado la revolución industrial 4.0 y 5.0 debido al impacto que ha generado para las empresas el automatizar la adquisición, controlar la calidad y generar valor a partir de los datos que se recolectan y asumir los retos tanto teóricos como prácticos que esto conlleva. Por esta razón, áreas como la informática, la estadística y la matemática han tomado un gran lugar en el desarrollo de los diferentes sectores económicos y científicos de la actualidad (Santos et al. 2017; Quiroga et al., 2017; Rodríguez, 2017; Özdemir & Hekim, 2018; Arévalo et al., 2018).

Por su parte, la estadística es una herramienta que sigue dando soporte a las nuevas metodologías matemáticas y computacionales para dar solución a la necesidad de obtener información útil proveniente de los grandes volúmenes de datos, en tareas como la predicción de comportamientos, identificación de patrones o en la visualización de la información contenida en ellos. Una actividad de gran importancia en la actualidad es la visualización de datos de alta dimensionalidad y su reducción de dimensión en un conjunto de variables representativas (Bishop, 2006; Johnson & Wichern, 2007; Lee & Verleysen, 2007; Diaz & Morales, 2012; Gerón, 2019). Sin embargo, en la captura de altos volúmenes de datos es muy probable que se presenten mezclas de datos de tipo cualitativo y cuantitativo, lo cual puede presentar un inconveniente a la hora de realizar el análisis.

Como se describió, se hacen necesarias herramientas potentes de análisis de datos mixtos ya que en la práctica es común encontrar que las técnicas de reducción de dimensión implementadas en paquetes de análisis estadístico y procesamiento de datos, en su mayoría, tengan una predominancia hacia metodologías cuantitativas. Técnicas de corte cualitativo, como el análisis de correspondencias múltiples no son utilizadas en una gran cantidad de áreas del conocimiento ya que los datos cualitativos, al no ser numéricos, no se procesan de forma sencilla por técnicas netamente matemáticas y, por lo tanto, se debe tener especial cuidado a la hora de ser analizados junto a variables numéricas, es decir, con las bases de datos mixtas (Laub & Sampson, 1998; Almalki, 2016).

Ilustrando lo anotado en el párrafo anterior, se tiene el Análisis de Correspondencias Múltiples (MCA - Multiple Correspondence Analysis), una de las técnicas de análisis multivariado de variables nominales más importantes en el área de la estadística multivariada. Para su implementación es necesario el uso de tablas de contingencia construidas a partir de las frecuencias absolutas o relativas de los individuos que toman valores en cada uno de los respectivos niveles de la variable y se centra en obtener la mejor representación para dos conjuntos de datos, los dispuestos en las filas o en las columnas de la respectiva matriz de datos. Resumiendo, se centra en obtener un número representativo de dimensiones (factores), de manera que la primera dimensión explique la mayor parte de la asociación total entre filas y columnas, el segundo la mayor cantidad del residuo de la asociación no explicada por el primer factor y así sucesivamente, siguiendo la filosofía de un PCA clásico, para conservar la mayor cantidad de variabilidad posible de los datos en las primeras dimensiones de las componentes (Lebart et al., 1984; Greenacre & Blasius, 2006; Abdi & Valentin, 2007; Johnson & Wichern, 2007; Diaz & Morales, 2012; Di Franco, 2016). El uso de las tablas de contingencia implica un alto consumo de memoria a la hora de realizar los cálculos computacionales que requiere la técnica. Además, la técnica, al igual que el PCA, basa su paso de proyección de datos sobre una descomposición espectral, lo que conlleva a que la captura de la información de la asociación de las variables, en principio, se realice de forma lineal.

Para encontrar la solución a los problemas expuestos anteriormente, y otros no mencionados, metodologías recientes, principalmente del área de la ciencia de datos, han optado por el uso de codificadores nominales u ordinales que eviten el uso de técnicas como el MCA en grandes volúmenes de datos y en su lugar se opten por técnicas de carácter netamente cuantitativo para procesar distintos conjuntos de variables sin importar su naturaleza. Entonces, una alternativa es utilizar técnicas de carácter cuantitativo para la reducción de dimensión apoyándose en el uso de codificadores de carácter nominal como el OneHot Enconder y como el Ordinal Encoder. El primer método consiste en la asignación de una variable dummy a cada una de las categorías de las variables nominales presentes en la muestra y trabajar con una nueva matriz de ceros y unos en lugar de el vector original de datos. Por su parte, el Ordinal Encoder codifica cada categoría de las variables ordinales en orden jerárquico sobre los números naturales. Siendo así la categoría jerárquicamente más baja asignada con el número uno, la siguiente el número dos y así sucesivamente. Ambas técnicas han tenido muy buena

acogida en la ciencia de datos, a pesar de las implicaciones que puede tener el trabajar con variables que solo ocupen los valores 0 y 1 como si fueran numéricas o suponer que la distancia numérica entre las categorías de una variable ordinal es siempre la misma, ya que las técnicas de reducción de dimensión no lineal presentan una alta capacidad para hacer frente a este tipo de problemas (Bishop, 2006; Choong & Lee, 2017; Potdar, Pardawala & Pai, 2017; Gerón, 2019).

Otro aspecto que ha ganado mayor relevancia en las últimas décadas para los métodos de reducción de dimensión ha sido la capacidad de captar relaciones no lineales entre los datos. Su gran importancia proviene del hecho de que la gran mayoría de los conjuntos de datos no guardan una relación lineal entre sus variables y, por lo tanto, las transformaciones lineales de los datos realizadas por los métodos clásicos no capturan su verdadera naturaleza (Tenenbaum, De Silva & Langford, 2000; Lee & Verleysen, 2007; Venna, et al., 2010). Esta relación entre las variables es una de las principales características de las técnicas de reducción de dimensión la cual, junto a otras características, se asocia con el modelo utilizado y es clave a la hora de seleccionar qué técnica es adecuada para cada uno de los escenarios que se plantean en los diferentes tipos de problemas. Las tres características más importantes de los modelos de reducción de dimensión son:

- El tipo de modelo asumido: Los métodos pueden asumir que las relaciones entre los datos pueden ser de carácter lineal o no lineal. Por ejemplo, métodos clásicos como el PCA asumen que las relaciones entre los datos son lineales. Sin embargo, esta es una característica muy poco potente para la descripción de las relaciones entre los datos, de manera que muchas otras técnicas de carácter no lineal como el Kernel PCA, el Isomap, entre otros, se han desarrollado en pro de subsanar esta debilidad (Lee & Verleysen, 2007).

- El tipo de algoritmo utilizado: Los métodos pueden tener solución de carácter cerrado al problema de optimización que se plantea por medio de procesos algebraicos simples. Sin embargo, en muchos casos es necesario el uso de diferentes técnicas de optimización que se basan en procesos iterativos con el fin de encontrar el óptimo global del problema planteado a pesar de que puede representar un alto costo computacional (Lee & Verleysen, 2007).

- El criterio de optimización definido: La selección del criterio de optimización a menudo determina cuáles funcionalidades ofrecerá el método,

interviene en el modelo de los datos y siempre orienta la implementación de un tipo de algoritmo en particular. Típicamente, el criterio a ser optimizado es escrito como una fórmula matemática. Uno de los criterios más comúnmente usados es el error de reconstrucción que compara el error entre los valores medidos y la inversa de los valores transformados. Otros criterios como la conservación de la varianza inicial observada en los datos proyectados son típicamente utilizados. Desde un punto de vista más geométrico y topológico, la proyección del objeto debe preservar su estructura, por ejemplo, conservando las distancias por pares, medidas entre las observaciones en el conjunto de datos. Si el objetivo es la separación de variables latentes, el criterio puede ser la no correlación. Este criterio puede enriquecerse aún más haciendo que las variables latentes estimadas sean lo más independientes posible (Lee & Verleysen, 2007).

Para ilustrar algunos de los diferentes caminos que pueden tomar las técnicas de reducción de dimensión en el ámbito no lineal se hace un breve recuento de diferentes métodos, iniciando por el escalamiento multidimensional al ser uno de los métodos lineales clásicos más importantes y pasando por diversos métodos de carácter no lineal basados en esta técnica y el análisis de componentes principales lineal, en este caso el análisis de componentes principales se estudiará en el Capítulo 2, de forma que no se incluirá en esta instancia.

- El escalamiento multidimensional (MDS - multidimensional scaling) es actualmente una familia de métodos más que un simple proceso bien definido. El escalamiento se refiere a métodos que construyen una configuración de puntos en un espacio métrico objetivo a partir de información sobre distancias entre puntos, y permite el escalado cuando el espacio objetivo es euclidiano. Su construcción es realizada sobre supuestos estrictamente lineales. Ha sido ampliamente utilizado y desarrollado en ciencias humanas como sociología, antropología, economía y también particularmente en un subcampo de la psicología llamado psicometría. En este último dominio, el MDS se utiliza como herramienta para la representación geométrica de conceptos, objetos u opiniones. Su método de optimización es exacto y puramente algebraico ya que su solución óptima se obtiene en forma cerrada. Es un método espectral, ya que la operación principal en su procedimiento se obtiene una descomposición en valores singulares de una matriz de Gram (Young & Householder, 1938; Torgerson, 1952; Shepard, 1962;

Kruskal, 1964; de Leeuw & Heiser, 1982; Lee & Verleysen, 2007;Cox, M. A., & Cox, T. F. 2008; Hout, Papesh & Goldinger, 2013; Ghojogh et al., 2020).

- El Isomap es un método que trabaja con una optimización algebraica exacta. Opera como el MDS métrico al descomponer una matriz de Gram en autovalores y autovectores, por lo tanto, a menudo se califica como un método espectral. Es el método de reducción no lineal más simple, para lograrlo la técnica usa la distancia gráfica como una aproximación de la distancia geodésica. La distancia geodésica es una medida entre puntos a lo largo de una variedad, recibe este nombre por analogía con las curvas dibujadas sobre la superficie de la tierra y se diferencia en gran medida de las distancias euclidianas ya que estas miden en principio la menor distancia posible entre los datos, tomando atajos como los aviones en sus trayectorias sin importar la existencia de carreteras. Por su parte, la distancia gráfica depende menos de la proyección particular de los datos en el espacio. Es decir, la curvatura del espacio no modifica (o modifica fuertemente) el valor de la distancia medida. La única diferencia con el MDS métrico es la métrica usada para medir la distancia entre pares de puntos, el Isomap usa la distancia gráfica en vez de la distancia euclidiana en el proceso algebraico del MDS métrico. Solo por la introducción de la distancia gráfica, el MDS métrico puramente lineal se torna en un método no lineal. De forma que estas capacidades no lineales se deben exclusivamente al uso de la distancia gráfica; otras partes del método, como el modelo subyacente del procedimiento de optimización, se derivan del MDS métrico clásico y siguen siendo puramente lineales (Tenenbaum, 1998; Tenenbaum, De Silva, & Langford, 2000; Balasubramanian et al., 2002; Lee & Verleysen, 2007; Zhang, Chow & Zhao, 2012; Du, 2019; Ghojogh et al., 2020).

- El Kernel PCA (KPCA) es una técnica de reducción de dimensión más cercana al clásico MDS métrico que el PCA. Más allá de la equivalencia entre el PCA y el MDS métrico clásico, se puede justificar la selección del nombre de PCA por el hecho de que es más conocido en el campo en el que el KPCA ha sido desarrollado. El KPCA extiende las propiedades del MDS a variedades no lineales, sin importar su significado geométrico. De hecho, en este sentido el Kernel PCA presenta una gran semejanza con el Isomap ya que ambos generalizan el MDS métrico a variedades no lineales de manera similar, aunque

las ideas subyacentes difieren completamente. El modelo del método es no lineal gracias a la función kernel que mapea los datos en un camino explícito. Por construcción, el KPCA comparte muchas ventajas y desventajas con el PCA y el MDS métrico. En contraste con esos métodos, el KPCA puede tratar con variedades no lineales. Sin embargo, el KPCA no es utilizado a menudo como técnica de reducción de dimensión. Las razones son que el método no está motivado por argumentos geométricos y, por lo tanto, la interpretación geométrica de varios kernels resulta difícil (Schölkopf, Smola Müller, 1998; Williams, 2001; Williams, 2002; Shawe-Taylor et al., 2005; Lee & Verleysen, 2007; Hoffmann, 2007; Sriperumbudur, & Sterge, 2017; Datta, Ghosh, S. & Ghosh, A., 2018).

- El mapeo no lineal de Sammon (Sammon's nonlinear mapping) establece un mapeo entre un espacio altamente dimensional y uno de baja dimensión. El método Sammon no determina exactamente un mapeo continuo entre dos espacios ya que su propósito es reducir la dimensionalidad a un conjunto finito de puntos. Es un método no lineal y discreto. Su proceso de optimización es aproximado y puede llegar a un mínimo local. No incluye un estimador de dimensionalidad intrínseca y la dimensionalidad de la proyección es asignada por el usuario. Embeddings incrementales o por capas no son posibles para el método, por lo tanto, deberá ser ejecutado separadamente para cada dimensionalidad específica (Sammon, 1969; Pekalska et al., 1999; Yang, 2004; Lee & Verleysen, 2007; Du, 2019; Ghojogh et al., 2020).

- Los mapas auto-organizados (SOM - Self-Organizing Maps) es quizás el método más conocido en el campo de las redes neuronales artificiales junto al perceptrón multicapa (MLP - multi-layer perceptron). Se centra en la combinación de dos subtareas concurrentes: cuantificación vectorial y representación topográfica (es decir, reducción de dimensión). Por lo general, los SOMs son implementados como algoritmos fuera de línea. Sin embargo, versiones en línea pueden ser fácilmente derivadas (Von der Malsburg, 1973; Rumelhart, Hinton & Williams, 1985; Rumelhart, Hinton, & Williams, 1986; Hinton, 1986; Lee & Verleysen, 2007; Werbos, 2008; Minsky & Papert, 2017; Miljković, 2017).

- El embedding localmente lineal (LLE - Locally linear embedding) propone una aproximación basada en mapeos conformables. Un mapeo conformal es una transformación que conserva los ángulos locales. Es un método por lotes fuera de línea que trabaja con operaciones al-

gebraicas simples. Como muchos otros métodos espectrales el LLE es capaz de producir proyecciones incrementalmente por medio de añadir o retirar valores propios. El mapeo provisto por el LLE es explícito y discreto y a diferencia de métodos como MDS asume que los datos son localmente lineales, no globalmente. Además, el LLE asume que las variedades pueden ser mapeadas a un plano usando un mapeo conformal. A pesar de que tanto el MDS como el LLE usan una descomposición espectral, que es puramente lineal, las capacidades no lineales del LLE vienen de su primer paso, el cálculo de los vecinos más cercanos (Roweis & Saul, 2000; Saul & Roweis, 2003; Lee & Verleysen, 2007; Chen, & Liu, 2011; Du, 2019).

- Los eigenmaps Laplacianos (LE - Laplacian Eigenmaps) pertenecen a la familia de técnicas de reducción no lineal basadas en descomposición espectral. El método se enfoca en remediar algunas deficiencias de otros métodos espectrales como el Isomap y el LLE. A diferencia del Isomap, el LE desarrolla un enfoque local para el problema de la reducción de dimensión no lineal. En ese sentido, el LE está estrechamente relacionado con el LLE, aunque aborda el problema de una manera diferente, en lugar de reproducir pequeños parches lineales alrededor de cada dato, el LE se basa en conceptos teóricos de grafos como el operador laplaciano. El método busca la minimización de distancias locales, es decir, distancias entre puntos de datos vecinos y para evitar la solución trivial en la que todos los puntos se asignan a un solo punto, la minimización es restringida (Belkin, & Niyogi, 2001; Belkin & Niyogi, 2003; Lee & Verleysen, 2007; Carreira-Perpinán & Lu, 2007; Li, B., Li, Y. R. & Zhang, 2019; Du, 2019).

- Los autocodificadores (Autoencoders) tradicionales basados en redes neuronales consisten de dos redes neuronales artificiales apiladas una encima de otra. La red codificadora es responsable por codificar la entrada de los datos en algunas variables latentes. La red decodificadora es usada para decodificar esos parámetros en orden de retornar en una recreación precisa de los datos de entrada. Los parámetros de este algoritmo son entrenados por medio de unas actualizaciones realizadas por un gradiente descendiente en pro de minimizar la pérdida de reconstrucción entre los datos de entrada y salida (Ng, 2011; Wetzel, 2017; Tschannen, Bachem, & Lucic, 2018).

- Los autocodificadores variacionales (Variational Autoencoders) son una versión moderna de los autocodificadores que imponen restric-

> ciones adicionales a las representaciones codificadas. Estas restricciones transforman los autocodificadores en un algoritmo que aprende un modelo de variable latente de sus datos de entrada. Mientras las redes neuronales de autocodificadores tradicionales aprenden una función arbitraria para codificar y decodificar los datos de entrada, los autocodificadores variacionales aprenden los parámetros de una distribución de probabilidad modelada de los datos. Después de aprender la distribución de probabilidad, uno puede muestrear parámetros de él y luego dejar que la red del codificador genera muestras muy parecidas a los datos de entrenamiento. Para lograr esto, los autocodificadores variacionales emplean el supuesto de que uno puede muestrear los datos de entrada de una distribución gaussiana de parámetros latentes. Los pesos del modelo son entrenados simultáneamente optimizando dos funciones de pérdida, una pérdida de reconstrucción y la divergencia Kullback-Leibler entre la distribución latente aprendida y una gaussiana unitaria previa (Doersch, 2016; Wetzel, 2017; Kingma & Welling, 2019; Khemakhem, et al., 2020).

Otras técnicas de reducción de dimensión con un enfoque no lineal importantes son t-SNE, Maximum Variance Unfolding, Gaussian process latent variable models, Probabilistic nonlinear PCA, Kernel entropy component analysis, para más información sobre estas y las técnicas abordadas anteriormente consultar Lawrence (2003), Lawrence & Hyvärinen (2005), Weinberger & Saul (2006), Van der Maaten & Hinton (2008), Jenssen (2009), Scholz (2012), Wattenberg, Viégas & Johnson (2016), Li & Chen (2016) y Waagen, et al., (2021). Vale resaltar que de las técnicas anteriormente expuestas, las utilizadas principalmente bajo una naturaleza explícitamente no métricas son el MCA y el MDS. Para el caso de las demás técnicas es necesario incluir como uno de los pasos del proceso de análisis el uso de codificadores como OneHot Encoder u Ordinal Encoder.

Desprendidas de la filosofía del PCA lineal existe en la literatura dos técnicas con un cierto nivel de capacidad no lineal y una metodología implícita en el procesamiento de variables cualitativas. Estas técnicas son conocidas como PCA mixto (PCA Mix) y PCA no lineal (NLPCA - Nonlineal PCA), ambos se basan en algoritmos de optimización iterativa con un paso de cuantificación de variables cualitativas y otro paso de descomposición espectral. La mayor diferencia entre ellas radica en el que el PCA Mix centra su metodología de procesamiento de variables cualitativas en un MCA, teniendo mayor sentido para variables de carácter nominal, mientras que el NLPCA

centra su metodología de cuantificación en la definición de un problema de optimización que mejor ajuste la cuantificación de las variables a su naturaleza. Como fue anotado anteriormente, este proceso de cuantificación es conocido como escalamiento óptimo (OS) y junto a la metodología del análisis de homogeneidad por mínimos cuadrados alternantes (HOMALS - Homogeneity Analysis For Alternating Least Squares) permite la transformación no solo de variables nominales sino también ordinales (De Leeuw & Van Rijckevorsel, 1980; Gifi, 1990; Meulman, 1998; Michailidis & De Leeuw, 1998; Meulman et al., 2004; Linting et al., 2007; Lintning & van der Kooij, 2012; Mair, 2018).

Como ya fue descrito, este trabajo se centra en el PCA lineal y en el NLPCA o PCA no lineal. Dentro de este contexto, una motivación práctica, es la utilización de técnicas de reducción de dimensión en el área de control de calidad. Principalmente, por el uso de técnicas de carácter no lineal que permitan el manejo de datos mixtos. Estas herramientas son comúnmente conocidas como cartas de control estadístico y permiten identificar comportamientos erráticos en la variabilidad de las medidas realizadas sobre los productos a controlar. Normalmente, las cartas de control se enfocan en el análisis de variables, haciendo referencia a variables cuantitativas, atributos y variables cualitativas. Sin embargo, son pocos los trabajos que se han enfocado en unir estos dos tipos de variables en pro de generar cartas de control de carácter mixto.

En el contexto de cartas de control, Tuerhong & Kim (2014) proponen una carta de control multivariada basada en la distancia de Gower que puede utilizar una mezcla de datos numéricos y datos categóricos. Ahsan, Mashuri, et al., (2018) proponen una nueva metodología basada en el PCA Mix unida a una estimación de densidad de kernel (Kernel Density Estimation) para estimar los límites de control. El PCA Mix es un método de procesamiento multivariado mixto que fundamenta su principio en el procesamiento de datos en una PCA para datos cuantitativos y en un MCA para datos cualitativos de carácter categórico uniendo ambos en un proceso iterativo de optimización. Jin & Loosveldt (2019) presentan una aproximación similar a la anterior carta de control incluyendo indicadores de carácter cualitativo junto al análisis de indicadores de carácter cuantitativo construyendo un estadístico de Hotelling T^2 a partir de una transformación previa de los datos por medio de un PCA Mix. Posteriormente, siguiendo la naturaleza no normal de las componentes obtenidas a partir del PCA Mix, se utiliza un método de bootstrap no paramétrico para obtener el límite para el estadísti-

co T^2.

Por lo anterior, se hace necesario contemplar nuevas metodologías que permitan la inclusión de variables de carácter ordinal dentro de las metodologías de control de calidad, como también el uso de metodologías con una fuerte capacidad no lineal para detectar las relaciones entre los datos y que su poder para identificar fluctuaciones en la variabilidad de los datos sea mayor.

1.2.1. Justificación

En la literatura existen varias investigaciones en las que en sus análisis utilizan análisis de componentes principales, pero no se han hecho estudios en donde se fundamente la teoría de componentes principales no lineales y se realicen aplicaciones a la luz la teoría que incorpora el análisis de datos mixtos, presentes en muchas investigaciones. Siendo así, se propone plantear una fundamentación matemática sólida de la metodología NLPCA basada en el análisis PCA. En la misma dirección, la selección de la técnica a estudiar es impulsada al identificar la falta de desarrollo en el campo del control de la calidad sobre cartas de control estadístico que permitan a la vez el control de características de calidad de carácter numérico, nominal y ordinal y que, además, cuenten con un cierto grado de capacidad no lineal en la captura de las relaciones entre las variables estudiadas. Algunas técnicas influenciadas fuertemente en la filosofía de las metodologías clásicas como el PCA han tomado lugar en un pequeño nicho de programas de análisis de datos comerciales como técnicas multivariadas de carácter mixtas. Una de estas son los ya nombradas PCA Mix y NLPCA, y es esta última respecto a la cuál tomará lugar el presente estudio. Esta técnica, además de permitir analizar datos de carácter cualitativo en conjunto con variables numéricas, captura información de las relaciones no lineales entre los datos por medio de la transformación óptima y controlada de las variables nominales y ordinales.

De esta manera el NLPCA se utilizará como técnica de acercamiento a la propuesta de construcción de cartas de control de carácter mixto con la inclusión de variables ordinales. Sin embargo, el problema aquí radica en que, a pesar de que el NLPCA se encuentra implementado en paquetes estadísticos como SAS, por medio de la técnica PRINQUAL, en SPSS, por la técnica PRINCIPALS, y en R, en el paquete GIFI, la documentación teórica es escasa (Lintning & van der Kooij, 2012). Su marco teórico se desarrolla en torno a publicaciones alineadas con el sistema Gifi que toman, por lo general, la técnica HOMALS como punto de partida para introducir el NLPCA, generando dificultades en la conexión de ideas y características con el PCA lineal (Gifi, 1990; Michailidis & De Leeuw, 1998; Mair, 2018). Además, la notación empleada es dispersa y diferente en cada publicación o paquete, quedando difícil el estudio y comprensión de estos temas. Por esto se ve la necesidad de estudiar el NLPCA a la luz de una técnica previamente conocida, y utilizada ampliamente en la construcción de cartas de control multivariadas, como lo es el PCA desde un contexto estadístico, matemático y algorítmico, que facilite la comprensión teórica y práctica de ambas técnicas.

Para ilustrar las metodologías abordadas en el presente trabajo, existen datos reales de investigaciones que dan la oportunidad de fundamentar la teoría y tener conclusiones útiles de la investigación. En este aspecto, específicamente, se utilizarán datos del Proyecto "Corazones de Baependi" (Processo Fapesp 2007/58150-7) del laboratorio de Genética y Cardiología Molecular (Incor/USP), que tiene varios frentes de estudio, entre ellos el estudio de determinantes genéticos que regulan enfermedades cardiovasculares (Oliveira et al., 2008, Giolo et al., 2009). Las variables que se utilizaran en este estudio están asociadas a enfermedades cardiovasculares, medidas antropométricas e índices de carácter psicosocial.

1.3. Objetivos

1.3.1. Objetivo general

Establecer una relación directa entre la teoría de educción de dimensión mixta NLPCA y su homónimo cuantitativo PCA.

1.3.2. Objetivos específicos

- Presentar una base teórica para el estudio de NLPCA fundamentada en la teoría existente de PCA.
- Utilizar el lenguaje de programación Python y el Programa estadístico SPSS para realizar las aplicaciones que ilustran la teoría formalizada en un escenario de interés práctico.
- Proponer una metodología para la construcción de cartas de control multivariadas mixtas que permitan el control de variables ordinales.

1.3.3. Estructura del documento

El documento está organizado de la siguiente forma:

- En el Capítulo 2, se presentarán los principios del PCA desde un punto de vista matemático y estadístico junto a la construcción de cartas de control.
- En el Capítulo 3, se iniciará el estudio del NLPCA partiendo por el OS y ALS, para posteriormente abordar el problema de reducción de dimensión para datos mixtos por medio del NLPCA y finalizar la sección con una primera propuesta de la construcción de cartas de control basadas en datos óptimamente escalados y las cartas de control basadas en NLPCA para características de calidad mixtas.
- En el cuarto Capítulo se presenta la aplicación de las técnicas, PCA y NLPCA, sobre la base de datos perteneciente al proyecto "Corazones de Baependi"(Processo Fapesp 2007/58150-7).
- En los Capítulos 5 y 6 se presenta la conclusión con los principales resultados y contribuciones de este trabajo así como algunas sugerencias para trabajos futuros.

Capítulo 2

Análisis de componentes principales lineales - PCA

El análisis de componentes principales es una técnica estadística ampliamente utilizada en la actualidad para el análisis de datos multivariados. El objetivo principal de esta herramienta es reducir la dimensionalidad de un conjunto de observaciones con una gran cantidad de variables, ayudándose del estudio de la estructura de varianzas-covarianzas (o de correlación) entre las variables de entrada. A partir de la proyección de las observaciones iniciales sobre las direcciones de máxima varianza se obtendrá un nuevo espacio de representación de los elementos iniciales en el que se puede eliminar fácilmente aquellas componentes con menor varianza, garantizando la mínima pérdida de información. También este análisis busca transformar las variables originales obteniéndose combinaciones lineales de ellas que pueden o no ser interpretables en el contexto del estudio.

Esta técnica multivariada se desarrolló inicialmente por Pearson a finales del siglo XIX (Pearson, 1901) a modo de solución para algunos problemas que eran de interés para la biometría de la época. Posteriormente fue estudiada por Hotelling en los años 30, quien se basó en los trabajos de Pearson y en las investigaciones sobre ajustes ortogonales por mínimos cuadrados (Hotelling, 1933). Sin embargo, es hasta la aparición de los computadores que se da el auge de la aplicación en las más diversas áreas.

La base matemática que se utiliza para desarrollar el PCA es el álgebra lineal. A continuación se va a presentar el desarrollo teórico que realiza el PCA con el objetivo final de disminuir las dimensiones de un conjunto

de datos de entrada mediante una transformación lineal que los proyecte sobre las direcciones de máxima varianza. También dentro de la teoría se verá como el PCA está altamente relacionado con la técnica algebraica de la descomposición en valores singulares (SVD - *Singular Value Decomposition*). Además, la compresión de como están relacionados el PCA y la SVD ayudará al mejor entendimiento de las posibles aplicaciones.

2.1. Formulación Matemática del PCA

Sea $\mathbf{X} \in \mathbb{R}^{n\times p}$ una matriz aleatoria conformada por las mediciones de p características o variables $\mathbf{x}_j \in \mathbb{R}^n$, medidas en n individuos $\mathbf{x}_{(i)} \in \mathbb{R}^p$, con $j = 1, 2, ..., p$ e $i = 1, 2, ..., n$. De forma que $\mathbf{X}$ puede ser escrita en términos de las variables o vectores columna como $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_p\end{bmatrix}$, en términos de los individuos o vectores fila como $\mathbf{X}' = \begin{bmatrix}\mathbf{x}_{(1)} & \mathbf{x}_{(2)} & \dots & \mathbf{x}_{(n)}\end{bmatrix}$ o en términos de cada medida como

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}. \tag{2.1}$$

El análisis de componentes principales (PCA - *Principal Component Analysis*) pretende explicar la estructura de la covarianza de la matriz $\mathbf{X}$ a través de unas pocas combinaciones lineales de sus variables, con el objetivo de reducir el tamaño de los datos a analizar sin una pérdida significativa de la variabilidad contenida en ellos. También el PCA busca identificar las direcciones de mayor variabilidad de los datos con el fin de usarlas como ejes de un nuevo subespacio, por lo tanto, es necesario encontrar una base de vectores unitarios $\mathbf{v}_j \in \mathbb{R}^p$, donde $j = 1, 2, ..., q$ con $q \leq p$, sobre los cuales se puedan proyectar los datos $\mathbf{x}_{(i)}$. Los ejes obtenidos a partir de la proyección de los datos sobre el nuevo subespacio son las componentes principales $\mathbf{y}_j \in \mathbb{R}^n$, y se obtendrán tantas componentes como elementos tenga la base $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_q\}$. Cada componente principal será una combinación lineal de las características $\mathbf{x}_j$ de la forma $\mathbf{y}_j = \mathbf{X}\mathbf{v}_j = v_{1j}\mathbf{x}_1 + v_{2j}\mathbf{x}_2 + ... + v_{pj}\mathbf{x}_p$ donde los coeficientes son las entradas de los vectores $\mathbf{v}_j$, gracias a esto, cada vector recibe el nombre de vector de pesos y la matriz $\mathbf{V}_q = \begin{bmatrix}\mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_q\end{bmatrix}$ de matriz de pesos. Por su parte, la matriz de componentes principales puede ser escrita tanto en términos de las componentes $\mathbf{Y}_q = \begin{bmatrix}\mathbf{y}_1 & \mathbf{y}_2 & \dots & \mathbf{y}_q\end{bmatrix}$ como de los vectores de puntajes de los individuos $\mathbf{y}_{(i)} \in \mathbb{R}^q$, es decir, en términos de los vectores fila $\mathbf{Y}'_q = \begin{bmatrix}\mathbf{y}'_{(1)} & \mathbf{y}'_{(2)} & \dots & \mathbf{y}'_{(n)}\end{bmatrix}$, con $\mathbf{Y}_q = \mathbf{X}\mathbf{V}_q$

de dimensión $n \times q$.

Para realizar la búsqueda de los ejes de máxima variabilidad es necesario definir una medida de dispersión de las observaciones. Dos medidas comúnmente utilizadas en el PCA son la matriz de covarianzas $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ y la matriz de correlaciones $\boldsymbol{\rho} \in \mathbb{R}^{p\times p}$. La matriz de covarianzas nos permite saber qué tanto se dispersan en promedio los datos alrededor de su centro de masa, sin embargo, esta medida puede ser afectada por el uso de escalas de medición muy desiguales a la hora de capturar su variabilidad, dándole prioridad a aquellas variables con órdenes de magnitud mayor. Para solucionar este problema, se recomienda utilizar la matriz de correlaciones en lugar de la matriz de covarianzas, ya que cada variable es corregida en magnitud por su respectiva desviación estándar. En estadística multivariada tanto las medidas de varianza como de correlación utilizan el vector de medias $\boldsymbol{\mu} \in \mathbb{R}^p$ como punto de referencia, ya que en él se ubica el centro de masa de los datos. Las medidas de dispersión indicadas anteriormente comparten como característica principal el ser matrices simétricas semidefinidas positivas, esto es gracias a que las propiedades de la multiplicación matricial y la transpuesta permiten que el producto $\mathbf{X}_c'\mathbf{X}_c$ cumple con la forma cuadrática $\mathbf{v}'\mathbf{X}_c'\mathbf{X}_c\mathbf{v} \geq 0$, donde $\mathbf{X}_c$ es la matriz corregida por su media y $\mathbf{v}$ es un vector unitario de tamaño p. La formalización de las ideas presentadas en esta sección se describen con mayor profundidad en el apéndice A.

2.2. Primera componente principal

Ya que el PCA consiste en encontrar un conjunto de p vectores unitarios ortogonales que maximicen la proyección de la variabilidad de los datos contenida en $\mathbf{X}_c'\mathbf{X}_c$, es posible abordar esta búsqueda como un problema de optimización enfocado en maximizar la proyección de $\mathbf{X}_c'\mathbf{X}_c$ sobre un conjunto de vectores $\mathbf{v}_j$, es decir, maximizar la distancia dada por la forma cuadrática $\mathbf{v}_j'\mathbf{X}_c'\mathbf{X}_c\mathbf{v}_j$, donde los vectores $\mathbf{v}_j$ se encuentran sujetos a condiciones de norma unitaria $\mathbf{v}_j'\mathbf{v}_j = 1$ y ortogonalidad $\mathbf{v}_j'\mathbf{v}_l = 0$, para $j, l = 1, 2, \ldots, p$ con $j \neq l$. Las condiciones de ortogonalidad aseguran que no haya redundancia en la información capturada por los nuevos ejes e impide la dependencia lineal entre los vectores de la base al evitar que se correlacionen entre ellos. Sea $\mathbf{X}_c'\mathbf{X}_c \in \mathbb{R}^{p\times p}$ una matriz cuadrada semidefinida positiva y $\mathbf{v}_1 \in \mathbb{R}^p$ un vector tal que $\mathbf{v}_1'\mathbf{v}_1 = 1$. Se tiene como objetivo buscar el valor de $\mathbf{v}_1$ que permita maximizar la proyección de la matriz $\mathbf{X}_c'\mathbf{X}_c$ sobre él. Para esto se plantea el siguiente problema de optimización

$$\min_{\mathbf{v}_1} ||\mathbf{X}_c - \hat{\mathbf{X}}_c||_F^2 \text{ sujeto a } \mathbf{v}_1'\mathbf{v}_1 = 1, \tag{2.2}$$

o su equivalente

$$\max_{\mathbf{v}_1} \mathbf{v}_1'\mathbf{X}_c'\mathbf{X}_c\mathbf{v}_1 \text{ sujeto a } \mathbf{v}_1'\mathbf{v}_1 = 1, \tag{2.3}$$

cuyo Lagrangiano es

$$L = \mathbf{v}_1'\mathbf{X}_c'\mathbf{X}_c\mathbf{v}_1 - \lambda_1(\mathbf{v}_1'\mathbf{v}_1 - 1). \tag{2.4}$$

Al calcular el gradiente respecto a $\mathbf{v}_1$ y λ_1 e igualar a cero se obtiene

$$\nabla L(\mathbf{v}_1) = 0,$$

$$\nabla L(\mathbf{v}_1) = 2\mathbf{X}_c'\mathbf{X}_c\mathbf{v}_1 - 2\lambda_1\mathbf{v}_1,$$

$$\mathbf{X}_c'\mathbf{X}_c\mathbf{v}_1 = \lambda_1\mathbf{v}_1,$$

donde el vector que captura la máxima variabilidad de los datos en el nuevo subespacio es el vector propio $\mathbf{v}_1$ relacionado con el valor propio λ_1 de la matriz $\mathbf{X}_c'\mathbf{X}_c$.

2.3. Otras componentes principales

Por lo general, no es de utilidad conservar únicamente la primera componente principal ya que muchas veces el objetivo de la investigación puede requerir conservar la mayor cantidad de información posible o, si el problema es de visualización, del uso de dos o tres componentes con el fin de obtener al menos una idea gráfica del comportamiento de los datos. De aquí, que sea necesario buscar las demás componentes principales teniendo como prioridad la restricción de ortogonalidad entre los vectores unitarios $\mathbf{v}_j$. Así, las nuevas restricciones son $\mathbf{v}_j'\mathbf{v}_j = 1$ y $\mathbf{v}_j'\mathbf{v}_l = 0$ con $j, l = 1, 2, ..., p$ y $j \neq l$. De esta manera, se define el vector $\mathbf{v}_j$ como la solución del problema de optimización con restricciones

$$\max_{\mathbf{v}_j} \mathbf{v}_j'\mathbf{X}_c'\mathbf{X}_c\mathbf{v}_j \text{ sujeto a } \mathbf{v}_j'\mathbf{v}_j = 1 \text{ y } \mathbf{v}_j'\mathbf{v}_l = 0,\ l = 1, 2, ..., j-1.$$

La anterior proposición indica que $\mathbf{v}_j$ es un vector propio unitario de $\mathbf{X}_c'\mathbf{X}_c$ correspondiente al j-ésimo valor propio más grande.

Por inducción sobre j: Ya se ha mostrado gracias a la proposición anterior que el paso $j = 1$, donde no hay restricciones de ortogonalidad, es verdadero. Entonces, se asume que la proposición es verdadera para los primeros j vectores de pesos $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_j$, y se considera el problema de encontrar $\mathbf{v}_{j+1}$.

Por hipótesis de inducción, se sabe que $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_j$ son vectores propios unitarios ortogonales de $\mathbf{X}_c'\mathbf{X}_c$. Vale notar que λ_j es el j-ésimo valor propio más grande de $\mathbf{X}_c'\mathbf{X}_c$ y cumple con la definición $\mathbf{X}_c'\mathbf{X}_c\mathbf{v}_j = \lambda_j\mathbf{v}_j$ donde $\mathbf{v}_j$ es el vector propio correspondiente. Así el Lagrangiano de la función objetivo es

$$L(\mathbf{v}_j) = \mathbf{v}_j'\mathbf{X}_c'\mathbf{X}_c\mathbf{v}_j - \lambda_j(\mathbf{v}_j'\mathbf{v}_j - 1) - \sum_{l=1}^{j-1}\eta_l\mathbf{v}_j'\mathbf{v}_l, \tag{2.5}$$

y su gradiente

$$\nabla L(\mathbf{v}_j) = 2\mathbf{X}_c'\mathbf{X}_c\mathbf{v}_j - 2\lambda_j\mathbf{v}_j - \sum_{l=1}^{j-1}\eta_l\mathbf{v}_l.$$

Para el paso $j + 1$ se tiene

$$L(\mathbf{v}_{j+1}) = \mathbf{v}_{j+1}'\mathbf{X}_c'\mathbf{X}_c\mathbf{v}_{j+1} - \lambda_{j+1}(\mathbf{v}_{j+1}'\mathbf{v}_{j+1} - 1) - \sum_{l=1}^{j}\eta_l\mathbf{v}_{j+1}'\mathbf{v}_l,$$

$$\mathbf{0} = \nabla L(\mathbf{v}_{j+1}) = 2\mathbf{X}_c'\mathbf{X}_c\mathbf{v}_{j+1} - 2\lambda_{j+1}\mathbf{v}_{j+1} - \sum_{l=1}^{j}\eta_l\mathbf{v}_l. \tag{2.6}$$

Esto implica que si $\mathbf{v}_{j+1}$ es ortogonal a $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_j$, entonces

$$\begin{aligned}
0 &= \mathbf{v}_m'\mathbf{0}, \\
&= 2\mathbf{v}_m'\mathbf{X}_c'\mathbf{X}_c\mathbf{v}_{j+1} - 2\lambda_{j+1}\mathbf{v}_m'\mathbf{v}_{j+1} - \sum_{l=1}^{j}\eta_l\mathbf{v}_m'\mathbf{v}_l, \\
&= 2(\mathbf{X}_c'\mathbf{X}_c\mathbf{v}_m)'\mathbf{v}_{j+1} + \eta_m, \\
&= 2(\lambda_m\mathbf{v}_m)'\mathbf{v}_{j+1} + \eta_m \\
&= 2\lambda_m\mathbf{v}_m'\mathbf{v}_{j+1} + \eta_m, \\
&= \eta_m,
\end{aligned}$$

para todo $m = 1, 2, ..., j$. Reemplazando estos valores dentro de la ecuación 2.6 se nota que $\mathbf{v}_{j+1}$ debe satisfacer $\mathbf{X}_c'\mathbf{X}_c\mathbf{v}_{j+1} = \lambda_{j+1}\mathbf{v}_{j+1}$, es decir, $\mathbf{v}_{j+1}$

es un vector propio de $\mathbf{X}_c'\mathbf{X}_c$ con valor propio λ_{j+1}.

Dado que el objetivo del problema es maximizar la función de pérdida, siempre se busca que el valor propio obtenido sea aquel valor propio más grande cuyo vector propio respete las restricciones de ortogonalidad. Por lo tanto, $\mathbf{v}_{j+1}$ debe ser el vector propio de $\mathbf{X}_c'\mathbf{X}_c$ correspondiente al $(j+1)$-ésimo valor propio más grande. Si se presentan valores propios iguales es posible utilizar el proceso de descomposición de Gram-Schmidt con el fin de encontrar una base ortogonal equivalente (Nasiriany et al., 2019).

2.4. Descomposición en valores singulares - SVD

El origen de los valores singulares se encuentra en el intento de los geómetras del siglo XIX por conseguir en lenguaje actual, la reducción de una forma cuadrática a forma diagonal mediante cambios de base ortogonales (Steward, 1993).

Sea $\mathbf{A} \in \mathbb{R}^{n\times p}$, entonces existe una matriz ortogonal $\mathbf{U} \in \mathbb{R}^{n\times n}$ y una matriz ortogonal $\mathbf{V} \in \mathbb{R}^{p\times p}$ tal que

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}', \tag{2.7}$$

donde $\mathbf{\Sigma}$ en este caso es llamada matriz de valores singulares y es una matriz diagonal de entradas $\sigma_j \geq 0$ con $j = 1, 2, ..., \text{mín}(n, p)$, y las constantes σ_j son los valores singulares de $\mathbf{A}$.

Se define el rango de la matriz $rank(\mathbf{A})$ como el número de valores singulares no nulos de la matriz $\mathbf{A} \in \mathbb{R}^{n\times p}$. Si todas las filas y columnas de la matriz $\mathbf{A}$ son linealmente independientes entonces $r = rank(\mathbf{A}) = \text{mín}\{n, p\}$. La descomposición en valores singulares también puede ser expresada como una expansión de la matriz $\mathbf{A}$ que depende del rango r de la matriz. Esto es expuesto en el teorema de Eckart - Young, el cual será utilizado como una herramienta fundamental en los resultados del análisis de componentes principales tanto lineales como no lineales.

Teorema. *Suponga una matriz* $\mathbf{A} \in \mathbb{R}^{n\times p}$ *con rango* $rank(\mathbf{A}) \leq min(n, p)$ *y sea* $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}' = \sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j'$ *su descomposición en valores singulares.*

Entonces

$$\mathbf{A}_r = \sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j' = \mathbf{U} \begin{bmatrix} \sigma_1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & 0 & 0 & \dots & 0 \\ 0 & 0 & \sigma_r & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 \end{bmatrix} \mathbf{V}',$$

donde $r \leq rank(\mathbf{A})$*, es la mejor aproximación de rango* r *a* $\mathbf{A}$ *en el sentido que*

$$||\mathbf{A} - \mathbf{A}_r||_F \leq ||\mathbf{A} - \tilde{\mathbf{A}}||_F,$$

para cualquier $\tilde{\mathbf{A}}$ *tal que* $rank(\tilde{\mathbf{A}}) \leq r$.

Específicamente, existen r constantes positivas $\sigma_1, \sigma_2, ..., \sigma_r$, r vectores unitarios ortogonales $\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_r$ de tamaño $n \times 1$ y r vectores unitarios ortogonales $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_r$ de tamaño $p \times 1$, tal que

$$\mathbf{A}_r = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r',$$

donde $\mathbf{U}_r = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_r \end{bmatrix}$, $\mathbf{V}_r = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_r \end{bmatrix}$ y $\mathbf{\Sigma}_r$ es una matriz diagonal con entradas σ_j. Aquí $\mathbf{A}\mathbf{A}'$ tiene parejas de valor-vector propio $(\sigma_j^2, \mathbf{u}_j)$, entonces

$$\mathbf{A}\mathbf{A}'\mathbf{u}_j = \sigma_j^2 \mathbf{u}_j,$$

con $\sigma_1^2, \sigma_2^2, ..., \sigma_r^2 > 0$ y $\sigma_{r+1}^2, \sigma_{r+2}^2, ..., \sigma_p^2 = 0$ para $n > p$. Alternativamente, los vectores $\mathbf{v}_j$ son los vectores propios de $\mathbf{A}'\mathbf{A}$ con los mismo valores propios no nulos σ_j^2.

2.5. Descomposición en valores propios - autovalores

Sea $\mathbf{A}$ una matriz de tamaño $n \times p$ y λ_j un valor propio de $\mathbf{A}'\mathbf{A}$. Si $\mathbf{v}_j \in \mathbb{R}^p$ es un vector unitario no nulo tal que

$$\mathbf{A}'\mathbf{A}\mathbf{v}_j = \lambda_j \mathbf{v}_j,$$

entonces $\mathbf{v}_j$ se dice que es un vector propio o vector característico de la matriz $\mathbf{A}'\mathbf{A}$ asociado con el valor propio λ_j. Donde los valores propios de $\mathbf{A}'\mathbf{A}$

son reales y no negativos $\lambda_j \geq 0$.

Ya que para cualquier vector propio $\mathbf{v}_j$, se tiene que

$$\mathbf{A}'\mathbf{A}\mathbf{v}_j = \lambda_j \mathbf{v}_j,$$

si se multiplica a ambos lados por $\mathbf{v}_j$, se obtiene la igualdad

$$\mathbf{v}_j'\mathbf{A}'\mathbf{A}\mathbf{v}_j = \lambda_j \mathbf{v}_j'\mathbf{v}_j,$$

que indica que $||\mathbf{A}\mathbf{v}_j||_F^2 = \lambda_j ||\mathbf{v}_j||_F^2$, por tanto $\lambda_j \geq 0$.

De modo que $\mathbf{A}'\mathbf{A}$ puede ser diagonalizada ortogonalmente por la forma

$$\mathbf{A}'\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}',$$

donde $\mathbf{V} \in \mathbb{R}^{p\times p}$ es una matriz ortogonal conformada por los vectores unitarios $\mathbf{v}_j \in \mathbb{R}^p$ y $\Lambda \in \mathbb{R}^{p\times p}$ es una matriz diagonal conformada por los valores propios λ_j con $j = 1, 2, ..., p$, donde $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$.

De forma que es posible obtener la matriz de pesos $\mathbf{V}$ por medio de la descomposición en valores propios de $\mathbf{S} = \mathbf{V}\Lambda\mathbf{V}'$ o $\mathbf{R} = \mathbf{V}\Lambda\mathbf{V}'$, donde Λ contiene los valores propios de la matriz en estudio. En este punto, es necesario resaltar que la descomposiciones por valores propios de las matrices $\mathbf{S}$ y $\mathbf{R}$ no son equivalentes, debido a la diferencia en el orden de magnitud que manejan los datos de cada matriz (Ipsen, 2009; Johnson & Wichern, 2007).

Ya que la expansión de la matriz $\mathbf{A}$ por medio de la descomposición en valores singulares escrita en términos de las matrices $\mathbf{U}, \mathbf{V}, \mathbf{\Sigma}$ es $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}'$ la descomposición en valores singulares de $\mathbf{A}\mathbf{A}'$ será $\mathbf{A}\mathbf{A}' = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}'$ donde los vectores columna de $\mathbf{U}$ son los n vectores propios unitarios ortogonales de $\mathbf{A}\mathbf{A}'$. Por otro lado, la descomposición matricial de $\mathbf{A}'\mathbf{A}$ será $\mathbf{A}'\mathbf{A} = \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}'$ donde los vectores columna de $\mathbf{V}$ son los p vectores propios unitarios ortogonales de $\mathbf{A}'\mathbf{A}$. De forma que, por la definición de valores propios se tiene que $\Lambda = \mathbf{\Sigma}^2$ y, por lo tanto, $\lambda_j = \sigma_j^2$. De esta manera, es posible obtener la descomposición en valores propios de $\mathbf{X}_c'\mathbf{X}_c$ a partir de la descomposición en valores singulares de la matriz $\mathbf{X}_c$ (Nasiriany et al., 2019; Ipsen, 2009).

2.6. Proyección de los datos en el nuevo espacio

Por lo general, aquellos estudios que involucran análisis de componentes principales suelen enfocarse en las relaciones presentes entre los individuos,

de forma que se trabaja en el espacio $\mathbb{R}^p$. Es por esto que se utiliza como medida de variabilidad los estadísticos desprendidos de la multiplicación $\mathbf{X}_c'\mathbf{X}_c$, tales como la covarianza y la correlación entre los individuos. aclarar que si el estudio tiene como objetivo evaluar la relación entre los individuos en lugar de la relación entre las variables como es utilizado tradicionalmente la proyección de los datos se realizan en el espacio $\mathbb{R}^n$ usando los usando descriptivos desprendidos del producto $\mathbf{X}_c\mathbf{X}_c'$ el cual funciona como una medida de dispersión de las variables, por medio de las llamadas coordenadas principales. De esta forma, se hace necesario estudiar la obtención tanto de las componentes principales como de las coordenadas principales a partir de $\mathbf{X}_c'\mathbf{X}_c$ y $\mathbf{X}_c\mathbf{X}_c'$ respectivamente, teniendo en cuenta la conexión entre ambas y las aplicaciones en las cuales el uso de cada una es necesario.

Análisis en el espacio $\mathbb{R}^p$

Como ya se indicó, en el espacio $\mathbb{R}^p$ el principal propósito es analizar las relaciones entre los individuos a partir de $\mathbf{X}_c'\mathbf{X}_c$, utilizando medidas de variabilidad como la correlación o la covarianza, en el caso de esta última tenemos que su descomposición en valores y vectores propios es denotada por $\mathbf{\Sigma} = \mathbf{V}\Lambda\mathbf{V}'$ donde $\mathbf{V}$ es la matriz con los p vectores propios y Λ es la matriz de valores propios.

Las componentes principales obtenidas de la descomposición espectral de la matriz $\mathbf{\Sigma}$ son las columnas de la matriz $\mathbf{Y}$ dada por

$$\mathbf{Y} = \mathbf{X}_c\mathbf{V}. \tag{2.8}$$

En la matriz de componentes principales las coordenadas de los n individuos proyectados sobre en el nuevo eje factorial $\mathbf{v}_j$ son las n componentes del vector $\mathbf{y}_j = \mathbf{X}\mathbf{v}_j$, es decir, la j-ésima componente principal.

Análisis en el espacio $\mathbb{R}^n$

Por otra parte, en el espacio $\mathbb{R}^n$ el objetivo es evaluar la relación entre las variables o características, para ello se hace uso de la descomposición en valores y vectores propios de la matriz $\mathbf{X}_c\mathbf{X}_c'$ denotada por $\mathbf{\Psi}$, donde $\mathbf{\Psi}$ es la matriz de covarianzas entre las filas de la matriz $\mathbf{X}$, obteniendo $\mathbf{\Psi}_{n\times n} = \mathbf{U}\mathbf{\Lambda}_\Psi\mathbf{U}'$ donde $\mathbf{U}$ es la matriz con los n vectores propios y $\mathbf{\Lambda}_\Psi$ es la matriz de valores propios. Las proyecciones de los datos sobre los ejes $\mathbf{u}_i$ reciben el nombre de coordenadas principales, obtenidas de la matriz $\mathbf{\Psi}_{n\times n}$ y son las columnas de la matriz $\mathbf{F}$ dadas en la ecuación

$$\mathbf{F} = \mathbf{X}_c'\mathbf{U}, \tag{2.9}$$

donde las coordenadas de las p características proyectadas en el nuevo eje factorial $\mathbf{u}_i$ son las p entradas del vector $\mathbf{f}_i = \mathbf{X}'\mathbf{u}_i$, es decir, la i-ésima coordenada principal.

Como muestra Gower (1996), el análisis de componentes principales sobre una matriz simétrica $p \times p$ y el análisis de coordenadas principales sobre una matriz $n \times n$ se consideran duales uno al otro cuando ambos conducen a un conjunto de puntos con las mismas distancias entre ellos. La descomposición en valores singulares de una matriz rectangular $\mathbf{X}_c$ de tamaño $n \times p$ de rango n permite estudiar dicha dualidad (Johnson & Wichern, 2007).

Equivalencia entre las dos metodologías

Para estudiar la dualidad entre las componentes principales y las coordenadas principales es necesario recordar que la relación entre las matrices de vectores propios ortogonales $\mathbf{V}$ y $\mathbf{U}$ proviene de la descomposición de la matriz $\mathbf{X}_c$ en la forma

$$\mathbf{X}_c = \mathbf{U}\Lambda^{1/2}\mathbf{V}', \tag{2.10}$$

donde $\Lambda^{1/2}$ es la matriz de valores singulares de $\mathbf{X}_c$, $\mathbf{U}$ representa los vectores propios asociados a $\mathbf{X}_c\mathbf{X}_c'$ y $\mathbf{V}$ los vectores propios asociados a $\mathbf{X}_c'\mathbf{X}_c$, como anotado al inicio de esta sección. Multiplicando por $\mathbf{V}$ a la derecha de cada elemento de la ecuación 2.10 se obtiene

$$\mathbf{X}_c\mathbf{V} = \mathbf{U}\Lambda^{1/2}\mathbf{V}'\mathbf{V}, \tag{2.11}$$

y haciendo uso de la ortogonalidad de la matriz $\mathbf{V}$ se sabe que $\mathbf{V}'\mathbf{V} = \mathbf{I}_p$, entonces

$$\mathbf{U} = \mathbf{X}_c\mathbf{V}\Lambda^{-1/2}, \tag{2.12}$$

de forma que el k-ésimo vector propio de la matriz $\mathbf{U}$ se conoce por

$$\mathbf{u}_k = \lambda_k^{-1/2}\mathbf{X}_c\mathbf{v}_k.$$

Al recordar que las componentes principales $\mathbf{Y}$ se calculan por medio de $\mathbf{X}_c\mathbf{V}$, como descrito en la ecuación 2.8, entonces reemplazando en la ecuación 2.12, la matriz $\mathbf{U}$ es escrita de la siguiente forma

$$\mathbf{U} = \mathbf{Y}\Lambda^{-1/2}. \tag{2.13}$$

Sustituyendo $\mathbf{U}$ de la ecuación 2.13 en la ecuación 2.9 se obtiene una forma de calcular las coordenadas principales a partir de las componentes principales, a saber

$$\mathbf{F} = \mathbf{X}_c'\mathbf{Y}\Lambda^{-1/2}. \tag{2.14}$$

Para entender en qué casos es necesario recurrir a las coordenadas principales en vez de las componentes principales, hay que fijarse en el rango de la matriz, ya que una matriz tendrá tantos valores singulares como el número mínimo de columnas o filas linealmente independientes que posea. La cantidad de valores singulares no nulos con los que cuenta la matriz reflejan la cantidad de información contenida en los datos, mientras que los valores singulares iguales a cero representan la cantidad de redundancia presente en ellos. Es decir, si tenemos una matriz de datos $\mathbf{X}$ de tamaño $n \times p$ donde $p > n$ y hacemos la descomposición en valores propios de $\mathbf{X}_c'\mathbf{X}_c = \mathbf{\Sigma} = \mathbf{V}\Lambda\mathbf{V}'$ la cantidad de vectores propios obtenidos será p, sin embargo, la cantidad de valores propios diferentes de cero será n ya que $rank(\mathbf{X}) = n$, o sea, hay $p-n$ valores propios iguales a cero lo que indica que $p - n$ vectores propios están capturando información redundante de los datos (Ipsen, 2009). En la práctica esto depende de cuántas variables se deben medir y de cuántas muestras se pueden obtener de la población objetivo. Tradicionalmente, es común ver experimentos en el que es mucho más grande el número de mediciones n que el número de variables o características medidas p. Si suponemos que todas las variables medidas son linealmente independientes entre sí, es decir, una variable o una combinación lineal de variables no explican el comportamiento de otra, la matriz de datos $\mathbf{X}$ tendrá rango columna completo $rank(\mathbf{X}) = p$, de forma que contará con p valores singulares $\sqrt{\lambda_j}$, con p vectores propios $\mathbf{v}_j$ y, por lo tanto, con p componentes principales $\mathbf{y}_j$. En el caso en el que sea necesario medir un gran número de variables simultáneamente y sea imposible obtener un número mayor de muestras, suponiendo que todas las filas, los individuos, son linealmente independientes la matriz de datos tendrá rango fila completo $rank(\mathbf{X}) = n$, es decir, tendrá n valores singulares $\sqrt{\lambda_i}$, n vectores propios u_i y, por lo tanto, n coordenadas principales $\mathbf{f}_i$.

Como conclusión general, se establece esta dualidad, donde los vectores de la matriz $\mathbf{V}$ son funciones de la matriz $\mathbf{U}$ y viceversa. Con este resultado se busca evitar problemas computacionales que surgen de la descomposición espectral de la matriz de datos en un camino estadístico cuya dimensión puede ser extremadamente grande. Por ejemplo, en estudios en genética cuyos datos son de altísima dimensión en las variables, se hace la descomposición espectral por individuos (que generalmente son pocos) y se generaliza a las

variables (Duarte et al., 2015). Sin embargo, en la actualidad técnicas como Randomized PCA (Feng, et al. 2018) resuelven este problema con gran facilidad, gracias a su capacidad de generar un modelo de reducción de dimensión por medio de un muestreo aleatorio de los datos que permite una construcción aproximada de la descomposición de la matriz.

2.6.1. Biplots

Los Biplots son comunmente usados para mostrar una representación conjunta de las filas y columnas de una matriz $\mathbf{X}$. El Biplot aproxima la distribución de una muestra multivariada en un espacio de dimensión reducida, normalmente de dimensión dos, y superpone sobre la misma representaciones de las variables sobre las que se mide la muestra (Gower, 1996). Se basa en el mismo principio utilizado en la técnica de reducción de dimensionalidad a través de la Descomposición del Valor Singular (SVD) de la matriz $\mathbf{X}$ como es anotado en la Sección 2.5, ecuación 2.7. La diferencia fundamental es que en este caso el objetivo es reproducir los datos originales a través de la representación conjunta de filas y columnas de una matriz (Gabriel, 1971).

En la representación gráfica, el Biplot considera solamente dos vectores singulares ($k = 2$), por tanto la matriz $\mathbf{X}_{n\times p}$ es aproximada por

$$\begin{aligned} \mathbf{X}_{n\times p} \approx \mathbf{U}_{(2)}\Lambda_{(2)}^{1/2}\mathbf{V}'_{(2)} &= (\mathbf{U}_{(2)}\Lambda_{(2)}^{1/2-c/2})(\Lambda^{c/2}\mathbf{V}'_{(2)}), & (2.15)\\ &= \mathbf{G}\mathbf{H}', & (2.16) \end{aligned}$$

donde $\mathbf{U}_{(2)}$ y $\mathbf{V}_{(2)}$ indica solamente dos vectores propios de las matrices $\mathbf{U}$ y $\mathbf{V}$, respectivamente. Siguiendo la notación descrita en Gabriel (1971) la matriz $\mathbf{G}$ y $\mathbf{H}$ son $\mathbf{G} = \mathbf{U}_{(2)}\ \Lambda_{(2)}^{1/2-c/2}$ y $\mathbf{H}' = \Lambda_{(2)}^{c/2}\mathbf{V}'_{(2)}$ (como ilustrado en la ecuación 2.15 y 2.16) con $c \in [0, 1]$.

Note que $\mathbf{G}$ es la representación de las n filas de la matriz $\mathbf{X}$ en un espacio de dos dimensiones y $\mathbf{H}$ la representación de las p columnas de $\mathbf{X}$ en este mismo espacio. Los valores más comúnmente usados para c son 0 , 1/2 y 1.

Es posible afirmar que el biplot puede ser construido por un plano bidimensional por la representación conjunta de n puntos de individuos ($\mathbf{U}_1\lambda^{1/2-c/2}$ $\mathbf{U}_2\lambda^{1/2-c/2}$) y p puntos que corresponden a las variables ($\mathbf{V}_1\lambda^{c/2}$ $\mathbf{V}_2\lambda^{c/2}$). Para $c = 0$, las filas están en las coordenadas principales y las columnas en las coordenadas estándar, denominada forma biplot, que favorece la visualización de los individuos; para $c = 1$, las filas están en coordenadas estándar

y las columnas en coordenadas principales, llamadas biplot de covarianza, lo que favorece la visualización de las variables. Cuando $c = 1/2$, el biplot favorece la visualización del efecto de interacción (Greenacre, 1984).

Es importante notar que el biplot es un análisis de componentes principales, en el cual la información de las columnas (variables) se agrega en el mismo gráfico en el que se representan las filas (individuos). Dado que, en los biplots, los individuos y las variables están representados en el mismo gráfico, tiene sentido evaluar las asociaciones entre los individuos y las variables como la preponderancia de una variable para explicar un individuo, o la contribución de los individuos a los valores de una variable. La variabilidad explicada por los ejes Biplot es similar a la variabilidad explicada por el análisis de componentes principales, donde los ejes se obtienen en la dirección de mayor variabilidad.

Un enfoque detallado de Biplots para variables cuantitativas puede ser estudiado en Gower (2015). Por otro lado para variables cualitativas, que pueden ser nominales u ordinales, se recomienda Gower (2016).

2.7. Elección de las componentes principales

Para definir cuántas componentes principales conservar en un estudio es importante tener claro el papel del PCA en él. Si el fin es ver el comportamiento de los datos en dos o tres dimensiones es necesario conservar las dos o tres primeras componentes principales y desistir de las demás. Si por el contrario, lo que se busca es conservar la mayor cantidad de información contenida en los datos, se puede utilizar como criterio la variabilidad conservada por cada una de las componentes.

Sea $\mathbf{\Sigma}$ la matriz de covarianza asociada con la matriz aleatoria $\mathbf{X}$. $\mathbf{\Sigma}$ tiene las parejas de vectores-valores propios $(\lambda_j, \mathbf{v}_j)$ para todo $j = 1, 2, ..., p$ donde $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$. Entonces, la j-ésima componente principal está dada por la combinación lineal

$$\mathbf{y}_j = \mathbf{X}_c\mathbf{v}_j = v_{1j}\mathbf{x}_1 + v_{2j}\mathbf{x}_2 + ... + v_{pj}\mathbf{x}_p,$$

donde $j = 1, 2, ..., p$ y las varianzas de las componentes principales son dadas por

$$Var(\mathbf{y}_j) = Var(\mathbf{X}_c\mathbf{v}_j) = \mathbf{v}_j'\mathbf{\Sigma}\mathbf{v}_j = \mathbf{v}_j'(\lambda_j\mathbf{v}_j) = \lambda_j(\mathbf{v}_j'\mathbf{v}_j) = \lambda_j, \qquad (2.17)$$

que es un resultado relevante, la varianza de las componentes principales están asociadas a los respectivos valores propios. Por otro lado, para $j \neq l$, y considerando la hipótesis de ortogonalidad $\mathbf{v}_j'\mathbf{v}_l = 0$, se encuentra que la covarianza entre las componentes principales es cero, específicamente

$$Cov(\mathbf{y}_j, \mathbf{y}_l) = Cov(\mathbf{X}_c\mathbf{v}_j, \mathbf{X}_c\mathbf{v}_l) = \mathbf{v}_j'\mathbf{\Sigma}\mathbf{v}_l = \mathbf{v}_j'(\lambda_l\mathbf{v}_l) = \lambda_l(\mathbf{v}_j'\mathbf{v}_l) = 0. \quad (2.18)$$

Para conocer cuánta variabilidad conserva cada una de las componentes principales es necesario hallar la varianza total VT de las componentes principales. Recordando que la variabilidad total de los datos es la suma de las varianzas de cada una de las variables, se tiene

$$VT = \sum_{j=1}^{p} \sigma_{jj} = tr(\mathbf{\Sigma}) = tr(\mathbf{V}\Lambda\mathbf{V}') = tr(\mathbf{V}'\mathbf{V}\Lambda) = tr(\Lambda) = \sum_{j=1}^{p} \lambda_j. \quad (2.19)$$

Así que la variabilidad total conservada por las componentes principales es la misma variabilidad presente en los datos originales solo que con una distribución diferente entre los ejes del nuevo subespacio. Los resultados obtenidos en 2.17, 2.18 y 2.19, pueden ser estudiados a profundiad en Johnson & Wichern (2007).

Teniendo en cuenta los resultados anteriores, es posible crear un indicador que permita medir cuánta variabilidad conserva cada una de las componentes principales, el cual recibirá el nombre de varianza proporcional VP

$$VP = \frac{\lambda_j}{VT}, \quad (2.20)$$

con $j = 1, 2, ...p$ la acumulación de la varianza proporcional se llamará varianza proporcional acumulada VPA

$$VPA = \frac{\sum_{j=1}^{q} \lambda_j}{VT}, \quad (2.21)$$

con $q \leq p$.

Usualmente se recomienda a los investigadores conservar la cantidad de componentes que representen entre un 80 % o un 90 % de la varianza total, siendo la primera componente a conservar aquella que tiene asociado el mayor valor propio, la segunda aquella con el segundo valor propio más grande y así sucesivamente. Un apoyo visual común para decidir cuantas componentes principales retener es el gráfico de sedimentación, el cual consiste en

ordenar los valores propios de mayor a menor en el eje x y graficar sus magnitudes en el eje y. Por lo general, se recomienda conservar las componentes principales asociadas a los valores propios que presentan gran magnitud y una diferencia notoria con el valor propio siguiente y no se recomienda retener aquellos vectores propios cuyos valores propios presentan poca magnitud y una pequeña diferencia con el valor siguiente de forma que presentan una tendencia constante con los valores subsecuentes. En general, a las componentes asociadas con los últimos valores propios las podemos relacionar con el ruido presente en los datos, es decir, aquella variabilidad que no explica su comportamiento (Diaz et al., 2012; Peña, 2002; Rencher, 2002).

2.8. Reconstrucción de los datos originales

El PCA suele ser utilizado para retirar aquella información que no ayuda en la explicación del comportamiento de los datos. De forma que muchas veces el interés final es obtener una estimativa de los datos sin el ruido presente en las últimas componentes. Recordando que las componentes principales son las proyecciones de los datos sobre los ejes de máxima variabilidad, es decir $\mathbf{Y} = \mathbf{X}_c\mathbf{V}$, si el interés es reducir el número dimensiones en el nuevo espacio con el fin de entender mejor el comportamiento de los datos se pueden obtener las q primeras componentes principales de la forma $\mathbf{Y}_q = \mathbf{X}_c\mathbf{V}_q$, con $q \leq p$, donde la matriz $\mathbf{V}_q$ se compone de los vectores correspondientes a los q valores propios más grandes, $\mathbf{V}_q = \begin{bmatrix}\mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_q\end{bmatrix}$.

Es posible reconstruir los datos originales a partir de las componentes principales por medio de la operación inversa $\mathbf{X}_c = \mathbf{Y}\mathbf{V}'$, de forma que se puede obtener una estimativa de los datos originales $\tilde{\mathbf{X}}_q$ que no contenga aquella variabilidad presente en las últimas componentes. Para esto, es necesario obtener los datos originales a partir de $\mathbf{Y}_q$ por medio de la operación $\tilde{\mathbf{X}}_q = \mathbf{Y}_q\mathbf{V}'_q = \mathbf{X}_c\mathbf{V}_q\mathbf{V}'_q$. Ya que el PCA puede ser interpretado como la minimización de la distancia perpendicular entre los puntos que representan los individuos y los ejes del nuevo espacio donde serán proyectados. La proyección ortogonal de un vector $\mathbf{x}_{(i)}$ corregido por su media sobre el subespacio generado por un vector unitario $\mathbf{v}_j$ es igual a $\mathbf{v}_j$ escalado por la proyección de $\mathbf{x}_{(i)}$ sobre $\mathbf{v}_j$

$$\tilde{\mathbf{x}}_{(i)} = P_{\mathbf{v}_1}\mathbf{x}_{(i)} = (\mathbf{x}'_{(i)}\mathbf{v}_j)\mathbf{v}_j. \tag{2.22}$$

De forma que otra derivación de PCA parte del hecho de buscar minimizar el

error de reconstrucción de los datos, obteniendo el problema de optimización

$$\min_{\tilde{\mathbf{X}}_q} ||\mathbf{X}_c - \tilde{\mathbf{X}}_q||_F^2 = \min_{\mathbf{V}_q} ||\mathbf{X}_c - \mathbf{Y}_q\mathbf{V}_q'||_F^2, \tag{2.23}$$

$$\min_{\tilde{\mathbf{X}}_q} ||\mathbf{X}_c - \tilde{\mathbf{X}}_q||_F^2 = \min_{\mathbf{V}_q} ||\mathbf{X}_c - \mathbf{X}_c\mathbf{V}_q\mathbf{V}_q'||_F^2. \tag{2.24}$$

Haciendo uso de las propiedades de la Norma de Frobenius se tiene

$$\sigma(\tilde{\mathbf{X}}_q) = ||\mathbf{X}_c - \tilde{\mathbf{X}}_q||_F^2 = \sum_{j=1}^{q}\sum_{i=1}^{n} ||\mathbf{x}_{(i)} - P_{\mathbf{v}_j}\mathbf{x}_{(i)}||_F^2, \tag{2.25}$$

donde $i = 1, 2, ..., n$ y $j = 1, 2, ..., q$ con $q \leq p$. Para cualquier $\mathbf{x}_{(i)} \in \mathbb{R}^p$, se conoce que $\mathbf{x}_{(i)} - P_{\mathbf{v}_j}\mathbf{x}_{(i)}$ es perpendicular a $P_{\mathbf{v}_j}\mathbf{x}_{(i)}$, entonces por el Teorema de Pitágoras se tiene

$$||\mathbf{x}_{(i)} - P_{\mathbf{v}_j}\mathbf{x}_{(i)}||_F^2 + ||P_{\mathbf{v}_j}\mathbf{x}_{(i)}||_F^2 = ||\mathbf{x}_{(i)}||_F^2, \tag{2.26}$$

de manera que

$$\sum_{j=1}^{q}\sum_{i=1}^{n} ||\mathbf{x}_{(i)} - P_{\mathbf{v}_j}\mathbf{x}_{(i)}||_F^2 = \sum_{j=1}^{q}\sum_{i=1}^{n}(||\mathbf{x}_{(i)}||_F^2 - ||P_{\mathbf{v}_j}\mathbf{x}_{(i)}||_F^2),$$

$$\sum_{j=1}^{q}\sum_{i=1}^{n} ||\mathbf{x}_{(i)} - P_{\mathbf{v}_j}\mathbf{x}_{(i)}||_F^2 = \sum_{j=1}^{q}\sum_{i=1}^{n} ||\mathbf{x}_{(i)}||_F^2 - \sum_{j=1}^{q}\sum_{i=1}^{n} ||(\mathbf{x}_{(i)}'\mathbf{v}_j)\mathbf{v}_j||_F^2,$$

$$\sum_{j=1}^{q}\sum_{i=1}^{n} ||\mathbf{x}_{(i)} - P_{\mathbf{v}_j}\mathbf{x}_{(i)}||_F^2 = \sum_{j=1}^{q}\sum_{i=1}^{n} ||\mathbf{x}_{(i)}||_F^2 - \sum_{j=1}^{q}\sum_{i=1}^{n} (\mathbf{x}_{(i)}'\mathbf{v}_j)'(\mathbf{x}_{(i)}'\mathbf{v}_j).$$

Por lo tanto, minimizar el error de proyección equivale a maximizar las proyecciones de los datos sobre los vectores generadores del nuevo subespacio, específicamente

$$\min_{\tilde{\mathbf{X}}_q} \sigma(\tilde{\mathbf{X}}_q) = \max_{\mathbf{v}_j} \sum_{j=1}^{q}\sum_{i=1}^{n} \mathbf{v}_j'\mathbf{x}_{(i)}\mathbf{x}_{(i)}'\mathbf{v}_j. \tag{2.27}$$

Ya que el problema de optimización se puede resolver para cada uno de los vectores $\mathbf{v}_j$ vale recordar las restricciones de ortogonalidad y norma unitaria con los demás vectores $\mathbf{v}_l$, con $l = 1, 2, ..., j-1$, de forma que el Lagrangiano de la función objetivo para cada $\mathbf{v}_j$ es

$$L(\mathbf{v}_j) = \sum_{i=1}^{n} \mathbf{v}_j'\mathbf{x}_{(i)}\mathbf{x}_{(i)}'\mathbf{v}_j - \lambda_j(\mathbf{v}_j'\mathbf{v}_j - 1) - \eta_l \sum_{l=1}^{j-1} \mathbf{v}_j'\mathbf{v}_l, \tag{2.28}$$

o escrito de mejor forma

$$L(\mathbf{v}_j) = \mathbf{v}_j'\mathbf{X}_c'\mathbf{X}_c\mathbf{v}_j - \lambda_j(\mathbf{v}_j'\mathbf{v}_j - 1) - \eta_l \sum_{l=1}^{j-1} \mathbf{v}_j'\mathbf{v}_l, \tag{2.29}$$

que es el Lagrangiano del problema planteado en la Sección 2.3, específicamente en la ecuación 2.5 y cuya solución cumple con la descomposición en valores y vectores propios de la matriz $\mathbf{X}_c'\mathbf{X}_c$ tomando únicamente los r primeros vectores columna de $\mathbf{V}$ (Nasiriany et al., 2019).

Cabe resaltar que el planteamiento del PCA como la minimización del error de reconstrucción de los datos es un paso fundamental en el planteamiento de técnicas de reducción de dimensión de datos mixtos como el PCA no lineal. La teoría asociada al PCA no lineal será descrita en el Capítulo 3.

2.9. Interpretación de las componentes principales

Para interpretar cada componente principal es necesario examinar la magnitud de los coeficientes de las variables originales, a saber los componentes de los vectores propios, cuanto mayor sea el valor absoluto del coeficiente, más importante será la variable correspondiente en el cálculo del componente. Por otro lado, se puede ver la correlación entre las componentes principales y las variables originales, como es descrito a continuación.

2.9.1. Análisis de los componentes de los vectores de pesos

Cada una de las entradas de los vectores $\mathbf{v}_j' = \begin{bmatrix} v_{1j} & v_{2j} & ... & v_{pj} \end{bmatrix}$ también ameritan inspección. Recordando que los vectores propios $\mathbf{v}_j$ son vectores unitarios por lo que su norma $||\mathbf{v}_j|| = \sqrt{v_{1j}^2 + v_{2j}^2 + ... + v_{pj}^2}$ debe tener una magnitud de 1. Bajo este hecho, sacando el cuadrado de la norma se tiene que la suma de cada una de las entradas del vector al cuadrado será igual a 1, $\sum_{l=1}^{p} v_{lj}^2 = 1$. De forma que se obtiene un criterio para identificar cuáles variables tienen mayor importancia en la combinación lineal de las componentes principales, ya que la cantidad v_{lj} cumple el papel de coeficiente de la combinación lineal $\mathbf{y}_j$ correspondiente al vector $\mathbf{x}_l$ y, v_{lj}^2 indica cuál es la importancia relativa del vector $\mathbf{x}_l$ respecto a los otros vectores en la componente $\mathbf{y}_j$. Por ejemplo, si el coeficiente $v_{lj} = 0{,}50$ quiere decir que un 25 % de la información de la componente $\mathbf{y}_j$ es proporcionada por el vector $\mathbf{x}_l$.

2.9.2. Coeficientes de correlación

Otra característica importante de los coeficientes de las componentes principales v_{lj} es su proporcionalidad directa con el coeficiente de correlación entre $\mathbf{y}_j$ y $\mathbf{x}_l$.

Sean $\mathbf{y}_1 = \mathbf{X}\mathbf{v}_1$, $\mathbf{y}_2 = \mathbf{X}\mathbf{v}_2$, ...,$\mathbf{y}_p = \mathbf{X}\mathbf{v}_p$ las componentes principales obtenidas a partir de $\mathbf{\Sigma}$, entonces

$$\rho_{\mathbf{y}_j,\mathbf{x}_l} = \frac{v_{lj}\sqrt{\lambda_j}}{\sqrt{\sigma_{ll}}}, \tag{2.30}$$

para $j, l = 1, 2, ..., p$, son los coeficientes de correlación entre las componentes $\mathbf{y}_j$ y las variables $\mathbf{x}_l$.

Para la obtención del coeficiente de correlación en 2.30 es necesario fijar un $\mathbf{a}_l' = \begin{bmatrix} 0 & \dots & 0 & 1 & 0 & \dots & 0 \end{bmatrix}$ tal que $\mathbf{x}_l = \mathbf{X}\mathbf{a}_l$ y $Cov(\mathbf{x}_l, \mathbf{y}_j) = Cov(\mathbf{X}\mathbf{a}_l, \mathbf{X}\mathbf{v}_j) = \mathbf{a}_l'\mathbf{\Sigma}\mathbf{v}_j$, como anotado en la ecuación 2.18. Ya que $\mathbf{v}_j$ es un vector propio de $\mathbf{\Sigma}$ por la ecuación característica, se tiene que $\mathbf{\Sigma}\mathbf{v}_j = \lambda_j\mathbf{v}_j$, por lo tanto, $Cov(\mathbf{x}_l, \mathbf{y}_j) = \mathbf{a}_l'\lambda_j\mathbf{v}_j = \lambda_j\mathbf{v}_{lj}$. Entonces, como $Var(\mathbf{y}_j) = \lambda_j$ y $Var(\mathbf{x}_l) = \sigma_{ll}$ se tiene

$$\rho_{\mathbf{y}_j,\mathbf{x}_l} = \frac{Cov(\mathbf{x}_l, \mathbf{y}_j)}{\sqrt{Var(\mathbf{y}_j)}\sqrt{Var(\mathbf{x}_l)}} = \frac{\lambda_j v_{lj}}{\sqrt{\lambda_j}\sqrt{\sigma_{ll}}} = \frac{v_{lj}\sqrt{\lambda_j}}{\sqrt{\sigma_{ll}}},$$

para $j, l = 1, 2, ..., p$, llegando así al resultado planteado en 2.30.

Lo anterior permite conocer, de igual forma, la contribución de cada variable $\mathbf{x}_l$ a la componente $\mathbf{y}_j$. Sin embargo, no indica la importancia de una variable $\mathbf{x}$ en la componente $\mathbf{y}$ en presencia de las otras $\mathbf{x}$; por lo general, si una variable tiene un coeficiente v_{lj} grande su coeficiente de correlación sera de igual forma grande, de manera que el uso de uno u otro dependerá del gusto del investigador (Johnson & Wichern, 2007; Diaz & Morales, 2012).

2.10. Últimas componentes principales

Las últimas componentes son aquellas con valores propios y varianzas cercanas a cero lo que implica que la componente principal es una combinación lineal aproximadamente constante entre las variables de la muestra. De esta manera un valor propio extremadamente pequeño puede ser indicio de

colinealidad entre las variables. Rencher (1998) sugiere que las variables con mayor correlación de aquellas componentes de mínima varianza sean retiradas de la muestra de manera que no distorsionen las primeras componentes principales.

A partir de la ortonormalidad de la matriz $\mathbf{V}$ es posible expresar la matriz identidad de la forma

$$\mathbf{I}_p = \mathbf{v}_1\mathbf{v}_1' + \mathbf{v}_2\mathbf{v}_2' + \cdots + \mathbf{v}_p\mathbf{v}_p'.$$

de manera que se puede expresar la observación $\mathbf{x}_{(i)}$ en términos de la información captada por cada componente como

$$\mathbf{x}_{(i)}-\mu = \mathbf{I}_p(\mathbf{x}_{(i)}-\mu) = \mathbf{v}_1(\mathbf{v}_1'(\mathbf{x}_{(i)}-\mu))+\mathbf{v}_2(\mathbf{v}_2'(\mathbf{x}_{(i)}-\mu))+\cdots+\mathbf{v}_p(\mathbf{v}_p'(\mathbf{x}_{(i)}-\mu)),$$

donde $\mathbf{v}_j'(\mathbf{x}_{(i)} - \mu) = y_{ij}$ es la j-ésima componente principal evaluada en la i-ésima observación. De forma que

$$\mathbf{x}_{(i)} - \mu = y_{i1}\mathbf{v}_1 + y_{i2}\mathbf{v}_2 + \cdots + y_{ip}\mathbf{v}_p,$$

donde $i = 1, 2, \ldots, n$. Por lo tanto, se pueden emplear los últimos $p - q$ elementos de la combinación lineal $y_{i,p-m}\mathbf{v}_{p-m}+y_{i,p-m+1}\mathbf{v}_{p-m+1}+\cdots+y_{ip}\mathbf{v}_p$ como un residual o medida del error de reconstrucción de la observación $\mathbf{x}_{(i)}$ sobre los q primeros ejes principales. De manera que si se obtiene la norma del error de reconstrucción se obtiene la medida de distancia de la observación al origen en el nuevo espacio

$$d_i^2 = y_{i,p-m}^2 + y_{i,p-m+1}^2 + \cdots + y_{ip}^2,$$

para cada una de las observaciones $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. Si una observación tiene un valor demasiado grande de d_i^2 indicará un ajuste pobre de las primeras $p - m - 1$ componentes principales, lo que se puede asociar al hecho de que la medición es atípica con relación a la estructura de correlación de los datos (Diaz & Morales, 2012).

2.11. Cartas de control basadas en PCA

Controlar y mejorar la calidad se ha tornado en una estrategia muy importante para aquellas organizaciones que buscan una ventaja competitiva en los mercados actuales. Este tipo de estrategias son importantes no sólo en el sector industrial sino también en el sector de la salud, gubernamental y financiero. La gestión de calidad consiste en un conjunto de actividades

que involucra cada uno de los sectores de la empresa tanto los frentes administrativos como operativos. Ambos frentes son sensibles de capturar un alto flujo de datos que permiten generar indicadores claves de rendimiento y este tipo de datos permiten conocer que tan estable es un proceso respecto a su variabilidad común una vez se ha establecido el nivel de calidad deseado.

Con esto en mente, no es sorpresa que la gestión de calidad se haya apoyado en herramientas estadísticas para tener un control de corte cuantitativo sobre sus procesos. Las cartas de control son las principales herramientas estadísticas utilizadas con este fin. Fundamentalmente, en la producción de bienes las cartas de control analizan la información recolectada de pruebas, destructivas o no destructivas, que se realizan sobre un conjunto de piezas muestreadas aleatoriamente. En el caso de los procesos administrativos o de la prestación de servicios se utilizan datos claves de las distintas operaciones que reflejen la influencia directa de la variabilidad de los procesos. Cuando se producen bienes continuamente o cuando se monitorea constantemente un proceso es necesario recolectar datos para evaluar la capacidad y estabilidad de los procesos. Cuando un proceso se considera estable las variaciones presentadas en él son generadas por causas comunes que están siempre presentes mas no por fuentes externas que representarían un mayor nivel de variación (Jhonson & Wichern, 2007; Montgomery, 2012). De esta forma, el propósito de cualquier carta de control es identificar la ocurrencia de causas especiales de variación que lleven al proceso a un comportamiento inusual y que indiquen la necesidad de una corrección. Sin embargo, el uso de las cartas de control en un proceso también pueden sugerir acciones de mejora en él.

Las cartas de control suelen ser diseñadas en pro de controlar la variabilidad de variables cuantitativas o variables cualitativas. Las cartas de control enfocadas en variables cuantitativas se llaman *cartas de control para variables*, en el caso de las variables cualitativas reciben el nombre de *cartas de control para atributos*. Esta sección se enfocará en la presentación de las cartas de control de carácter cuantitativo multivariado, principalmente aquellas elaboradas a partir del uso del PCA. Para más información sobre la filosofía del control de calidad y las cartas de control univariadas consultar Montgomery (2012) y para información sobre las cartas de control multivariadas cuantitativas no basadas en PCA consultar el apéndice A5 y Johnson & Wichern (2007).

Desde un punto de vista estadístico las cartas de control se fundamentan en la construcción de regiones de confianza que permitan identificar comporta-

mientos atípicos en los datos. Esto se realiza por medio de la medición de la distancia de cada observación respecto a su media. Si una observación se encuentra a una distancia de la media mayor que la distancia definida por los límites de control se dice que la observación está fuera de control. Por otro lado, si una proporción de las observaciones muestreadas ubicadas por fuera de los límites de control es mayor que el nivel de significancia definido para el análisis, se dice que el proceso no es estable y que se presentan causas especiales de variabilidad. De manera que las cartas de control pueden ser utilizadas en diversos tipos de estudios fuera del contexto de la gestión de calidad, gracias a su capacidad para identificar comportamientos atípicos en los datos de una muestra. En este caso los límites de control definirán la distancia a partir de la cual el investigador podrá considerar que una observación tiene un comportamiento inusual. Estos límites de control se basan en la distancia de Mahalanobis construida a partir de las componentes principales obtenidas del PCA. Como se presenta en el Apéndice A5 los formatos de control basados en en datos cualitativos suelen cumplir con los formatos elipsoidal y T^2 utilizados comúnmente en muestras multivariadas. Se le recomienda al lector leer los apéndices A3, A4 y A5, en donde se presentan los elementos de estadística inferencial necesarios para la construcción de las cartas de control multivariadas.

Para que un proceso sea estable en el tiempo se requiere que las características medidas solo sean influenciadas por variaciones generadas a partir de causas comunes, de esta manera al almacenar la mayor cantidad de variación de los datos las dos primeras componentes deben ser estables e identificar los efectos de las fuentes de variación especial de forma adecuada, es decir, que sean lo suficientemente sensibles a los cambios de variabilidad no comunes presentados durante el proceso. Sin embargo, ya que las otras componentes pueden contener gran parte de la variabilidad de los datos es importante estudiarlas de igual manera. Por lo tanto, se puede monitorear la calidad a partir de las componentes principales en un proceso de dos partes. La primera parte consiste en la construcción de una carta de formato elipse para las dos primeras componentes y la segunda parte en la construcción de una carta de control formato T^2 para las componentes restantes.

2.11.1. Cartas para el análisis de estabilidad de una muestra de observaciones

Para construir los contornos de control para las componentes principales es posible reescribir la distancia de Mahalanobis de los datos originales en

términos de las componentes y posteriormente aplicar sus propiedades al nuevo problema.

Sea $\mathbf{X}$ una matriz aleatoria distribuida como una multinormal $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ con descomposición en valores propios asociada a la matriz de covarianza $\boldsymbol{\Sigma} = \mathbf{V}\Lambda\mathbf{V}'$ cuya inversa es $\boldsymbol{\Sigma}^{-1} = \mathbf{V}\Lambda^{-1}\mathbf{V}'$. La distancia de Mahalanobis para la observación $\mathbf{x}_{(i)}$ está dada por

$$(\mathbf{x}_{(i)} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_{(i)} - \boldsymbol{\mu}) = (\mathbf{x}_{(i)} - \boldsymbol{\mu})'\mathbf{V}\Lambda^{-1}\mathbf{V}'(\mathbf{x}_{(i)} - \boldsymbol{\mu}),$$

$$(\mathbf{x}_{(i)} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_{(i)} - \boldsymbol{\mu}) = \mathbf{y}'_{(i)}\Lambda^{-1}\mathbf{y}_{(i)},$$

$$(\mathbf{x}_{(i)} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_{(i)} - \boldsymbol{\mu}) = \sum_{j=1}^{p} \frac{y_{ij}^2}{\lambda_j},$$

de forma que

$$\frac{y_{i1}^2}{\lambda_1} + \frac{y_{i2}^2}{\lambda_2} + ... + \frac{y_{ip}^2}{\lambda_p} \sim \chi_p^2,$$

lo que presenta una ecuación de una elipse con sus ejes alienados a los ejes coordenados. Estadísticamente, esta ecuación refuerza la idea de que las componentes generadas por el PCA son no correlacionadas entre sí. En el Apéndice A.2 se demuestra que la distancia de Mahalanobis se distribuya como una chi cuadrado.

Ya que para este caso las componentes principales son generadas sobre la matriz de covarianza muestral $\mathbf{S}$, tiene sentido indicar que tanto los vectores y valores propios como la matriz de puntuaciones son estimadas. En estos casos es común recurrir al Teorema del Límite Central asumiendo un tamaño muestral grande para acotar tanto la carta de control de formato elipse como la carta de control de formato T^2 con un estadístico chi cuadrado. Para más información sobre el estadístico T^2 se le recomienda al lector leer el Apéndice A.2. Las cartas de formato elipsoidal y T^2 son fundamentalmente similares en cuanto a construcción, ya que ambas se basan en la distancia de Mahalanobis. Sin embargo, la primera solo analiza las dos primeras componentes, las cuales a pesar de conservar una gran cantidad de la variabilidad pueden no representar adecuadamente el comportamiento de los datos, por lo tanto, las cartas de control T^2 se utilizan para más de dos componentes y analizar la variabilidad de las $p - 2$ componentes restantes.

Cartas de formato elipse

Al ser una carta bivariada se seleccionan las dos primeras componentes para la construcción de la elipse de control, principalmente porque ambas componentes retienen la mayor cantidad de variabilidad de los datos. De manera que aquellos elementos $(\hat{y}_{i1}, \hat{y}_{i2})$ que se encuentren por fuera de la región de control construida con una confiabilidad del 95 %, definida por

$$\frac{\hat{y}_{i1}^2}{\hat{\lambda}_1} + \frac{\hat{y}_{i2}^2}{\hat{\lambda}_2} \leq \chi_2^2(0{,}05), \tag{2.31}$$

se consideran como fuera de control o no estables. Las cartas de control se caracterizan por ser una herramienta visual, en este caso consiste de una elipse construida por la ecuación 2.31 y la dispersión de los datos, todos los datos que se encuentren ubicados por fuera de la elipse se encuentra por fuera de la carta de control. El uso de la herramienta se ilustrará en la Sección 4.2.

Cartas de formato T^2

Para estudiar la variabilidad contenida en las $p-2$ variables restantes se define la distancia T_i^2 para cada dato por medio de

$$T_i^2 = \frac{\hat{y}_{i3}^2}{\hat{\lambda}_3} + \frac{\hat{y}_{i4}^2}{\hat{\lambda}_4} + ... + \frac{\hat{y}_{ip}^2}{\hat{\lambda}_p},$$

y se define el límite de control inferior, LCI, en

$$LCI = 0,$$

y el límite de control superior, LCS, en

$$LCS = \chi_{p-2}^2(0{,}05).$$

Visualmente se define una recta a la altura del límite de control superior y se representan las distancias T_i^2 como puntos en orden temporal, si los datos se tratan de observaciones independientes temporalmente no es necesario respetar un orden alguno. De esta forma, aquellas observaciones por encima del límite de control superior se consideran fuera de control. Esta herramienta es utilizada e ilustrada gráficamente en la aplicación presentada en la Sección 4.2.

Para información sobre la construcción de cartas de control para observaciones futuras basadas en PCA se le recomienda al lector acudir a Johnson et al. (2007).

Capítulo 3

Análisis de componentes principales no lineales - NLPCA

El PCA tiene dos limitaciones importantes para su aplicación. La primera, asume que las relaciones entre las variables son lineales, y la segunda, el método solo puede ser utilizado cuando todos los datos son cuantitativos. En áreas como las ciencias sociales y de la salud estos supuestos no siempre se pueden sostener (Linting et al., 2007). Para evitar estas limitaciones, autores como Guttman (1941), Kruskal (1965), Shepard (1966), Kruskal & Shepard (1974), Young et al. (1978), Gifi (1990), Michailidis & De Leeuw (1998), han desarrollado a lo largo de los años una alternativa al PCA conocida como Análisis de componentes principales no lineales (NLPCA - *Nonlineal Principal Components Analysis*). Este método cumple con los mismos objetivos del PCA lineal, reducción de dimensión, con la ventaja de que permite analizar niveles de medida mixtos (nominal, ordinal y numérico) que pueden no estar relacionados de forma lineal con las variables originales.

En el NLPCA, las componentes son combinaciones lineales del conjunto de variables originales, con la diferencia de que se pueden incluir variables cualitativas, que deben ser escaladas de la mejor forma para ser correctamente analizadas. Para cumplir con esta condición, el NLPCA se apoya en el escalamiento óptimo (OS - *Optimal Scaling*) y la técnica de mínimos cuadrados alternantes (ALS - *Alternating Least Squares*). Ambas técnicas permiten al NLPCA cuantificar las variables a medida que se obtienen las respectivas componentes. El OS obtiene información numérica a partir de los datos

cualitativos respetando las posibles relaciones no lineales que se encuentren entre las variables, siempre y cuando se sigan las restricciones planteadas por su naturaleza o nivel de análisis. Es decir, si la variable es numérica, ordinal, nominal simple o nominal múltiple. El nivel de análisis cuantitativo habla de aquellas características que pueden ser explicadas por un número real. El nivel de análisis ordinal por su parte permite agrupar a los individuos observados asignando una característica cualitativa a la vez que define una relación de orden entre ellos. Vale aclarar que no hay un distanciamiento claro en términos numéricos de los diferentes niveles de las variables. Por último, las variables nominales agrupan a aquellos individuos que comparten una característica en común sin definir una relación de orden entre ellos. El nivel de análisis nominal múltiple identifica el comportamiento promedio de todos aquellos individuos que pertenecen a una categoría de la variable analizada, mientras que el nivel nominal simple analiza a cada individuo según la categoría a la que pertenece. En el caso del nivel de análisis nominal múltiple es necesario el uso del análisis de homogeneidad por mínimos cuadrados alternantes (HOMALS - *Homogeneity Analysis For Alternating Least Squares*), también conocido como análisis de correspondencias múltiple, para el cálculo de una medida representativa de todos los individuos pertenecientes a cada categoría. Por otro lado, la técnica de mínimos cuadrados alternantes, ALS, permite alternar los pasos de cuantificación y de descomposición espectral que conforman el algoritmo de NLPCA de manera iterativa. El problema de reducción de PCA cuenta con una solución exacta proveniente de la descomposición espectral de la matriz de distancias de los datos originales lo que facilita su cálculo y análisis de propiedades, a diferencia del NLPCA que es una técnica computacional de naturaleza iterativa que, como veremos más adelante, no cuenta con una solución cerrada (Gifi, 1989; Linting et al., 2007).

De manera que se iniciará el capítulo abordando las metodologías de escalamiento óptimo y mínimos cuadrados alternantes con el fin de plantear las bases algorítmicas para el estudio de la metodología HOMALS de la cual se desprende el NLPCA en sus variantes basadas en la teoría de pérdida de reunión y teoría de pérdida conjunta, que basan su principio teórico en el PCA y HOMALS, terminando con la presentación de un conjunto de propuestas teóricas para la elaboración de cartas de control de calidad basadas en datos escalados óptimamente y el análisis de componentes no lineales.

3.1. Escalamiento óptimo

El escalamiento óptimo consiste en la asignación óptima de valores cuantitativos a escalas cualitativas, en otras palabras, convierte variables cualitativas a cuantitativas. Si se toma el modelo de regresión lineal como un caso particular, donde se desea predecir una variable respuesta a partir de un número de predictores, el escalamiento óptimo busca crear un modelo que permita relacionar las variables cualitativas con su respuesta cuantitativa por medio de funciones no lineales que admitan este tipo de variables predictoras. Las variables ordinales pasan por un proceso de regresión monotónica que permite obtener una función que respeta el orden de las categorías sin necesariamente conservar un comportamiento lineal. En el caso de las variables cuantitativas, pasan por un proceso de regresión lineal donde se conserva la relación de distancia entre las medidas, sin necesariamente ser los mismos valores, por lo que según el paquete estadístico que se utilice este paso puede o no ser evitado en el desarrollo del algoritmo. Cuando son variables que no conservan una jerarquía entre sus categorías y que funcionan como etiquetas no es necesario imponer una restricción de orden. De esta manera, para el nivel de medida nominal simple se propone realizar una transformación no monotónica de los datos donde no se presenta ningún tipo de restricción de orden entre ellos. Además de los niveles de medida presentados hasta ahora, hay un nivel de especial interés conocido como nominal múltiple, en el que cada categoría recibe una cuantificación representativa en cada una de las dimensiones del nuevo subespacio. Esta transformación permite, también, identificar los centroides de los datos pertenecientes a cada categoría (Gifi, 1990; Constantini et al., 2010; Linting et al., 2007). La principal diferencia del OS como técnica de escalamiento de variables cualitativas en comparación con el OneHot Encoder y el Ordinal Encoder, es que en el OS el proceso realizado a la hora de asignar las cuantificaciones a las variables consiste en resolver un conjunto de problemas de optimización definidos y restringidos según la naturaleza de la ellas, mientras que en el OneHot Encoder se construye una vector de 1 y 0 indicando la existencia de una medida en una categoría de una variable específica y en el Ordinal Encoder se asignan numeros naturales para indicar el orden jerárquico de las categorías de la variable. Este es un proceso computacionalmente más costoso que el uso de, por ejemplo, el OneHot Encoder ya que su primer consiste justamente en la obtención de una matriz de 1 y 0 a partir del uso de este codificador para posteriormente incluir cada una de las variables en un proceso de optimización iterativa distinto. Se podría esperar, por lo tanto, que las cuantificaciones obtenidas por este proceso sean mayormente

representativas del comportamiento natural de las variables que el que brindan los codificadores como OneHot Encoder y Ordinal, quienes siguen una simple regla de dedo para la asignación de las codificaciones (Bishop, 2006; Choong & Lee, 2017; Potdar, Pardawala & Pai, 2017; Gerón, 2019).

El manejo matemático de las transformaciones utilizadas por el OS se presentará en el apartado de análisis de componentes principales no lineales. Se debe aclarar que el escalamiento óptimo solo corresponde a un paso del proceso del NLPCA, el otro paso hace referencia a la aplicación de una descomposición en valores singulares (o valores propios) sobre una matriz de distancias de los datos cuantificados con lo que se construyen las componentes principales. Al tratarse de dos pasos que reducen sus respectivas funciones de pérdida de forma iterativa, una correspondiente a la del PCA y otra que varía según el nivel de análisis de cada variable cualitativa, se espera que se reduzca proporcionalmente la función de pérdida del problema general en cada iteración, es decir, la del NLPCA. Para ello se utiliza el algoritmo de optimización de mínimos cuadrados alternantes (Constantini et al., 2010; Linting et al., 2007).

3.2. Algoritmo de mínimos cuadrados alternantes

Cuando un problema de optimización no tiene soluciones de forma cerrada, al optimizar su función de pérdida en términos de una matriz objetivo $\mathbf{X}$, se suele utilizar un algoritmo iterativo de optimización. Este tipo de algoritmos permite actualizar los valores de la matriz $\mathbf{X}$ de manera que el valor de la función de pérdida es mejorado en cada paso, asegurando encontrar un óptimo local o global después de muchas iteraciones, con base a una cota definida previamente como criterio de parada. Estos algoritmos mejoran monotónicamente el valor de la función de pérdida con cada iteración. Por lo general, hay dos configuraciones de algoritmos iterativos: maximización iterativa y minimización de mínimos cuadrados alternantes por medio de relajación de bloques.

- **Maximización iterativa:** Consiste en encontrar una actualización para una matriz de parámetros $\mathbf{X}$ que maximice la función de pérdida $\sigma(\mathbf{X})$ por medio de la minimización de otra función que incluya a la matriz $\mathbf{X}$ y posteriormente definir una $\mathbf{X}_c = \mathbf{X}$ que contendrá el estimador de $\mathbf{X}$ para la iteración actual. La función que depende de $\mathbf{X}_c$ será denotada como $m_c(\mathbf{X})$.

Dado $\mathbf{X}_c$ como un estimador inicial de la matriz $\mathbf{X}$, se busca una función de maximización $m_c(\mathbf{X})$ relativamente simple tal que $\sigma(\mathbf{X}) \leq m_c(\mathbf{X})$, para todo $\mathbf{X}$, y $m_c(\mathbf{X}_c) = \sigma(\mathbf{X}_c)$. Al minimizar $m_c(\mathbf{X})$ es posible encontrar una nueva actualización de $\mathbf{X}$, denotada $\mathbf{X}_u$.

Siempre se busca un $\mathbf{X}_u$ tal que $m_c(\mathbf{X}_u) \leq m_c(\mathbf{X}_c)$ ya que se tendría $\sigma(\mathbf{X}_u) \leq m_c(\mathbf{X}_u)$ para todo $\mathbf{X}_u$ y $m_c(\mathbf{X}_c) = \sigma(\mathbf{X}_c)$ de lo que sigue $\sigma(\mathbf{X}_u) \leq m_c(\mathbf{X}_u) \leq m_c(\mathbf{X}_c) = \sigma(\mathbf{X}_c)$. Por lo tanto, $\mathbf{X}_u$ es un mejor estimador que $\mathbf{X}_c$ para $\mathbf{X}$, de forma que se asigna a $\mathbf{X}_u$ como el nuevo $\mathbf{X}_c$ minimizando así el valor de la función de pérdida $\sigma(\mathbf{X})$. Si hay restricciones impuestas sobre $\mathbf{X}$ deben ser utilizadas en la minimización de $m_c(\mathbf{X})$.

De forma que la clave está en encontrar una función de maximización simple $m_c(\mathbf{X})$, es decir, una función fácil de minimizar.

La configuración general de un algoritmo de maximización iterativa para minimizar $\sigma(\mathbf{X})$ es:

- **Paso 1**: Seleccionar un criterio de parada ε, inicializar la matriz $\mathbf{X}$ y llamarla $\mathbf{X}_c$.
- **Paso 2**: Calcular $\sigma_c = \sigma(\mathbf{X}_c)$.
- **Paso 3**: Calcular la actualización $\mathbf{X}_u$, usando la solución para minimizar/maximizar $m_c(\mathbf{X})$.
- **Paso 4**: Calcular $\sigma_u = \sigma(\mathbf{X}_u)$.
- **Paso 5**: Si $(\sigma_c - \sigma_u) > \varepsilon\sigma_c$ entonces reemplazar $\mathbf{X}_c$ por $\mathbf{X}_u$ y retornar al paso 2; sino considerar que el algoritmo convergió.

- **Mínimos cuadrados alternantes**: Cuando se busca minimizar una función sobre diferentes matrices de parámetros, es necesario actualizar una matriz a la vez mientras se mantienen las otras fijas. Con el fin de que cada actualización mejore el valor de la función y sea localmente optimizada sobre el conjunto de las matrices de los parámetros.

  Suponga que el problema es minimizar $\sigma(\mathbf{A}, \mathbf{B}, \mathbf{C})$ sobre $\mathbf{A}$, $\mathbf{B}$ y $\mathbf{C}$, entonces la configuración de un algoritmo de mínimos cuadrados alternantes es:

  - **Paso 1**: Seleccionar un criterio de parada ε e inicializar $\mathbf{A}$, $\mathbf{B}$ y $\mathbf{C}$ como $\mathbf{A}_c$, $\mathbf{B}_c$ y $\mathbf{C}_c$, respectivamente.
  - **Paso 2**: Calcular $\sigma_c = \sigma(\mathbf{A}_c, \mathbf{B}_c, \mathbf{C}_c)$.

- **Paso 3a**: Calcular la actualización $\mathbf{A}_u$, minimizando $\sigma(\mathbf{A}, \mathbf{B}_c, \mathbf{C}_c)$ sobre $\mathbf{A}$ y manteniendo fijas $\mathbf{B}_c$ y $\mathbf{C}_c$.
- **Paso 3b**: Calcular la actualización $\mathbf{B}_u$, minimizando $\sigma(\mathbf{A}_u, \mathbf{B}, \mathbf{C}_c)$ sobre $\mathbf{B}$ y manteniendo fijas $\mathbf{A}_u$ y $\mathbf{C}_c$.
- **Paso 3c**: Calcular la actualización $\mathbf{C}_u$, minimizando $\sigma(\mathbf{A}_u, \mathbf{B}_u, \mathbf{C})$ sobre $\mathbf{C}$ y manteniendo fijas $\mathbf{A}_u$ y $\mathbf{B}_u$.
- **Paso 4**: Calcular $\sigma_u = \sigma(\mathbf{A}_u, \mathbf{B}_u, \mathbf{C}_u)$.
- **Paso 5**: Si el índice de pérdida relativo $(\sigma_c - \sigma_u) > \varepsilon\sigma_c$, entonces reemplazar $\mathbf{A}_c$, $\mathbf{B}_c$ y $\mathbf{C}_c$ por $\mathbf{A}_u$, $\mathbf{B}_u$ y $\mathbf{C}_u$ respectivamente y retornar al paso 2; si no, se considera que el algoritmo converge.

Los únicos pasos complejos en este algoritmo son los pasos del 3a al 3c ya que dependerán mucho del contexto del problema. A veces, pueden manejarse con soluciones de minimización de forma cerrada, en otros casos sólo se puede disminuir la función de pérdida sobre el conjunto de parámetros en cuestión utilizando, por ejemplo, el enfoque de maximización iterativa, expuesto anteriormente. Por lo tanto, los problemas de optimización que involucran varias matrices de parámetros a menudo se pueden resolver mediante una combinación de mínimos cuadrados alternantes y el enfoque de maximización iterativa (Kiers, 2002).

La aplicación del algoritmo de mínimos cuadrados alternantes consiste en realizar los pasos presentados anteriormente, siempre ajustando las iteraciones 3a, 3b y 3c al contexto del problema abordado. Es usual ver el uso del escalamiento óptimo en conjunto con un algoritmo mínimos cuadrados alternantes lo cual recibe el nombre de ALSOS (*Alternating Least Squares With Optimal Scaling*) y es utilizado en gran variedad de aplicaciones, algunas de ellas son presentadas por Young & De Leeuw (1976), De Leeuw et al. (1976), Takane et al. (1977), Sands & Young (1980), Perault & Young (1980), Var der Burg & De Leeuw (1988), Crawford et al. (2014) y Kuroda, Mori, & Iizuka (2020). El ALS en el contexto de NLPCA es también utilizado en en el análisis de homogeneidad, en la sección 3.4 se presentará un planteamiento sencillo del algoritmo para resolver el problema HOMALS. Vale resaltar que en el proceso de NLPCA como en HOMALS es fundamental la construcción de la matriz indicadora de las variables categóricas que se introduce a continuación.

3.3. Matriz indicadora

La herramienta primordial en los procesos de cuantificación de variables cualitativas nominales y ordinales es la matriz indicadora $\mathbf{G}$. Esta matriz permite conocer la pertenencia de cada uno de los individuos a un cierto nivel de la variable categórica. Para esto se genera una matriz indicadora $\mathbf{G}_j$, por medio de una codificación OneHot, donde por cada una de las variables de la matriz de datos $\mathbf{H}$, cada matriz indicadora tendrá tantas columnas como categorías tenga la variable $\mathbf{h}_j$, y la asignación de un individuo se realiza por medio de ubicar un 1 en la columna que representa a dicho nivel y en la fila que representa al individuo y 0 en caso contrario.

Sea $\mathbf{H} = \begin{bmatrix}\mathbf{h}_1 & \mathbf{h}_2 & \dots & \mathbf{h}_p\end{bmatrix}$ una matriz de datos cualitativos de tamaño $n \times p$ donde el vector columna $\mathbf{h}_j$ contiene las n mediciones sobre los k_j niveles de la variable j. Para cuantificar el vector $\mathbf{h}_j$ es necesario el uso de la matriz indicadora $\mathbf{G}_j$ de tamaño $n \times k_j$,

$$\mathbf{G}_j = \begin{bmatrix} g_{j11} & g_{j12} & \cdots & g_{j1k_j} \\ g_{j21} & g_{j22} & \cdots & g_{j2k_j} \\ \vdots & \vdots & \ddots & \vdots \\ g_{jn1} & g_{jn2} & \cdots & g_{jnk_j} \end{bmatrix},$$

donde g_{jim} es el indicador de pertenencia de una medición i a un nivel m, donde $m = 1, 2, ..., k_j$, de la variable $\mathbf{h}_j$, de forma que

$$g_{jim} = \begin{cases} 1, & \text{si el objeto } i \text{ pertenece al nivel } k_j, \\ 0, & \text{si el objeto } i \text{ no pertenece al nivel } k_j. \end{cases}$$

De esta forma, la matriz indicadora $\mathbf{G}$ asociada a la matriz de datos cualitativos $\mathbf{H}$ es

$$\mathbf{G} = \begin{bmatrix}\mathbf{G}_1 & \mathbf{G}_2 & \dots & \mathbf{G}_p\end{bmatrix}.$$

Otra matriz importante en el estudio categórico es la matriz diagonal $\mathbf{D} = diag(\mathbf{G}'\mathbf{G})$ que está conformada por submatrices diagonales $\mathbf{D}_j$ cuyos elementos corresponden a las frecuencias absolutas de cada una de las k_j categorías de la variable $\mathbf{h}_j$. A menudo cuando es necesario el uso del producto $\mathbf{G}'\mathbf{G}$ se opta por utilizar la matriz $\mathbf{D}$, este será el caso durante el desarrollo del capítulo desde que no se indique lo contrario. Para una vista más profunda sobre este tema consultar Gifi (1990).

3.4. Análisis de homogeneidad - HOMALS

El análisis de homogeneidad por mínimos cuadrados alternantes (HOMALS), también conocido como análisis de correspondencia múltiple, es en un sentido estricto, una técnica de análisis de datos puramente categóricos con una cierta función objetivo y una forma particular para encontrar la solución óptima por medio de un algoritmo de ALS (Gifi, 1990; Michailidis et al. 1998). En esencia, HOMALS juega dos papeles de gran importancia en el desarrollo del NLPCA, el primero es que presenta una idea directa de cómo cuantificar variables de origen categórico, y el segundo es que su función objetivo será primordial a la hora de abordar el modelo de centroides sobre el cual se realizan las cuantificaciones de las variables a un nivel de análisis nominal múltiple.

También podemos decir que HOMALS busca determinar cuantificaciones óptimas de los individuos y de las categorías de cada variable j de modo tal que se maximice la homogeneidad entre ellas, es decir, que sus categorías estén separadas entre sí lo más posible. Así, si están en la misma categoría las observaciones serán cercanas entre sí y si están en distintas categorías se encontrarán lo más alejadas posible. Al igual que un proceso de conglomerados en el que se utilice una medida de similitud dentro de los datos agrupados o una de disimilitud entre las agrupaciones.

Con el fin de obtener una componente unidimensional que resuma la información contenida en las relaciones no lineales de las variables de una matriz $\mathbf{X} \in \mathbb{R}^{n\times p}$ se define una transformación no lineal de la forma $\phi(\mathbf{x}_j) \in \mathbb{R}^n$. De manera que se busca minimizar la función de pérdida

$$\sigma(\mathbf{y}_1, \mathbf{v}_1, \phi) = p^{-1} \sum_{j=1}^{p} ||\mathbf{y}_1 - v_{j1}\phi_j(\mathbf{x}_j)||_F^2, \tag{3.1}$$

la cual define una ponderación diferencial entre las variables transformadas $\phi(\mathbf{x}_j)$, donde el vector $\mathbf{v}_1 \in \mathbb{R}^p$ es el vector de cargas y $\mathbf{y}_1 \in \mathbb{R}^n$ es el vector de puntuaciones que resume la máxima cantidad de información de las variables en el nuevo subespacio. Este último vector se encuentra sujeto comúnmente a las restricciones

$$\mathbf{y}_1'\mathbf{y}_1 = n, \tag{3.2}$$

$$\mathbf{y}_1'\mathbf{1}_n = 0, \tag{3.3}$$

donde la ecuación 3.2 hace referencia a la norma y la ecuación 3.3 a la centralidad del vector. Estas restricciones facilitan la interpretación de la solución obtenida al establecer el origen como punto de centralidad de los datos y normalizar la variabilidad de las puntuaciones en las diferentes dimensiones. De forma que el producto $\phi(\mathbf{X})\mathbf{v}_1$, que no se encuentra restringido, recolecta toda la información sobre las relaciones no lineales presentes en los datos. Vale resaltar que el uso de restricciones como $||v_{j1}\phi_j(\mathbf{x}_j)||^2 = 1$ o $||\phi_j(\mathbf{x}_j)||^2 = 1$ son igualmente válidas pero es necesario ir con especial cuidado al momento de interpretar los resultados. Además, es en este último caso cuando los pesos v_{j1} podrán ser verdaderamente diferenciados de $\phi_j(\mathbf{x}_j)$. A pesar de las ideas expuestas anteriormente la minimización de $\sigma(\mathbf{y}_1, \mathbf{v}_1, \phi)$ es trivial bajo estas restricciones. Al admitir todas las transformaciones no lineales, el espacio de las transformaciones admisibles crece hasta un espacio n dimensional y, por lo tanto, cualquier cuantificación de las variables será suficiente para obtener un ajuste perfecto aunque con una solución trivial. El comportamiento de estas transformaciones es similar al de una aplicación Kernel. Sin embargo, se busca que la base de esta transformación tenga un rango menor que n, para que sea posible restringir las transformaciones, por ejemplo, a polinomios de bajo orden o incluso splines. Ya que el únicamente imponer restricciones sobre las cuantificaciones no será suficientes para evitar soluciones triviales (Gifi, 1990).

Para el caso en el que las variables de la matriz de datos sean cualitativas categóricas o numéricas discretizadas en un número de categorías limitadas, es decir, se trabaje con una matriz $\mathbf{H}$, es razonable pensar en la construcción de una matriz indicadora $\mathbf{G}_j$ para cada variable $\mathbf{h}_j$. Los k_j vectores en $\mathbf{G}_j$ se toman como base para generar un subespacio k_j dimensional a partir del espacio n dimensional de todas las posibles transformaciones no lineales que se tienen inicialmente. De forma que la solución trivial de la función de pérdida 3.1 se previene gracias al uso del conjunto de restricciones presentadas anteriormente junto con el uso del subespacio generado por las matrices indicadoras. De esta forma, la incorporación de la base $\mathbf{G}_j$ implica que la transformación no lineal puede ser escrita como $\phi_j(\mathbf{h}_j) = \mathbf{G}_j\mathbf{w}_{j1}$, para algún vector de coeficientes $\mathbf{w}_{j1}$ de tamaño k_j. Entonces la función de pérdida 3.1 para HOMALS se convierte en

$$\sigma(\mathbf{y}_1, \mathbf{w}_1) = p^{-1}\sum_{j=1}^{p} ||\mathbf{y}_1 - \mathbf{G}_j\mathbf{w}_{j1}||_F^2, \tag{3.4}$$

bajo las restricciones

$$\mathbf{y}_1'\mathbf{y}_1 = n,$$
$$\mathbf{y}_1'\mathbf{1}_n = 0,$$

donde el rol de las cuantificaciones $\mathbf{w}_1' = \begin{bmatrix}\mathbf{w}_{11}' & \mathbf{w}_{21}' & \cdots & \mathbf{w}_{p1}'\end{bmatrix}$, de tamaño $K = \sum_{j=1}^{p} k_j$, es comparable pero no igual al rol de los pesos $\mathbf{v}_1$ en el PCA. Esto no significa que $\mathbf{w}_1$ no pueda ser interpretado como cargas; más bien, se selecciona para enfatizar que en contraste a $\mathbf{v}_1$ el vector $\mathbf{w}_1$ son pesos en un espacio de mayor dimensionalidad, abarcando todas las posibles transformaciones disponibles en este subespacio.

Si se busca obtener múltiples componentes es posible generalizar la función de pérdida 3.4 por medio de la obtención de una matriz de puntajes $\mathbf{Y}$ de tamaño $n \times q$ y una matriz de transformación no lineal $\mathbf{W}$ de tamaño $K \times q$ donde $q \leq p$, de forma que la función de pérdida a minimizar es

$$\sigma(\mathbf{Y}, \mathbf{W}) = p^{-1} \sum_{j=1}^{p} ||\mathbf{Y} - \mathbf{G}_j\mathbf{W}_j||_F^2, \tag{3.5}$$

bajo las restricciones

$$\mathbf{Y}'\mathbf{Y} = n\mathbf{I}_q, \tag{3.6}$$
$$\mathbf{Y}'\mathbf{1}_n = \mathbf{0}, \tag{3.7}$$

donde la primera condición, ecuación 3.6, además de restringir la longitud cuadrada de los vectores, indica que es necesario que cada componente sea ortogonal entre sí. La función de pérdida 3.5 será la función de pérdida básica de donde se desprenderán las funciones de pérdida para los demás tipos de datos. A esta función de pérdida se le suele conocer como función de pérdida de HOMALS, ya que un problema con $\mathbf{W}_j$ sin restricciones y $\mathbf{Y}$ restringida puede ser minimizado por el programa HOMALS.

De forma general se puede afirmar que la función de pérdida 3.4 es utilizada en problemas asociados con una sola componente y la función de pérdida 3.5 para múltiples soluciones o componentes. Como se resaltó en un principio, para solucionar este tipo de problemas de optimización se hace uso de un algoritmo basado en mínimos cuadrados alternantes donde los estimadores de puntuaciones y cuantificaciones usados en cada paso iterativo de optimización se obtienen por medio de las soluciones óptimas de las respectivas funciones de pérdida. Para esto es necesario obtener las derivadas parciales de la función de pérdida respecto al vector $\mathbf{y}_1$ o la matriz $\mathbf{Y}$ de puntuaciones

y el vector $\mathbf{w}_1$ o matriz $\mathbf{W}$ de cuantificaciones. De forma que las ecuaciones que permiten estimar tanto las puntuaciones como las cuantificaciones en el problema definido por la ecuación 3.4 son

$$\mathbf{y}_1 = p^{-1}\mathbf{G}\mathbf{w}_1, \tag{3.8}$$

$$\mathbf{w}_1 = \mathbf{D}^{-1}\mathbf{G}'\mathbf{y}_1. \tag{3.9}$$

Para esto recordemos que la norma de Frobenius $||\cdot||_F$ aplicada sobre la función 3.4 trabaja como la norma dos $||\cdot||_2$ ya que en efecto la dimensión de la operación resultante es $n \times 1$. Con esta aclaración, la solución de 3.4 para el vector de puntuaciones $\mathbf{y}_1$ se obtiene derivando parcialmente respecto a este vector. Como se verá a continuación

$$\frac{\partial\sigma(\mathbf{y}_1,\mathbf{w}_1)}{\partial\mathbf{y}_1} = \frac{\partial}{\partial\mathbf{y}_1}\sum_{j=1}^{p}(\mathbf{y}_1 - \mathbf{G}_j\mathbf{w}_{j1})'(\mathbf{y}_1 - \mathbf{G}_j\mathbf{w}_{j1}),$$

$$\frac{\partial\sigma(\mathbf{y}_1,\mathbf{w}_1)}{\partial\mathbf{y}_1} = \frac{\partial}{\partial\mathbf{y}_1}\sum_{j=1}^{p}\mathbf{y}_1'\mathbf{y}_1 - 2\frac{\partial}{\partial\mathbf{y}_1}\sum_{j=1}^{p}\mathbf{w}_{j1}'\mathbf{G}_j'\mathbf{y}_1 + \frac{\partial}{\partial\mathbf{y}_1}\sum_{j=1}^{p}\mathbf{w}_{j1}'\mathbf{D}_j\mathbf{w}_{j1},$$

$$\frac{\partial\sigma(\mathbf{y}_1,\mathbf{w}_1)}{\partial\mathbf{y}_1} = \sum_{j=1}^{p}2\mathbf{y}_1 - 2\sum_{j=1}^{p}\mathbf{G}_j\mathbf{w}_{j1},$$

igualando la derivada parcial al vector cero de tamaño n se obtiene

$$\frac{\partial\sigma(\mathbf{y}_1,\mathbf{w}_1)}{\partial\mathbf{y}_1} = \mathbf{0},$$

de donde el vector óptimo $\mathbf{y}_1$ tiene la forma

$$\mathbf{y}_1 = p^{-1}\sum_{j=1}^{p}\mathbf{G}_j\mathbf{w}_{j1},$$

$$\mathbf{y}_1 = p^{-1}\mathbf{G}\mathbf{w}_1.$$

La solución de 3.4 para el caso del vector de codificaciones $\mathbf{w}_{j1}$ se tiene

$$\frac{\partial\sigma(\mathbf{y}_1,\mathbf{w}_1)}{\partial\mathbf{w}_{j1}} = \frac{\partial}{\partial\mathbf{w}_{j1}}\sum_{l=1}^{p}(\mathbf{y}_1 - \mathbf{G}_l\mathbf{w}_{l1})'(\mathbf{y}_1 - \mathbf{G}_l\mathbf{w}_{l1}),$$

$$\frac{\partial\sigma(\mathbf{y}_1,\mathbf{w}_1)}{\partial\mathbf{w}_{j1}} = \frac{\partial}{\partial\mathbf{w}_{j1}}\sum_{l=1}^{p}\mathbf{y}_1'\mathbf{y}_1 - 2\frac{\partial}{\partial\mathbf{w}_{j1}}\sum_{l=1}^{p}\mathbf{w}_{l1}'\mathbf{G}_l'\mathbf{y}_1 + \frac{\partial}{\partial\mathbf{w}_{j1}}\sum_{l=1}^{p}\mathbf{w}_{l1}'\mathbf{D}_l\mathbf{w}_{l1},$$

$$\frac{\partial\sigma(\mathbf{y}_1,\mathbf{w}_1)}{\partial\mathbf{w}_{j1}} = -2\mathbf{G}_j'\mathbf{y}_1 + 2\mathbf{D}_j\mathbf{w}_{j1},$$

igualando la derivada parcial al vector cero de tamaño k_j se llega a

$$\frac{\partial \sigma(\mathbf{y}_1, \mathbf{w}_1)}{\partial \mathbf{w}_{j1}} = \mathbf{0},$$

y se obtiene la solución

$$\mathbf{w}_{j1} = \mathbf{D}_j^{-1}\mathbf{G}_j'\mathbf{y}_1,$$
$$\mathbf{w}_1 = \mathbf{D}^{-1}\mathbf{G}'\mathbf{y}_1.$$

Para el problema de soluciones múltiples basta con derivar parcialmente la ecuación 3.5 respecto a la matriz $\mathbf{Y}$ y la matriz $\mathbf{W}_j$. En el caso de $\mathbf{Y}$

$$\frac{\partial \sigma(\mathbf{Y}, \mathbf{W})}{\partial \mathbf{Y}} = \frac{\partial}{\partial \mathbf{Y}} \sum_{j=1}^{p} tr((\mathbf{Y} - \mathbf{G}_j\mathbf{W}_j)'(\mathbf{Y} - \mathbf{G}_j\mathbf{W}_j)),$$
$$\frac{\partial \sigma(\mathbf{Y}, \mathbf{W})}{\partial \mathbf{Y}} = \frac{\partial}{\partial \mathbf{Y}} \sum_{j=1}^{p} (tr(\mathbf{Y}'\mathbf{Y}) - 2tr(\mathbf{W}_j'\mathbf{G}_j'\mathbf{Y}) + tr(\mathbf{W}_j'\mathbf{D}_j\mathbf{W}_j)),$$
$$\frac{\partial \sigma(\mathbf{Y}, \mathbf{W})}{\partial \mathbf{Y}} = \sum_{j=1}^{p} 2\mathbf{Y} - 2\sum_{j=1}^{p} \mathbf{G}_j\mathbf{W}_j,$$

igualando la derivada parcial a la matriz cero de tamaño $n \times q$ se obtiene

$$\frac{\partial \sigma(\mathbf{Y}, \mathbf{W})}{\partial \mathbf{Y}} = \mathbf{0},$$

$$\mathbf{Y} = p^{-1} \sum_{j=1}^{p} \mathbf{G}_j\mathbf{W}_j,$$

$$\mathbf{Y} = p^{-1}\mathbf{G}\mathbf{W}. \tag{3.10}$$

A partir de la 3.5, para el caso de la matriz de cuantificaciones $\mathbf{W}_j$ se tiene

$$\frac{\partial \sigma(\mathbf{Y}, \mathbf{W})}{\partial \mathbf{W}_j} = \frac{\partial}{\partial \mathbf{W}_j} \sum_{l=1}^{p} tr(\mathbf{Y}'\mathbf{Y}) - 2\frac{\partial}{\partial \mathbf{W}_j} \sum_{l=1}^{p} tr(\mathbf{W}_l'\mathbf{G}_l'\mathbf{Y}) + \frac{\partial}{\partial \mathbf{W}_j} \sum_{l=1}^{p} tr(\mathbf{W}_l'\mathbf{D}_l\mathbf{W}_l),$$

$$\frac{\partial \sigma(\mathbf{Y}, \mathbf{W})}{\partial \mathbf{W}_j} = -2\mathbf{G}_j'\mathbf{Y} + 2\mathbf{D}_j\mathbf{W}_j.$$

igualando la derivada parcial al vector cero de tamaño k_j se obtiene

$$\frac{\partial \sigma(\mathbf{Y}, \mathbf{W})}{\partial \mathbf{W}_j} = \mathbf{0},$$

$$\mathbf{W}_j = \mathbf{D}_j^{-1}\mathbf{G}_j'\mathbf{Y},$$
$$\mathbf{W} = \mathbf{D}^{-1}\mathbf{G}'\mathbf{Y}. \tag{3.11}$$

Donde la ecuación 3.11 recibe el nombre de *primer principio del centroide* en la literatura y hace referencia a que la cuantificación de la categoría obtenida por medio de esta ecuación se encuentra en el centroide de las puntuaciones de los objetos (Benzécri, 1973). Mientras que la ecuación 3.10 muestra que el puntaje de un objeto es el promedio de las cuantificaciones de las categorías a las que pertenece (Michailidis et al., 1998). Una vez obtenidas las soluciones estacionarias del problema de optimización se utiliza el algoritmo de mínimos cuadrados alternantes con el fin de disminuir iterativamente las funciones de pérdida 3.4 y 3.5. Para el caso de la función de pérdida 3.5 con restricción $\mathbf{Y}'\mathbf{Y} = n\mathbf{I}_q$ se utilizan las soluciones estacionarias 3.10 y 3.11, en el siguiente algoritmo:

Algoritmo 1. HOMALS.

1. Inicialización aleatoria de la matriz de cuantificación: $\tilde{\mathbf{W}}$.
2. Actualización de la matriz de puntajes: $\tilde{\mathbf{Y}} \leftarrow p^{-1}\mathbf{G}\tilde{\mathbf{W}}$.
3. Restricciones:
   3.a. Centralización: $\hat{\mathbf{Y}} \leftarrow (\mathbf{I}_n - n^{-1}\mathbf{1}_n'\mathbf{1}_n)\tilde{\mathbf{Y}}$.
   3.b. Normalización: $\mathbf{Y}^+ \leftarrow \sqrt{n}\hat{\mathbf{Y}}(\hat{\mathbf{Y}}'\hat{\mathbf{Y}})^{-1/2}$.
4. Actualización de la matriz de cuantificación: $\mathbf{W}^+ \leftarrow \mathbf{D}^{-1}\mathbf{G}'\mathbf{Y}^+$.
5. Test de convergencia: Se define un valor de ε, por lo regular 10^{-6}, tal que cuando se cumpla la desigualdad $|\sigma(\tilde{\mathbf{Y}}, \tilde{\mathbf{W}}) - \sigma(\mathbf{Y}^+, \mathbf{W}^+)| < \varepsilon\sigma(\tilde{\mathbf{Y}}, \tilde{\mathbf{W}})$, se considera que el algoritmo ha convergido adecuadamente con las soluciones óptimas $\mathbf{Y}^* \leftarrow \mathbf{Y}^+$ y $\mathbf{W}^* \leftarrow \mathbf{W}^+$. En caso contrario, se retoma el algoritmo desde el paso 2 con $\tilde{\mathbf{W}} \leftarrow \mathbf{W}^+$.

En el paso 3 del algoritmo presentado es posible realizar la normalización de $\tilde{\mathbf{Y}}$ por medio de una descomposición de Gram-Schmidt, la cual permite descomponer la matriz $\tilde{\mathbf{Y}}$ como el producto de una matriz ortogonal $\mathbf{Z}$ de tamaño $n \times p$ y una matriz triangular superior $\mathbf{T}$ de tamaño $p \times p$, asignando a la matriz de puntajes $\mathbf{Y}^+$ la matriz $\sqrt{n}\mathbf{Z}$ con el objetivo de cumplir con la condición $\mathbf{Y}^{+\prime}\mathbf{Y}^+ = n\mathbf{I}_q$. De manera que el paso 3 se puede realizar como:

3. Restricciones:

3.a. Centralización: $\hat{\mathbf{Y}} \leftarrow (\mathbf{I}_n - n^{-1}\mathbf{1}_n'\mathbf{1}_n)\tilde{\mathbf{Y}}$.

3.b. Descomposición de Gram-Schmidt: $(\mathbf{Z}, \mathbf{T}) \leftarrow \hat{\mathbf{Y}}$.

3.c. Selección de base: $\mathbf{Y}^+ \leftarrow \sqrt{n}\mathbf{Z}$.

Esta solución es conocida en la literatura como la solución al programa HOMALS (Gifi, 1990; De Leeuw, 1984; De Leeuw & Van Rijckevorsel, 1980).

Es posible describir el comportamiento de la función de pérdida 3.5 de HOMALS en término de vectores y valores singulares como se muestra en Gifi (1990) y Michailidis et al. (1998) a pesar de que la solución HOMALS no permite obtener una solución cerrada al problema. Para esto se utiliza la matriz de correlación inducida por la matriz indicadora y la matriz diagonal de frecuencias absolutas

$$\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{G}'\mathbf{G}\mathbf{D}^{-1/2},$$

obteniendo la descomposición en valores singulares de $\mathbf{G}\mathbf{D}^{-1/2}$ de la forma

$$\mathbf{G}\mathbf{D}^{-1/2} = \mathbf{U}\Lambda^{1/2}\mathbf{V}'.$$

Ya que el programa HOMALS alcanza la solución estacionaria en las matrices $\mathbf{W}^*$ y $\mathbf{Y}^*$, que debe cumplir con la condición de ortogonalidad impuesta por $\mathbf{Y}^{*\prime}\mathbf{Y}^* = n\mathbf{I}_q$ se puede pensar en asignar

$$\mathbf{Y}^* = \sqrt{n}\mathbf{U}_q, \tag{3.12}$$

y de forma similar

$$\mathbf{W}^* = \sqrt{n}\mathbf{D}^{-1/2}\mathbf{V}_q\Lambda_q^{1/2}. \tag{3.13}$$

Reescribiendo la función de pérdida 3.5 se tiene

$$\sigma(\mathbf{Y}, \mathbf{W}) = tr(\mathbf{Y}'\mathbf{Y}) - p^{-1}tr(\mathbf{W}'\mathbf{D}\mathbf{W}),$$

reemplazando 3.12 y 3.13

$$\sigma(\mathbf{Y}^*, \mathbf{W}^*) = n(tr(\mathbf{U}_q'\mathbf{U}_q)) - np^{-1}(tr(\Lambda_q^{1/2}\mathbf{V}_q'\mathbf{D}^{-1/2}\mathbf{D}\mathbf{D}^{-1/2}\mathbf{V}_q\Lambda_q^{1/2})),$$

$$\sigma(\mathbf{Y}^*, \mathbf{W}^*) = n\left(q - \sum_{j=1}^{q}\frac{\lambda_j}{p}\right). \tag{3.14}$$

De forma que el ajuste de la función objetivo 3.5 se puede explicar en términos de los valores propios de la matriz de correlación.

Para el caso en el que solo sea de interés describir el comportamiento de la función de pérdida en una sola componente basta con redefinir las ecuaciones 3.12 y 3.13 en términos del vector columna de $\mathbf{U}$ y de $\mathbf{V}$ respectivo, de forma que

$$\mathbf{y}_j^* = \sqrt{n}\mathbf{u}_j, \tag{3.15}$$

y de forma similar

$$\mathbf{w}_j^* = \lambda_j^{1/2}\sqrt{n}\mathbf{D}^{-1/2}\mathbf{v}_j, \tag{3.16}$$

donde $j = 1, 2, ..., q$. Reescribiendo la ecuación 3.4 se obtiene

$$\sigma(\mathbf{y}_j, \mathbf{w}_j) = \mathbf{y}_j'\mathbf{y}_j - p^{-1}\mathbf{w}_j'\mathbf{D}\mathbf{w}_j,$$

de forma que al reemplazar en las ecuaciones 3.15 y 3.16 se tiene

$$\sigma(\mathbf{y}_j^*, \mathbf{w}_j^*) = n\mathbf{u}_j'\mathbf{u}_j - np^{-1}\lambda_j\mathbf{v}_j'\mathbf{D}^{-1/2}\mathbf{D}\mathbf{D}^{-1/2}\mathbf{v}_j,$$

$$\sigma(\mathbf{y}_j^*, \mathbf{w}_j^*) = n\left(1 - \frac{\lambda_j}{p}\right). \tag{3.17}$$

Vale resaltar que a pesar de que se logra describir el comportamiento de la función de pérdida por medio de los valores propios de la matriz de correlaciones, ecuaciones 3.17 y 3.14, solo se tendrá un ajuste perfecto cuando se pretenda representar la variabilidad de una variable en una dimensión, es decir para $q = 1$, $p = 1$ y, por lo tanto, $\lambda_1 = 1$.

Un resultado de especial interés es el hecho de que se puede relacionar la pérdida total con la medida de discriminación de ajuste de cada variable cuantificada por medio de la relación $\mathbf{q}_j = \mathbf{G}_j\mathbf{w}_j$ y las puntuaciones de la componente $\mathbf{y}_j$. En este caso, la correlación entre ambas cantidades juega el papel de medida de discriminación

$$\rho^2_{\mathbf{y}_j,\mathbf{q}_j} = \frac{Cov(\mathbf{y}_j, \mathbf{q}_j)^2}{Var(\mathbf{y}_j)Var(\mathbf{q}_j)} = \frac{(\mathbf{y}_j'\mathbf{q}_j)^2}{(\mathbf{y}_j'\mathbf{y}_j)(\mathbf{q}_j'\mathbf{q}_j)} = \frac{(\mathbf{y}_j'\mathbf{G}_j\mathbf{w}_j)^2}{(\mathbf{y}'\mathbf{y})(\mathbf{w}_j'\mathbf{G}_j'\mathbf{G}_j\mathbf{w}_j)},$$

ya que $\mathbf{D}_j\mathbf{w}_j = \mathbf{G}_j'\mathbf{y}_j$ se llega a

$$\rho^2_{\mathbf{y}_j,\mathbf{q}_j} = \frac{(\mathbf{w}_j'\mathbf{D}_j\mathbf{w}_j)^2}{n(\mathbf{w}_j'\mathbf{D}_j\mathbf{w}_j)} = n^{-1}(\mathbf{w}_j'\mathbf{D}_j\mathbf{w}_j),$$

si se reemplazan las soluciones óptimas estacionarias 3.15 y 3.16 se obtiene que la correlación puede ser expresada en términos de los valores propios de la matriz de correlaciones de $\mathbf{G}$,

$$\rho^2_{\mathbf{y}_j^*,\mathbf{q}_j^*} = \lambda_j, \tag{3.18}$$

reemplazando la ecuación 3.18 en 3.14 es posible expresar el ajuste de HOMALS en términos de las medidas de discriminación

$$\sigma(\mathbf{Y}^*, \mathbf{W}^*) = n \left(q - \sum_{j=1}^{q} \frac{\rho^2_{\mathbf{y}_j^*, \mathbf{q}_j^*}}{p} \right).$$

De donde se obtiene una forma práctica de calcular el ajuste del método HOMALS.

Por último se presentan algunas de las propiedades de la solución de HOMALS en la j-ésima dimensión se presentan a continuación (Michailidis et al., 1998):

- Las cuantificaciones de las categorías y los puntajes de los objetos son representados en un mismo espacio.
- Un punto de categoría es el centroide de los objetos pertenecientes a esa categoría, una consecuencia directa de 3.11.
- Objetos con el mismo patrón de respuesta (perfiles idénticos) reciben puntuaciones idénticas de los objetos, esto se ve a partir de 3.10. En general, la distancia entre dos puntos de objetos está relacionada a la similitud entre sus perfiles.
- Una variable discrimina mejor en la medida que sus puntos de categoría están más alejados (referente a 3.4).
- Si una categoría aplica únicamente a un objeto sencillo, entonces el punto del objeto y el punto de la categoría deben coincidir.
- Objetos con un único perfil deberán ser localizados lejos del origen del espacio conjunto, mientras objetos con perfiles similares al promedio deberán ser localizados cerca al origen (consecuencia directa de la propiedad anterior).
- Las cuantificaciones de las categorías de cada variable tienen una suma ponderada sobre categorías igual a cero. Esto sigue de emplear la normalización de los objetos.
- Las soluciones de HOMALS son anidadas. Esto significa que una solución q_2-dimensional tal que $q_2 > q_1$ compartirá las primeras q_1 dimensiones con la solución q_1-dimensional. Gracias a esto, las soluciones

para las subsecuentes dimensiones son ordenadas. Lo que significa que la primera dimensión tiene el mayor valor propio. La segunda dimensión tiene el máximo valor propio sujeto a la restricción de que $\mathbf{y}_1$ y $\mathbf{y}_2$ sean no correlacionadas y así sucesivamente.

- Los puntajes de los objetos son no correlacionados en las dimensiones subsecuentes. Sin embargo, las cuantificaciones de las categorías no necesitan ser correlacionadas necesariamente; de hecho, sus patrones de correlaciones podrían ser más bien impredecibles.

3.5. Análisis de componentes principales no métricos

Como se resaltó con anterioridad las dos principales limitaciones del PCA lineal es que asume que las relaciones entre variables son lineales y su implementación sólo es posible si todas las variables se trabajan bajo un nivel de análisis numérico (Linting et al., 2007). Es de aquí de donde parte la necesidad de utilizar una transformación específica para cada nivel de análisis específico. Debe mantenerse en mente que los diferentes niveles de análisis implican diferentes requerimientos. En el caso de un nivel de análisis nominal, el único requerimiento es que los individuos que pertenezcan a la misma categoría, en la variable original, obtengan el mismo valor de cuantificación. Este requerimiento es el más débil en cuanto a restricciones en el NLPCA. En el caso de un análisis de nivel ordinal, las cuantificaciones de los individuos deberían adicionalmente respetar el orden de las categorías originales; esto es, que la cuantificación de una categoría debe ser siempre menor o igual a la cuantificación de la categoría que tiene un mayor rango en los datos originales. El nivel de análisis numérico por su parte busca que tanto el orden de los datos originales como el distanciamiento entre los datos en una escala métrica se respete proporcionalmente. Sin embargo, si uno desea analizar las relaciones no lineales entre variables numéricas, se debe escoger un nivel de análisis no numérico (Meulman et al., 2004; Linting et al., 2007; Gifi, 1990).

A continuación se resaltan algunas diferencias fundamentales entre PCA y NLPCA. Para realizar la transformación de las variables y explorar las relaciones entre ellas, el NLPCA debe ejecutar tanto el proceso de escalamiento óptimo como la estimación del modelo de PCA simultáneamente por medio de la minimización de una función de pérdida. Esto se logra con

el uso de un algoritmo de mínimos cuadrados alternantes (ALS). Como el NLPCA utiliza un algoritmo de mínimos cuadrados con un paso de escalamiento óptimo y otro de proyección en el nuevo subespacio, las matriz de distancias, sea de covarianza o de correlación, no se calcula directamente de las variables originales sino de las variables escaladas óptimamente. Consecuentemente, opuesto al PCA, esta matriz en el NLPCA no es fija ya que depende del tipo de cuantificación escogida para cada una de las variables. Por lo tanto, en contraste a la solución del PCA, la solución del NLPCA no es cerrada respecto a la función de pérdida sino calculada iterativamente a partir de los datos, usando el proceso de escalamiento óptimo para cuantificar las variables de acuerdo con los niveles de análisis escogidos a la vez que se realiza su descomposición espectral. En consecuencia el objetivo del escalamiento óptimo es optimizar las propiedades de la matriz de distancias de las variables cuantificadas, ya que el uso de codificadores clásicos, como los codificadores OneHot y Ordinal, llevaría a una solución equivalente a la obtenida por la aplicación de un PCA lineal sobre el conjunto de variables codificadas por medio de ellos. Específicamente, maximizando los primeros q valores propios, donde q indica el número de componentes que son escogidas en el análisis (Gifi, 1990).

Al igual que el PCA, el NLPCA permite representar las cargas de las variables como vectores usando cada una de las entradas como coordenadas en el espacio de las componentes principales. Este hecho ha dado el nombre de modelo vectorial al modelo de NLPCA que se encuentra altamente influenciado por la filosofía del PCA, cuyo marco conceptual recibe el nombre de *teoría de la pérdida conjunta*. Por otro lado, también es posible representar el comportamiento promedio de los individuos de cada una de las categorías de una variable en el espacio de componentes haciendo uso de la teoría de HOMALS, este modelo es llamado el modelo de centroides y su marco conceptual se conoce como *teoría de la pérdida de reunión* (Meulman et al., 2004). En las siguientes dos secciones se definirán a detalle ambas teorías.

3.5.1. Teoría de la pérdida conjunta

Sea $\mathbf{Q}$ una matriz de tamaño $n \times p$ de datos óptimamente escalados provenientes de la matriz $\mathbf{H}$ de variables ordinales, categóricas y numéricas, una función de pérdida obtenida a partir de un análisis de PCA lineal es

$$\sigma_J(\mathbf{Q}, \mathbf{Y}, \mathbf{V}) = p^{-1} \sum_{j=1}^{q} ||\mathbf{q}_j - \mathbf{Y}\mathbf{v}_{(j)}||_F^2, \tag{3.19}$$

con $q \leq p$; donde la matriz de cargas $\mathbf{V}$ tiene dimensión $p \times q$, de forma que el vector $\mathbf{v}_{(j)}$ es un vector fila escrito como columna y la matriz $\mathbf{Y}$ es la matriz de puntuaciones de dimensión $n \times q$. Donde la matriz de cuantificaciones, la matriz de puntuaciones y la matriz de cargas son centradas y normalizadas, de manera que el problema está sujeto al siguiente conjunto de restricciones

$$Cov(\mathbf{Q}) = Cor(\mathbf{Q}), \tag{3.20}$$

$$\mathbf{Q}'\mathbf{1}_p = \mathbf{0}, \tag{3.21}$$

$$\mathbf{Y}'\mathbf{Y} = n\mathbf{I}_p, \tag{3.22}$$

$$\mathbf{Y}'\mathbf{1}_p = \mathbf{0},$$

$$\mathbf{V}'\mathbf{V} = \Lambda, \tag{3.23}$$

$$\mathbf{V}'\mathbf{1}_p = \mathbf{0},$$

en 3.23, la matriz Λ contiene los valores propios asociados a $\mathbf{Q}'\mathbf{Q}$. Definiendo $\sigma_J(\mathbf{Q}, *, *)$ como el mínimo de $\sigma_J(\mathbf{Q}, \mathbf{Y}, \mathbf{V})$ respecto $\mathbf{Y}$ y $\mathbf{V}$ con $\mathbf{Q}$ fijo, se puede escribir la función de pérdida 3.19 en términos de los valores propios λ_j^2 de la matriz de correlaciones ρ de $\mathbf{Q}$, gracias al teorema de Eckart-Young, presentado en la sección 2.5. De forma que los $p - q$ valores propios más pequeños indican el ajuste de la solución de la misma forma que en el PCA lineal, es decir

$$\sigma_J(\mathbf{Q}, *, *) = p^{-1} \sum_{j=q+1}^{p} \lambda_j^2,$$

donde el objetivo es de igual forma minimizar la distancia obtenida entre la matriz de datos estimada y la matriz original. De manera que la función de pérdida puede tomar valores entre

$$0 \leq \sigma_J(\mathbf{Q}, *, *) \leq 1 - q/p,$$

con $\sigma_J(\mathbf{Q}, *, *) = 0$ si y sólo si $rank(\mathbf{R}) = q$ y $\sigma_J(\mathbf{Q}, *, *) = 1 - q/p$ si y sólo si $\mathbf{R}$ es la matriz identidad, para más información consultar Gifi (1990). Si se desea una sola dimensión en el nuevo subespacio, es decir $q = 1$. se busca el $\mathbf{w}_j$ tal que el valor propio mayor de la matriz de correlación de los vectores $\mathbf{q}_j = \mathbf{G}_j\mathbf{w}_j$ es maximizado. Si $q = p - 1$, entonces se busca minimizar el valor propio menor de la matriz de correlación. Para valores intermedios de q se debe distinguir entre la dimensionalidad q, que es definida en términos de los números de columnas de $\mathbf{Y}$ y $\mathbf{V}$ en la función de pérdida, y el número de soluciones sucesivas r. En el análisis de componentes se tiene $q \geq 1$ y $r = 1$, de donde se minimiza σ_J para un valor de q dado. Por lo tanto, el

análisis de componentes da una cuantificación simple a diferencia del análisis de homogeneidad que obtiene múltiples cuantificaciones $r \geq 1$ para una sola dimensión, $q = 1$. También es posible combinar las dos aproximaciones y construir una técnica con ambos $q \geq 1$ y $r \geq 1$. Esto equivale a aplicar el análisis de componentes r veces, cada vez imponiendo una condición de ortogonalidad adicional (Gifi, 1990).

El hecho de incluir la cuantificación de los datos categóricos en la minimización de la función de pérdida $\sigma_J(\mathbf{Q}, *, *)$ sobre $\mathbf{q}_j$, que pertenece a un subespacio definido por la proyección de los vectores de datos sobre un cono C_j con las restricciones $\mathbf{q}_j'\mathbf{1}_n = 0$ y $\mathbf{q}_j'\mathbf{q}_j = 1$, significa que se transforman o cuantifican las variables de forma tal que la suma de los $p-q$ valores propios de $\mathbf{R}$ más pequeños se minimiza o, equivalentemente, tal que la suma de los q valores propios más grandes se maximiza. Los conos C_j son subespacios donde los datos deben ser proyectados con el fin de obtener las cuantificaciones óptimas que mejor se comporten para su escala de medición. Para datos categóricos, los conos C_j son subespacios definidos por

$$C_j = \{\mathbf{q}_j | \mathbf{q}_j = \mathbf{G}_j \mathbf{w}_j\},$$

como visto en la metodología de HOMALS. Si los datos son numéricos y se requiere que todas las transformaciones sean lineales, entonces el cono se define como

$$C_j = \{\mathbf{q}_j | \mathbf{q}_j = \alpha_j \mathbf{h}_j + \beta_j\},$$

lo cual representa una regresión lineal respecto a los datos originales con las cuantificaciones obtenidas. Sin pérdida de generalidad, se puede suponer que $||\mathbf{h}_j||^2 = 1$ y $\mathbf{h}_j'\mathbf{1} = 0$ para todo j. Como se requiere que $||\mathbf{q}_j||^2 = 1$ y $\mathbf{q}_j'\mathbf{1} = 0$, se obtiene que

$$\sigma_J(\mathbf{Q}, \mathbf{Y}, \mathbf{V}) = p^{-1} \sum_{j=1}^{p} ||\mathbf{h}_j - \mathbf{Y}\mathbf{v}_{(j)}||_F^2 = p^{-1} ||\mathbf{H} - \mathbf{Y}\mathbf{V}'||_F^2. \tag{3.24}$$

Aquí no hay libertad de escoger $\mathbf{Q}$ diferente de $\mathbf{H}$ y, consecuentemente, el análisis equivale a calcular los valores y vectores propios de la matriz de correlación de $\mathbf{H}$, es decir, realizar un PCA sobre $\mathbf{H}$. Lo cual es apoyado por lo presentado en la sección 2.8 donde se presenta de forma explícita la similitud entre ambas funciones de pérdida, ecuaciones 2.23 y 3.24.

Si se trabaja con datos ordinales es posible usar matrices indicadoras en un sentido obvio, utilizándola en la escritura de la función de pérdida σ_J siendo

función de las cuantificaciones categóricas $\mathbf{w}_j$. Entonces la función pérdida 3.19 se torna

$$\sigma_J(\mathbf{W}, \mathbf{Y}, \mathbf{V}) = p^{-1} \sum_{j=1}^{p} ||\mathbf{G}_j \mathbf{w}_j - \mathbf{Y}\mathbf{v}_{(j)}||_F^2,$$

la cual debe ser minimizada bajo las condiciones

$$\mathbf{1}_n' \mathbf{G}_j \mathbf{w}_j = 0, \tag{3.25}$$

$$\mathbf{w}_j' \mathbf{D}_j \mathbf{w}_j = 1, \tag{3.26}$$

$$\mathbf{w}_j \in C_j, \tag{3.27}$$

donde C_j es un cono en un espacio k_j dimensional con $k_j << n$, donde el objetivo es conservar el orden de los datos tal que

$$q_{ij} = \sum_{s=1}^{q} y_{is} v_{js},$$

restringido a

$$h_{ij} > h_{kj} \rightarrow q_{ij} > q_{kj},$$
$$h_{ij} = h_{kj} \rightarrow q_{ij} = q_{kj},$$

con $i, k = 1, 2, ..., n$ y $j = 1, 2, ..., p$. Esto implica que los datos deben ser cuantificados por medio de una regresión monotónica la cual respeta la jerarquía de las categorías, sin necesariamente usar una relación lineal. De Leeuw & Takane (1976) y De Leeuw & Van Rijckervorsel (1980) discuten un proceso enfocado en un sistema de niveles de medida que puede ser usado para definir muchos tipos diferentes de conos C_j.

El modelo planteado hasta ahora es utilizado por programas como MINFAEX, PRINQUAL, PRINCIPALS y PRINCALS encontrados en paquetes estadísticos como: SPSS, SAS y R (Meulman et al., 2004; Mair, 2018; Meulman, 1998; Kuhfeld, 1990; Wold et al., 1987; Abdi & Williams, 2010; Dunteman, 1989; Gifi, 1985; Gifi, 1989). Su principio de funcionamiento se expone en el algoritmo 2, con especial énfasis en el conjunto de pasos necesarios para que se cumplan las diversas restricciones impuestas sobre la solución del problema. Vale aclarar, además, que cada uno de los programas nombrados anteriormente tienen sus particularidades lógicas que ayudan al proceso computacional y que, por lo tanto, pueden diferir en mayor o menor medida del presentado a continuación.

Algoritmo 2. NLPCA - Datos categóricos simples.

0. Inicialización aleatoria de los vectores de cuantificaciones y la matriz de puntuaciones: $\tilde{\mathbf{w}}_j$, $\tilde{\mathbf{Y}}$.

1. Actualización de la matriz de datos escalados para $j = 1, 2, ..., p$:

   1.a. Datos numéricos: $\tilde{\mathbf{q}}_j \leftarrow \mathbf{h}_j$.

   1.b. Datos nominales simples y ordinales: $\tilde{\mathbf{q}}_j \leftarrow \mathbf{G}_j\tilde{\mathbf{w}}_j$.

   1.1. Centralización: $\mathbf{Q}^\circ \leftarrow (\mathbf{I}_n - n^{-1}\mathbf{1}_n'\mathbf{1}_n)\tilde{\mathbf{Q}}$.

   1.2. Normalización: $\hat{\mathbf{Q}} \leftarrow \sqrt{n}\mathbf{Q}^\circ diag(\mathbf{Q}^{\circ\prime}\mathbf{Q}^\circ)^{-1/2}$.

2. Actualización de la matriz de pesos y puntuaciones:

   2.1. Descomposición en valores propios: $(\mathbf{V}_e, \Lambda) \leftarrow eig(\hat{\boldsymbol{\rho}})$.

   2.2. Actualización de la matriz de cargas: $\hat{\mathbf{V}} \leftarrow \mathbf{V}_e\Lambda^{1/2}$.

   2.3. Actualización de la matriz de puntuaciones: $\hat{\mathbf{Y}} \leftarrow \hat{\mathbf{Q}}\mathbf{V}_e\Lambda^{-1/2}$.

3. Actualización de la matriz de cuantificación:

   3.a. Datos nominales simples: $\hat{\mathbf{w}}_j \leftarrow \mathbf{D}_j^{-1}\mathbf{G}_j'\hat{\mathbf{Y}}\hat{\mathbf{v}}_{(j)}$.

   3.b. Datos ordinales: Regresión monotónica sobre $\hat{\mathbf{w}}_j$ con $\tilde{\mathbf{q}}_j$ como variable respuesta y $\mathbf{h}_j$ como variable regresora.

   3.1. Centralización: $\mathbf{w}_j^\circ \leftarrow (\mathbf{I}_q - \mathbf{G}_j^{-1}\mathbf{1}_q\mathbf{1}_q'\mathbf{G}_j)\hat{\mathbf{w}}_j$.

   3.2. Normalización: $\tilde{\mathbf{w}}_j \leftarrow \mathbf{w}_j^\circ(\mathbf{w}_j^{\circ\prime}\mathbf{D}_j\mathbf{w}_j^\circ)^{-1/2}$.

4. Test de convergencia.

Se debe resaltar además que para el funcionamiento del algoritmo es necesario definir la cantidad de componentes q que se desean hallar, este valor definirá la dimensión columna de $\mathbf{Y}_q$ y $\mathbf{V}_q$. Esto se verá reflejado en el capítulo 4 durante el desarrollo de la aplicación.

El algoritmo anterior inicia estimando todos los vectores $\mathbf{w}_j$ y la matriz $\mathbf{Y}$ de forma que satisfagan sus respectivas restricciones de normalidad y centralidad. Entonces, en el primer paso se calcula $\mathbf{q}_j = \mathbf{G}_j\mathbf{w}_j$ con el fin de cuantificar los datos y posteriormente minimizar en el segundo paso la función de pérdida que plantea la descomposición espectral de los datos. Se

procura hacer especial énfasis en este punto, de manera que a continuación se presentará detalladamente una solución a este paso desde el punto de vista del PCA.

Para iniciar se asocia la función de pérdida del PCA lineal con las restricciones típicas a este paso, es decir

$$\sigma(\mathbf{Y}, \mathbf{V}) = ||\mathbf{Q} - \mathbf{Y}\mathbf{V}'||_F^2,$$

sujeto a

$$\mathbf{V}_e'\mathbf{V}_e = \mathbf{I}_p,$$

de donde se obtiene la solución cerrada

$$\frac{\mathbf{Q}'\mathbf{Q}}{n} = \mathbf{V}_e\Lambda\mathbf{V}_e', \tag{3.28}$$

por medio de la descomposición en valores y vectores propios. Sin embargo, una de las diferencias claves encontradas entre el PCA y el NLPCA se presenta en este punto, ya que la matriz de cargas se obtiene escalando cada vector propio por su respectivo valor singular

$$\mathbf{V} = \mathbf{V}_e\Lambda^{1/2}, \tag{3.29}$$

de manera que la matriz de puntuaciones obtiene la forma

$$\tilde{\mathbf{Y}} = \mathbf{Q}\mathbf{V}_e\Lambda^{1/2},$$

la cual debe ser normalizada para cumplir con la condición $\mathbf{Y}'\mathbf{Y} = n\mathbf{I}_p$.

La ecuación 3.28 se puede escribir como

$$\mathbf{V}_e'\mathbf{Q}'\mathbf{Q}\mathbf{V}_e = n\Lambda, \tag{3.30}$$

al calcular la matriz de distancias de la matriz de puntuaciones se obtiene la expresión

$$\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}} = \Lambda^{1/2}\mathbf{V}_e'\mathbf{Q}'\mathbf{Q}\mathbf{V}_e\Lambda^{1/2},$$

utilizando el resultado 3.30 para simplificar la expresión se tiene

$$\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}} = n\Lambda^2,$$

y la normalización de $\tilde{\mathbf{Y}}$ es

$$\mathbf{Y} = \sqrt{n}\tilde{\mathbf{Y}}(\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}})^{-1/2},$$

$$\mathbf{Y} = \mathbf{Q}\mathbf{V}_e\Lambda^{-1/2}. \tag{3.31}$$

Por esta razón, si en el NLPCA todas las variables son numéricas, las soluciones del NLPCA y PCA se encuentran estrechamente relacionadas. Ya que como se presentó en la ecuación 3.24 si los datos son numéricos no es necesario realizar transformación alguna. Sin embargo, las restricciones definidas para la matriz de cargas y la matriz de puntucaciones impiden que las soluciones sean exactamente iguales a las del PCA lineal tradicional.

En el caso de variables nominales simples, la selección de las cuantificaciones auxiliares $\tilde{\mathbf{w}}_j$ se obtienen por la minimización de la función objetivo

$$\sigma_J(\mathbf{w}_j, \mathbf{Y}, \mathbf{v}_{(j)}) = ||\mathbf{G}_j\mathbf{w}_j - \mathbf{Y}\mathbf{v}_{(j)}||_F^2,$$

respecto a $\mathbf{w}_j$ sin restricciones, para un $\mathbf{Y}$ y $\mathbf{V}$ fijos. De esta forma, derivando respecto a $\mathbf{w}_j$ e igualando al vector cero se obtiene

$$\tilde{\mathbf{w}}_j = \mathbf{D}_j^{-1}\mathbf{G}_j'\mathbf{Y}\mathbf{v}_{(j)}.$$

Para el caso de variables ordinales se realiza una regresión monotónica para buscar a $\mathbf{w}_j$ de forma que se respete el orden de las categorías de las variables $\mathbf{h}_j$ en las variables cuantificadas $\mathbf{q}_j$. Se regresa al primer paso y se repite el proceso hasta que el algoritmo converja.

3.5.2. Teoría de pérdida de reunión

Desde el punto de vista del sistema Gifi el NLPCA es visto como un análisis de homogeneidad en el cual se imponen restricciones de orden uno a los vectores de cuantificación, para más información sobre el sistema Gifi consultar Michailidis et al. (1998). Además, la teoría de pérdida conjunta indica que el análisis de componentes principales tiene el inconveniente de que las cuantificaciones múltiples deben ser calculadas sucesivamente a diferencia del análisis de homogeneidad en el que pueden ser obtenidas simultáneamente. El uso de la teoría de pérdida de reunión no tiene esta desventaja. Para verlo se debe replantear la función de pérdida del NLPCA

$$\sigma_M(\mathbf{Y}, \mathbf{W}) = p^{-1}\sum_{j=1}^{p} ||\mathbf{Y} - \mathbf{G}_j\mathbf{W}_j||_F^2, \tag{3.32}$$

sujeta a

$$\mathbf{Y}'\mathbf{1}_p = \mathbf{0},$$

$$\mathbf{Y}'\mathbf{Y} = n\mathbf{I}_p,$$

y en dónde implícitamente $\mathbf{Y} = \mathbf{QV}$ y a su vez $\mathbf{Q} = \mathbf{GW}$ con las restricciones impuestas a la función de pérdida 3.19.

La principal ventaja del uso de la teoría de pérdida de reunión es obtener cuantificaciones multidimensionales para una misma variable, a diferencia del PCA y la teoría de pérdida conjunta que solo permiten obtener cuantificaciones unidimensionales. Para verlo es necesario entender que el paso de cuantificaciones múltiples a simples se logra por medio de imponer restricciones de rango uno a las matrices de cuantificación. Estas restricciones son escritas como

$$\mathbf{W}_j = \mathbf{w}_j\mathbf{v}'_{(j)},$$

con los requerimientos adicionales

$$\mathbf{1}'_n\mathbf{D}_j\mathbf{w}_j = 0,$$

$$\mathbf{w}'_j\mathbf{D}'_j\mathbf{w}_j = 1,$$

$$\mathbf{w}_j \in C_j.$$

De manera que el no uso de las restricciones de orden uno, es decir restringir las cuantificaciones a ser unidimensionales, implica que en teoría no es posible acceder a la información de los vectores de cargas de las componentes principales. La solución obtenida de este análisis es

$$\mathbf{Q}_j = \mathbf{G}_j\mathbf{W}_j.$$

Esta ecuación permite plantear el método de escalamiento óptimo de las variables definidas como nominales múltiples siguiendo la función de pérdida

$$\sigma_M(\mathbf{Q}, \mathbf{W}) = p^{-1}\sum_{j=1}^{p} ||\mathbf{Q} - \mathbf{G}_j\mathbf{W}_j||^2_F,$$

bajo la generalización de las restricciones 3.20 y 3.21. Estas soluciones son los centroides de las individuos que pertenecen a cada una de las categorías correspondientes, como se verá durante el desarrollo del capítulo 4.

Es de esperarse que las teorías de pérdida conjunta y de reunión se encuentren conectadas de alguna manera. En este caso es posible mostrar que ambas funciones de pérdida se relacionan por medio de

$$\sigma_M(\mathbf{Y}, \mathbf{W}, \mathbf{V}) = \sigma_J(\mathbf{Y}, \mathbf{W}, \mathbf{V}) + n(p-1), \tag{3.33}$$

se aprecia entonces que minimizar $\sigma_M(\mathbf{Y}, \mathbf{W}, \mathbf{V})$ o $\sigma_J(\mathbf{Y}, \mathbf{W}, \mathbf{V})$ son problemas equivalentes con las mismas soluciones. Para esto se parte de la ecuación 3.32 y se reescribe para todas las variables como

$$\sigma_M(\mathbf{Y}, \mathbf{W}, \mathbf{V}) = p^{-1} \sum_{j=1}^{p} ||\mathbf{Y} - \mathbf{G}_j \mathbf{w}_j \mathbf{v}'_{(j)}||_F^2,$$

y al utilizar las propiedades de la norma de Frobenius se tiene

$$\sigma_M(\mathbf{Y}, \mathbf{W}, \mathbf{V}) = nq + p^{-1} \sum_{j=1}^{p} \mathbf{v}'_{(j)} \mathbf{v}_{(j)} - 2p^{-1} \sum_{j=1}^{p} \mathbf{v}_{(j)} \mathbf{Y}' \mathbf{G}_j \mathbf{w}_j,$$

en donde se han usado las condiciones $\mathbf{Y}'\mathbf{Y} = n\mathbf{I}_q$, por lo tanto $tr(\mathbf{Y}'\mathbf{Y}) = nq$, y $\mathbf{w}_j \mathbf{D}_j \mathbf{w}_j = 1$. Similarmente, la función de pérdida conjunta se puede escribir como

$$\sigma_J(\mathbf{Y}, \mathbf{W}, \mathbf{V}) = n + p^{-1} \sum_{j=1}^{p} \mathbf{v}'_{(j)} \mathbf{v}_{(j)} - 2p^{-1} \sum_{j=1}^{p} \mathbf{v}_{(j)} \mathbf{Y}' \mathbf{G}_j \mathbf{w}_j,$$

de donde se obtiene fácilmente la ecuación 3.33.

La primera ventaja de usar σ_M es que se pueden imponer las condiciones $\mathbf{W}_j = \mathbf{w}_j \mathbf{v}'_j$ para algunas variables y no para otras. Ya que una variable nominal puede ser simple o múltiple y, por lo tanto, tener asociado un vector $\mathbf{w}_j$ o una $\mathbf{W}_j$ de cuantificaciones asociada, pero no ambas. Si se imponen condiciones de orden uno para para todas las variables, entonces se estaría minimizando σ_J y, por lo tanto, se realiza un análisis de componentes principales y es posible obtener tanto $\mathbf{w}_j$ como $\mathbf{v}_j$. Si por el contrario no se impone ninguna restricción a las variables se estaría realizando un análisis de homogeneidad y solo es posible obtener $\mathbf{W}_j$ (Gifi, 1990). De forma que puede ser interesante mezclar ambas opciones e imponer cuantificaciones simples a unas variables y cuantificaciones múltiples a otras. Una modificación del algoritmo 2 es presentado a continuación incluyendo el proceso de cuantificación múltiple en él.

Algoritmo 3. NLPCA - Datos categóricos múltiples.

0. Inicialización aleatoria de los vectores y matrices de cuantificaciones y la matriz de puntajes: $\tilde{\mathbf{w}}_j$, $\tilde{\mathbf{W}}_j$, $\tilde{\mathbf{Y}}$.

1. Actualización de la matriz de datos escalados:

1.a. Datos numéricos: $\tilde{\mathbf{q}}_j \leftarrow \mathbf{h}_j$.

1.b. Datos nominales simples y ordinales: $\tilde{\mathbf{q}}_j \leftarrow \mathbf{G}_j\tilde{\mathbf{w}}_j$.

1.c. Datos nominales múltiples: $\tilde{\mathbf{Q}}_j \leftarrow \mathbf{G}_j\tilde{\mathbf{W}}_j$.

1.1. Centralización: $\mathbf{Q}^\circ \leftarrow (\mathbf{I}_n - n^{-1}\mathbf{1}_n'\mathbf{1}_n)\tilde{\mathbf{Q}}$.

1.2. Normalización: $\hat{\mathbf{Q}} \leftarrow \sqrt{n}\mathbf{Q}^\circ diag(\mathbf{Q}^{\circ\prime}\mathbf{Q}^\circ)^{-1/2}$.

2. Actualización de la matriz de pesos y puntuaciones:

2.1. Descomposición en valores propios: $(\mathbf{V}_e, \Lambda) \leftarrow eig(\hat{\boldsymbol{\rho}})$.

2.2. Actualización de la matriz de cargas: $\hat{\mathbf{V}} \leftarrow \mathbf{V}_e\Lambda^{1/2}$.

2.3. Actualización de la matriz de puntuaciones: $\hat{\mathbf{Y}} \leftarrow \hat{\mathbf{Q}}\mathbf{V}_e\Lambda^{-1/2}$.

3. Actualización de la matriz de cuantificación:

3.a. Datos nominales simples: $\hat{\mathbf{w}}_j \leftarrow \mathbf{D}_j^{-1}\mathbf{G}_j'\tilde{\mathbf{Y}}\hat{\mathbf{v}}_{(j)}$.

3.b. Datos ordinales: Regresión monotónica sobre $\hat{\mathbf{w}}_j$ con $\tilde{\mathbf{q}}_j$ como variable respuesta y $\mathbf{h}_j$ como variable regresora.

3.b.1. Centralización: $\mathbf{w}_j^\circ \leftarrow (\mathbf{I}_q - \mathbf{G}_j^{-1}\mathbf{1}_q\mathbf{1}_q'\mathbf{G}_j)\hat{\mathbf{w}}_j$.

3.b.2. Normalización: $\tilde{\mathbf{w}}_j \leftarrow \mathbf{w}_j^\circ(\mathbf{w}_j^{\circ\prime}\mathbf{D}_j\mathbf{w}_j^\circ)^{-1/2}$.

3.c. Datos nominales múltiples: $\tilde{\mathbf{W}}_j \leftarrow \mathbf{D}_j^{-1}\mathbf{G}_j'\tilde{\mathbf{Y}}$.

3.c.1. Centralización: $\mathbf{W}_j^\circ \leftarrow (\mathbf{I}_q - \mathbf{G}_j^{-1}\mathbf{1}_q\mathbf{1}_q'\mathbf{G}_j)\hat{\mathbf{W}}_j$.

3.c.2. Normalización: $\tilde{\mathbf{W}}_j \leftarrow \mathbf{W}_j^\circ(\mathbf{W}_j^{\circ\prime}\mathbf{D}_j\mathbf{W}_j^\circ)^{-1/2}$.

4. Test de convergencia.

En el algoritmo anterior la matriz de datos cuantificados $\mathbf{Q}_j$ corresponde a los datos que fueron tratados como nominales múltiples y es utilizada únicamente en el cálculo de la matriz de correlaciones de los datos cuantificados. Sin embargo, es necesario aclarar cual dimensión de la matriz de datos cuantificados se va a utilizar ya que cuenta con tantas columnas como dimensiones tenga la solución. De igual forma, hay que resaltar que a la hora de presentar los resultados no se asocian en la matriz de cargas $\mathbf{V}$ ni en la matriz de puntajes $\mathbf{Y}$ columnas correspondientes a las variables nominales múltiples, ya que $\mathbf{W}_j$ representa el centroide de cada una de las categorías pertenecientes a cada variable nominal múltiple, a pesar de que sí se usan durante los cálculos del algoritmo.

Es de vital importancia resaltar que el NLPCA puede no solo reducir dimensiones sino también crearlas a diferencia del PCA lineal. Principalmente por el uso de la teoría de HOMALS a la hora de cuantificar las variables nominales múltiples, ya que a cada variable nominal múltiple le corresponderá una matriz de cuantificación de tamaño $n \times k_j$, es decir, su matriz de cuantificación tendrá tantas columnas como categorías tenga la variable a transformar. Para el caso de los demás niveles de análisis las cuantificaciones corresponden a vectores y por lo tanto cada variable será representada por una sola dimensión. De manera que a la hora de realizar el paso correspondiente a la descomposición espectral se le recomienda al investigador incluir únicamente el primer vector de cuantificación de cada variable nominal múltiple para construir la matriz $\mathbf{Q}$ que se utilizará en el algoritmo, esto con el fin de no incluir mayor información que la contenida en los datos originales representada en variabilidad. Sin embargo, algoritmos como el incluido en *SPSS* puede utilizar esta información extra y de esta manera aumentar las dimensiones originales del estudio.

3.5.3. Correlación entre las componentes y las variables

El PCA estudia la interdependencia de las variables y las transformaciones no lineales maximizan el promedio de ésta. La optimalidad de esta propiedad se puede conocer estudiando la forma en que la variabilidad de las variables originales se distribuyen entre las componentes. Cuando las variables obtienen una transformación ordinal o nominal, la técnica maximiza la suma de los q valores propios más grandes de la matriz de correlación entre las variables transformadas. Al igual que en el PCA la suma de los valores propios es igual al total de la varianza contenida por las variables transformadas. Sin embargo, en el NLPCA la obtención de cada valor propio se realiza por medio del cálculo de la norma de los vectores de cargas correspondiente a cada componente, a diferencia del PCA en el que provienen de las varianzas de los vectores de puntuaciones (Linting, 2007). De forma que la matriz de cargas es en especial importante para el análisis de componentes principales no lineales ya que permite conocer diferentes medidas de ajuste entre las variables transformadas y las componentes. Todo esto gracias a la restricción de normalización impuesta sobre la matriz de cargas, ecuación 3.23.

Por lo tanto, un primer resultado importante para conocer la correlación entre las componentes y las variables proviene del hecho que la norma al cuadrado de un vector de cargas es igual al valor propio asociado a esa

componente. Es decir,

$$||\mathbf{v}_j||^2 = \lambda_j.$$

Para esto, se fija un vector $\mathbf{a}_l' = \begin{bmatrix} 0 & \dots & 0 & 1 & 0 & \dots & 0 \end{bmatrix}$ que indique el vector requerido de la matriz, y recordando el resultado 3.29, se tiene

$$||\mathbf{v}_j||^2 = ||\mathbf{V}\mathbf{a}_j||^2 = ||\mathbf{V}_e\Lambda^{1/2}\mathbf{a}_j||^2 = ||\lambda_j^{1/2}\mathbf{v}_{ej}||^2 = |\lambda_j^{1/2}|^2||\mathbf{v}_{ej}||^2 = \lambda_j.$$

Un segundo resultado de igual interés para el investigador es el coeficiente de correlación que indica el ajuste entre las componentes y las variables transformadas. En este caso el coeficiente de correlación entre una componente y una variable es igual al coeficiente del vector de cargas que las asocia,

$$\rho_{\mathbf{y}_j,\mathbf{q}_l} = v_{lj}.$$

Utilizando de nuevo el vector $\mathbf{a}_l$ tal que $\mathbf{q}_l = \mathbf{Q}\mathbf{a}_l$ se tiene que

$$Cov(\mathbf{y}_j, \mathbf{q}_l) = \frac{\mathbf{v}_j'\mathbf{Q}'\mathbf{Q}\mathbf{a}_l}{n} = \mathbf{a}_j'\Lambda^{-1/2}\mathbf{V}_e'\mathbf{V}_e\Lambda\mathbf{V}_e'\mathbf{a}_l,$$

$$Cov(\mathbf{y}_j, \mathbf{q}_l) = \lambda_j^{1/2}\mathbf{a}_j'\mathbf{V}_e'\mathbf{a}_l = \lambda_j^{1/2}v_{elj} = v_{lj},$$

con lo cual es posible calcular el coeficiente de correlación

$$\rho_{\mathbf{y}_j,\mathbf{q}_l} = \frac{Cov(\mathbf{y}_j, \mathbf{q}_l)}{\sqrt{Var(\mathbf{y}_j)}\sqrt{Var(\mathbf{q}_l)}} = \frac{v_{lj}}{(1)(1)} = v_{lj}.$$

De manera que la correlación entre las nuevas componentes y las variables cuantificadas viene dada por la carga que le corresponde a cada vector en la construcción de la componente, esto es así principalmente porque los vectores de carga no están normalizados a diferencia del PCA donde es una restricción clave en su construcción.

3.5.4. Algunos hechos importantes

En esta sección se pretende resaltar algunos hechos importantes en el NLPCA y remarcar algunas diferencias importantes con el PCA lineal, presentados en Constantini et al. (2010), Lintning et al.(2007) y Meulman et al. (2004):

- Vale resaltar que el mayor nivel de libertad en la cuantificación de una variable es el nivel de análisis nominal y el más restrictivo es el numérico. Por lo tanto, las variables cuantificadas acumularán mayor cantidad de varianza cuando se especifica el nivel de análisis nominal y la mayor cantidad de varianza acumulada en cada componente se verá cuando todas las variables sean analizadas bajo un nivel nominal.

- Los análisis de nivel ordinal y nominal descritos anteriormente pueden ser algo irregulares. Como una alternativa, es posible usar funciones suaves, en este caso splines, para obtener la transformación no lineal. A grandes rasgos una transformación spline monotónica es menos restrictiva que una transformación lineal pero más restrictiva que una ordinal, ya que sus restricciones no solo implican que conserven el orden, sino también que las transformaciones muestren una curva suave. Las splines pueden ser tanto no monotónicas como monotónicas para reemplazar las transformaciones nominales y ordinales vistas en el documento. El estudio de las transformaciones spline en el marco del NLPCA se abordará en estudios futuros.

- De igual forma, es posible utilizar transformaciones sobre los datos inicialmente cuantificados por medio de un OneHot Encoder que reemplacen al paso definido por el OS y que permitan proyectar a los datos a un espacio de mayor dimensión, para posteriormente obtener una proyección de los datos en un espacio de menor dimensión, por medio de un paso de descomposición espectral. Esto lleva al NLPCA a ser un caso particular del KPCA en donde cada variable recibe un tratamiento de transformación diferente al de las demás.

- Representación de variables como vectores: Un camino para representar una variable cuantificada consiste en presentar los puntos de las categorías en el espacio de las componentes principales. Este tipo de gráfico, consiste en representar la variable por medio de una línea recta que pasa a través del origen. Los puntos de las categorías son posicionados sobre el vector o recta y sus coordenadas se hallan multiplicando las cuantificaciones de las categorías por las correspondientes cargas sobre la primera y segunda componente. El orden de los puntos de las categorías sobre el vector de la variable está en concordancia con las cuantificaciones: el origen representa la media de la variable cuantificada de manera que las categorías con cuantificaciones mayores a la media se ubican en el lado derecho del origen en el cual los puntos de las cargas de las componentes están posicionadas y las categorías

con cuantificaciones menores que la media permanecen en dirección opuesta. Este tipo de representación responde a la teoría de pérdida conjunta y será utilizado en el capítulo 4 como apoyo en la cosntrucción de cartas de control.

- Representación de variables como centroides: A diferencia de la representación de variables como vectores la representación de variables como centroides se basa en la teoría de pérdida de reunión y, por lo tanto, es la única forma de representar las variables con nivel de análisis nominal múltiple. Cuando la transformación nominal simple es irregular y no puede ser interpretada fácilmente o cuando las categorías originales no pueden ser puestas en cualquier orden significativo, las cuantificaciones nominales múltiples son una gran alternativa. El objetivo de la cuantificación nominal múltiple es representar el comportamiento promedio de un conjunto de individuos mas no describir cada individuo por aparte. El objetivo es alcanzado por medio de asignar una cuantificación para cada categoría en cada una de las dimensiones generadas por las componentes. Las cuantificaciones múltiples para las categorías se obtienen por medio de promediar en cada dimensión las puntuaciones para todos los individuos que pertenecen a la misma categoría de una variable en particular. Consecuentemente, tales cuantificaciones serán diferentes para cada componente. Gráficamente, las cuantificaciones múltiples son las coordenadas de los centroides de las categorías de una variable en el espacio de las componentes principales. Se resaltan dos puntos: En contraste a las variables con otros niveles de medida, las variables múltiples nominales no obtienen cargas de componentes asociadas y, además, es importante darse cuenta de que se definen como múltiples nominales sólo a variables con un nivel de análisis nominal. Esta metodología será aplicada en el capítulo 4 como apoyo en el análisis del comportamiento promedio de los individuos discriminados por diferentes variables.

- Representación de individuos como puntos: Cada individuo obtiene una puntuación respecto a cada componente principal. Al igual que en el PCA lineal estas puntuaciones pueden ser utilizadas para construir un biplot en el que se puedan identificar patrones y relaciones entre las variables. Además, las variables nominales múltiples pueden ser graficadas dentro de este espacio funcionando como centroides de las diferentes categorías a las que pertenecen los individuos. Este gráfico es llamado Triplot, para más información consultar (Meulman et al.,

2004).

- Las diferencias cruciales es que el PCA lineal, las variables medidas son directamente analizadas, mientras, en PCA no lineal, la medida variable es cuantificadas durante los análisis (excepto cuando todas las variables son tratadas numéricamente). Otra diferencia concierne a la capacidad de anidamiento de los resultados de la solución.

3.6. Cartas de control basadas en datos óptimamente escalados

Al igual que en el PCA es posible construir un conjunto de herramientas a partir del NLPCA que apoye la gestión de calidad desde un punto de vista analítico y la identificación de patrones de variabilidad atípicos en los datos. Ya que el ALSOS y NLPCA permiten analizar variables cuantitativas y cualitativas en un mismo espacio, el lograr construir cartas de control con base en estas metodologías representa un potencial latente en la teoría de control de calidad para el procesamiento de variables cualitativas. Gracias a que en el proceso del NLPCA todas las variables pasan por un escalamiento óptimo se propone una metodología para la construcción de cartas de control basadas en datos óptimamente escalados, este es un aporte inédito al trabajo que se está presentando..

Para empezar es necesario recordar que la matriz de covarianza Σ obtenida a partir de la matriz $\mathbf{Q}$ de datos escalados contiene información sobre la variabilidad de las observaciones, la cual es fundamental a la hora de construir regiones de confianza. Es de esperar que una vez cuantificados los datos estos no sigan un comportamiento normal, sin embargo, si se cuenta con una tamaño muestral lo suficientemente grande es posible apoyarse en el Teorema del Límite Central y utilizar la teoría desarrollada en el Apéndice A.5 sobre la construcción de cartas de control para variables multivariadas. Es decir, se pueden ajustar las cartas de control de formato elipsoidal y T^2 para variables numéricas a datos cuantificados. En este punto es necesario resaltar que las variables cuantificadas óptimamente tendrán menor libertad que las variables numéricas a la hora de asignarle una medida a un individuo. Esto se debe a que las variables cualitativas cuentan con un conjunto limitado de posibles valores de cauntificación correspondiente a cada una de las categorías de la variable transformada. De manera que dos individuos pertenecientes a la misma categoría tendrán la misma cuantificación. Esto permitirá en el

caso multivariado hablar de comportamientos atípicos de grupos más que de individuos, ya que se le asignarán las mismas cuantificaciones a aquellos individuos con perfiles iguales.

3.6.1. Cartas para el análisis de estabilidad de una muestra de observaciones

Dada una muestra aleatoria $\mathbf{h}_{(1)}, \mathbf{h}_{(2)}, \ldots, \mathbf{h}_{(n)}$ de datos mixtos asociada a la matriz de datos cuantificados $\mathbf{Q}$ de tamaño $n \times p$, es necesario resaltar que las variables definidas como nominales múltiples contendrán como máximo tantas dimensiones en su cuantificación como categorías tengan las variables. De forma que es de especial importancia para la construcción de las cartas de control que cada una de las variables originales sea representada solo por un vector de cuantificaciones.

Sea una muestra aleatoria de gran tamaño $\mathbf{q}_{(1)}, \mathbf{q}_{(2)}, \ldots, \mathbf{q}_{(n)}$ con distribución $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, gracias al teorema del límite central. El residual utilizado para la construcción de las cartas de control es $\mathbf{q}_{(i)} - \bar{\mathbf{q}}$. De manera que

$$\mathbf{q}_{(i)} - \bar{\mathbf{q}} = \left(1 - \frac{1}{n}\right)\mathbf{q}_{(i)} - \frac{1}{n}\mathbf{q}_{(1)} - \cdots - \frac{1}{n}\mathbf{q}_{(i-1)} - \frac{1}{n}\mathbf{q}_{(i+1)} - \cdots - \frac{1}{n}\mathbf{q}_{(n)},$$

en donde el valor esperado y la covarianza son

$$E(\mathbf{q}_{(i)} - \bar{\mathbf{q}}) = \boldsymbol{\mu} - \boldsymbol{\mu} = \mathbf{0},$$

$$Cov(\mathbf{q}_{(i)} - \bar{\mathbf{q}}) = \left(1 - \frac{1}{n}\right)^2 \boldsymbol{\Sigma} + (n-1)n^{-2}\boldsymbol{\Sigma} = \frac{(n-1)}{n}\boldsymbol{\Sigma}.$$

Cada residual $\mathbf{q}_{(i)} - \bar{\mathbf{q}}$ tiene una distribución normal $N_p\left(\mathbf{0}, \frac{(n-1)}{n}\boldsymbol{\Sigma}\right)$. Recordando que bajo muestras de gran tamaño la distancia de Mahalanobis estimada se distribuye como una chi cuadrado, se tiene que

$$(\mathbf{q}_{(i)} - \bar{\mathbf{q}})'\mathbf{S}^{-1}(\mathbf{q}_{(i)} - \bar{\mathbf{q}}) \sim \chi^2_p,$$

lo cual está demostrado en la sección A.2.

Cartas de formato elipse

La carta de formato elipse es una carta de control bivariada y, por lo tanto, solo se pueden utilizar dos variables para su construcción. Por tal

razón se utilizan las dos características de la i-ésima unidad (q_{i1}, q_{i2}) con el límite de control definido por la región

$$(\mathbf{q}_{(i)} - \bar{\mathbf{q}})' \mathbf{S}^{-1} (\mathbf{q}_{(i)} - \bar{\mathbf{q}}) \leq \chi_2^2(0{,}05),$$

bajo una confiabilidad del 95 %. Al detectarse un punto fuera de control, se debe analizar cada variable independientemente. Para lo cual se construye una carta de control para variables con los siguientes límites de control superior (LCS) e inferior (LCI):

$$LCS = \bar{q}_j + 3s_{jj},$$

$$LCI = \bar{q}_j - 3s_{jj}.$$

Donde el valor central del intervalo corresponde a $\bar{q}_j$. Si los datos son no negativos el límite inferior se define en cero.

Cartas de control $\mathbf{T}^2$

Cuando se desee controlar más de dos características a la vez es necesario utilizar una carta T^2. Para esto, se calcula el estadístico T_i^2 para la i-ésima observación por medio de

$$T_i^2 = (\mathbf{q}_{(i)} - \bar{\mathbf{q}})' \mathbf{S}^{-1} (\mathbf{q}_{(i)} - \bar{\mathbf{q}}).$$

Posteriormente se grafican los individuos T_i^2 en un eje temporal con el límite de control inferior en cero y el límite de control superior en

$$LCS = \chi_p^2(0{,}05),$$

para una carta de control del 95 % de confiabilidad. Cuando se identifica una observación fuera de control se pueden utilizar una modificación de los intervalos de Bonferroni con el fin de describir de mejor manera el comportamiento de la unidad problemática. De manera que la j-ésima variable está fuera de control si q_{ij} no pertenece al intervalo

$$(\bar{q}_j - t_{n-1}(0{,}005/p)\sqrt{s_{jj}},\ \bar{q}_j + t_{n-1}(0{,}005/p)\sqrt{s_{jj}}),$$

donde p es el total de variables medidas.

3.6.2. Regiones de control para observaciones futuras individuales

La región de control para observaciones futuras se construye a partir de un conjunto de datos estables que permiten juzgar el comportamiento de nuevas observaciones independientes. Sea la muestra de datos $\mathbf{q}_{(1)}, \mathbf{q}_{(2)}, \ldots, \mathbf{q}_{(n)}$ utilizados para la creación de la región de confianza y la observación futura cuantificada $\mathbf{q}$ proveniente de la observación $\mathbf{h}$ con el vector indicador asociado $\mathbf{g}$, es posible adaptar los resultados presentados en la sección A resaltando el tratamiento de la nueva observación construida como $\mathbf{q} = \begin{bmatrix} q_1 & q_2 & \ldots & q_p \end{bmatrix}$ donde

$$q_j = \begin{cases} \mathbf{g}'\mathbf{w}_j, & \text{si la dimensión } j \text{ es nominal simple,} \\ \mathbf{g}'\mathbf{W}_{j1}, & \text{si la dimensión } j \text{ es nominal múltiple,} \\ \alpha_j h_j + \beta_j, & \text{si la dimensión } j \text{ es numérica.} \end{cases}$$

Vale recordar que las matrices y vectores de cuantificación son obtenidos durante el algoritmo de mínimos cuadrados alternantes del NLPCA. Específicamente para un nivel de análisis numérico la variable puede tomar tanto la cantidad h_j o una transformación lineal que se ajuste al cono definido para la variable. Sin embargo, el uso de una metodología o la otra dependerá de del algoritmo utilizado para procesar los datos. De esta forma la construcción de las cartas de control se presenta a continuación.

Sean $\mathbf{q}_{(1)}, \mathbf{q}_{(2)}, \ldots, \mathbf{q}_{(n)}$ distribuidas independientemente como $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ y sea $\mathbf{q}$ una observación futura de la misma distribución. Entonces el estadístico T^2 tiene la forma

$$T^2 = \frac{n}{n+1}(\mathbf{q} - \bar{\mathbf{q}})'\mathbf{S}^{-1}(\mathbf{q} - \bar{\mathbf{q}}),$$

el cual se distribuye como

$$T^2 \sim \frac{(n-1)p}{n-p} F_{p,n-p}.$$

De manera que un elipsoide de predicción p dimensional del $100(1-\alpha)\,\%$ de confiabilidad está dado por todos los $\mathbf{q}$ que satisfacen

$$(\mathbf{q} - \bar{\mathbf{q}})'\mathbf{S}^{-1}(\mathbf{q} - \bar{\mathbf{q}}) \leq \frac{(n^2-1)p}{n(n-p)} F_{p,n-p}(\alpha).$$

Cartas de control elipsoidales para futuras observaciones

Con $p = 2$, el 95 % la elipse de predicción se puede definir como

$$(\mathbf{q} - \bar{\mathbf{q}})'\mathbf{S}^{-1}(\mathbf{q} - \bar{\mathbf{q}}) \leq \frac{2(n^2-1)}{n(n-2)} F_{2,n-2}(0,05).$$

De forma que cualquier observación futura $\mathbf{q}$ se declara fuera de control si se encuentra fuera de la elipse.

Cartas de control T^2 para futuras observaciones

Para cada observación $\mathbf{q}$ se debe ubicar el punto

$$T^2 = \frac{n}{n+1}(\mathbf{q} - \bar{\mathbf{q}})'\mathbf{S}^{-1}(\mathbf{q} - \bar{\mathbf{q}}),$$

en orden temporal definiendo el límite de control inferior como cero y el límite de control superior como

$$LCS = \frac{(n-1)p}{(n-p)} F_{p,n-p}(0{,}05).$$

Los puntos por encima del límite de control superior representan causas potenciales de variación y sugieren que el proceso debería ser examinado para determinar si es necesario tomar una decisión correctiva o que un individuo presenta características altamente atípicas respecto a la población a la que pertenece.

3.7. Cartas de control basadas en NLPCA

Para las cartas de control basadas en NLPCA se debe tener presente que debido a las restricciones impuestas sobre la matriz de puntuaciones $\mathbf{Y}$, ecuación 3.22, la matriz de covarianza no es más que la matriz identidad, $\mathbf{\Sigma} = \mathbf{I}_q$. De forma que la distancia de Mahalanobis para las puntuaciones de las componentes es

$$\mathbf{y}'_{(i)}\mathbf{\Sigma}^{-1}\mathbf{y}_{(i)} = \mathbf{y}'_{(i)}\mathbf{I}_q\mathbf{y}_{(i)} = \sum_{i=1}^{q} y_{ij}^2 \sim \chi_q^2,$$

con este resultado es posible construir una primera propuesta para las cartas de control en formato elipse y formato T^2 basadas en NLPCA. La metodología recomendada en este caso es igual a la del PCA tradicional, es decir, se debe generar un número de componentes igual al total de las variables utilizadas en el estudio, con el fin de que se capture la totalidad de la variabilidad de los datos.

3.7.1. Cartas para el análisis de estabilidad de una muestra de observaciones

Sea $\mathbf{q}_{(1)}, \mathbf{q}_{(2)}, ..., \mathbf{q}_{(n)}$ una muestra de gran tamaño de datos escalados óptimamente. Las dos primeras componentes principales están dadas por $\hat{y}_{i1} = \hat{\mathbf{v}}_1(\mathbf{q}_{(i)} - \bar{\mathbf{q}})$ y $\hat{y}_{i2} = \hat{\mathbf{v}}_2(\mathbf{q}_{(i)} - \bar{\mathbf{q}})$ cuyas varianzas muestrales son próximas a la unidad. La distancia estimada de una observación $\mathbf{y}_{(i)}$ a su media está dada por

$$\hat{y}_{i1}^2 + \hat{y}_{i2}^2 + ... + \hat{y}_{ip}^2 \leq \chi_p^2(\alpha),$$

bajo muestras de gran tamaño.

Cartas de control de formato elipse

La carta de control elipsoidal para dos componentes obtenidas a partir del NLPCA consiste de los límites definidos por la región de confianza

$$\hat{y}_{i1}^2 + \hat{y}_{i2}^2 \leq \chi_2^2(\alpha),$$

con una significancia del $\alpha\,\%$ de manera que si la observación $\mathbf{y}_{(i)}$ se ubica por fuera de la región se dice que no se encuentra bajo control.

Cartas de control de formato T^2

Para el caso de la carta de control T^2 se calcula la distancia T_i^2 para cada una de las observaciones $\mathbf{y}_{(i)}$ por medio de

$$T_i^2 = \hat{y}_{i1}^2 + \hat{y}_{i2}^2 + ... + \hat{y}_{ip}^2,$$

con los límites de control

$$LCS = \chi_p^2(\alpha),$$
$$LCI = 0,$$

para una confiabilidad del $(1-\alpha)\,\%$.

El hecho de que la matriz de covarianza sea la identidad implica que se hable de una circunferencia de control de radio $\chi_2^2(\alpha)$ en vez de una elipse como tradicionalmente se trabaja, como se verá en la sección 4. Esta puede ser la mayor falencia de esta propuesta de construcción de cartas de control, ya que las puntuaciones no recogen la totalidad de la variabilidad almacenada en los datos cuantificados. En este caso, la variabilidad de los datos cuantificados

es recogida por la matriz de cargas a diferencia del PCA donde es contenida por la matriz de puntuaciones. De esta forma, se propone modificar las restricciones impuestas sobre la matriz de pesos y la matriz de puntuaciones en orden de que los resultados sean equivalentes a los obtenidos en el PCA en todos los aspectos referentes a la descomposición en valores singulares. Estas restricciones son

$$\mathbf{Y}'\mathbf{Y} = \Lambda,$$

$$\mathbf{V}'\mathbf{V} = \mathbf{I}_p.$$

Sin embargo, este cambio en las restricciones implica un cambio en el algoritmo de NLPCA. A continuación se presenta la modificación necesaria al algoritmo 3.

Algoritmo 4. NLPCA - Datos categóricos múltiples modificado.

0. Inicialización aleatoria de los vectores y matrices de cuantificaciones y la matriz de puntajes: $\tilde{\mathbf{w}}_j$, $\tilde{\mathbf{W}}_j$, $\tilde{\mathbf{Y}}$.

1. Actualización de la matriz de datos escalados:

   1.a. Datos numéricos: $\tilde{\mathbf{q}}_j \leftarrow \mathbf{h}_j$.

   1.b. Datos nominales simples y ordinales: $\tilde{\mathbf{q}}_j \leftarrow \mathbf{G}_j\tilde{\mathbf{w}}_j$.

   1.c. Datos nominales múltiples: $\tilde{\mathbf{Q}}_j \leftarrow \mathbf{G}_j\tilde{\mathbf{W}}_j$.

      1.1. Centralización: $\mathbf{Q}^\circ \leftarrow (\mathbf{I}_n - n^{-1}\mathbf{1}_n'\mathbf{1}_n)\tilde{\mathbf{Q}}$.

      1.2. Normalización: $\hat{\mathbf{Q}} \leftarrow \sqrt{n}\mathbf{Q}^\circ diag(\mathbf{Q}^{\circ\prime}\mathbf{Q}^\circ)^{-1/2}$.

2. Actualización de la matriz de pesos y puntuaciones:

   2.1. Descomposición en valores propios: $(\mathbf{V}_e, \Lambda) \leftarrow eig(\hat{\boldsymbol{\rho}})$.

   2.2. Actualización de la matriz de cargas: $\hat{\mathbf{V}} \leftarrow \mathbf{V}_e$.

   2.3. Actualización de la matriz de puntuaciones: $\hat{\mathbf{Y}} \leftarrow \hat{\mathbf{Q}}\mathbf{V}$.

3. Actualización de la matriz de cuantificación:

   3.a. Datos nominales simples: $\hat{\mathbf{w}}_j \leftarrow \mathbf{D}_j^{-1}\mathbf{G}_j'\tilde{\mathbf{Y}}\hat{\mathbf{v}}_{(j)}$.

   3.b. Datos ordinales: Regresión monotónica sobre $\hat{\mathbf{w}}_j$ con $\tilde{\mathbf{q}}_j$ como variable respuesta y $\mathbf{h}_j$ como variable regresora.

      3.b.1. Centralización: $\mathbf{w}_j^\circ \leftarrow (\mathbf{I}_q - \mathbf{G}_j^{-1}\mathbf{1}_q\mathbf{1}_q'\mathbf{G}_j)\hat{\mathbf{w}}_j$.

      3.b.2. Normalización: $\tilde{\mathbf{w}}_j \leftarrow \mathbf{w}_j^\circ(\mathbf{w}_j^{\circ\prime}\mathbf{D}_j\mathbf{w}_j^\circ)^{-1/2}$.

3.c. Datos nominales múltiples: $\tilde{\mathbf{W}}_j \leftarrow \mathbf{D}_j^{-1}\mathbf{G}_j'\tilde{\mathbf{Y}}$.

3.c.1. Centralización: $\mathbf{W}_j^{\circ} \leftarrow (\mathbf{I}_q - \mathbf{G}_j^{-1}\mathbf{1}_q\mathbf{1}_q'\mathbf{G}_j)\hat{\mathbf{W}}_j$.

3.c.2. Normalización: $\tilde{\mathbf{W}}_j \leftarrow \mathbf{W}_j^{\circ}(\mathbf{W}_j^{\circ\prime}\mathbf{D}_j\mathbf{W}_j^{\circ})^{-1/2}$.

4. Test de convergencia.

Es de esperarse que bajo este algoritmo las propiedades de las matrices $\mathbf{Y}$ y $\mathbf{V}$ sean equivalentes a las expuestas durante el Capítulo 2 y sea posible generalizar la mayoría de los resultados expuestos en él para el NLPCA. El impacto de la modificación propuesta será investigado en futuros estudios a profundidad.

Capítulo 4

Aplicación y visualización

4.1. Descripción de los datos

Para ilustrar las metodologías estudiadas se utilizarán datos obtenidos del proyecto "Corazones de Baependi"(Processo Fapesp 2007/58150-7) conducido por el Laboratorio de Genética y Cardiología Molecular (Incor/USP) (de Oliveira, et al., 2008). El proyecto tiene como objetivo general encontrar determinantes genéticos que modulan o regulan enfermedades cardiovasculares evaluando la influencia de factores genéticos y ambientales. En este sentido varios estudios han sido publicados y varias investigaciones están siendo desarrolladas en esa dirección. El proyecto se ha venido desarrollando en diferentes fases teniendo como foco una muestra de los habitantes de la ciudad de Baependi del estado Minas Gerais, Brasil (con un área de 752 km^2 y 18.307 habitantes, aproximadamente).

Una primera fase de recolección de datos fue llevada a cabo entre diciembre de 2005 y enero de 2006 en que un total de 81 familias fueron muestreadas, aproximadamente 1700 individuos en total. En esta primera fase del estudio el tamaño de las familias fue entre 3 y 156 miembros con una media de 21 individuos por familia, la edad de los individuos se encontró entre los 18 y los 100 años con media de 44 años y 57 % de todos los individuos fueron mujeres. En cada uno de los participantes, varios fenotipos (o variables) para análisis estadísticos fueron evaluados tales como glicemia, colesterol, presión arterial, triglicéridos entre otras. También fueron obtenidas medidas antropométricas como el peso, la altura y la circunferencia de la cintura. Se definieron aspectos de comportamiento como la frecuencia de consumo de tabaco, alcohol y realización de actividad física. De la misma

forma, medidas de presión sanguínea sistólica y diastólica fueron incluidos en el experimento. Mediciones de glucosa, colesterol y triglicéridos fueron evaluados por medio de técnicas de análisis sanguíneos.

Un primer seguimiento tuvo lugar en 2010 donde se agregaron recopilaciones de datos específicos. El seguimiento involucró tanto un muestreo de ADN renovado así como medidas de los principales factores de riesgo cardiovascular. En el seguimiento de 2010, se agregaron 548 personas y un nuevo seguimiento, que marca los 10 años del estudio, fue realizado en 2016. Para los estudios en el área de Psicología, variables como depresión y ansiedad, entre otras, fueron recolectadas entre abril de 2013 y marzo de 2016 en la estación de investigación permanente del estudio, de Baependi. Cabe anotar que el protocolo de estudio fue aprobado por el comité de Ética del Hospital de las Clínicas, Universidad de São Paulo, Brasil, y cumplió con los estándares éticos internacionales sobre experimentación humana. Una descripción detallada del estudio de Baependi, puede ser encontrado en Egan et al. (2016).

Vale resaltar que el Proyecto " Corazones de Baependi" fue el primer estudio familiar en Brasil sobre enfermedades cardiovasculares y actualmente se propone monitorear a 3.500 participantes que serán reevaluados cada 5 años. Lo relevante de este estudio es que tiene el privilegio de observar posibles factores, como los hábitos de estilo de vida, involucrados en el desarrollo de enfermedades crónicas a lo largo del tiempo, como lo son la obesidad, la diabetes, la presión arterial alta y el colesterol alto, entre muchas otras variables.

Varios estudios han surgido, algunos enfocados a investigar en qué proporción en promedio se transmiten las características de los padres a su descendencia, una medida ampliamente conocida en genética como herdabilidad. Un valor alto de herdabildiad es un indicador que existen genes regulando la característica en estudio, o sea la variable o fenotipo. Otros estudios se han enfocado en el área de psicología y algunos enfocados a encontrar índices que pueden ayudar a determinar individuos con enfermedades cardiovasculares.

A continuación se describen algunos de los estudios de forma general, enfocando en las metodologías estadísticas utilizadas.

- En los inicios del proyecto, un estudio dedicado al cálculo de medidas

de herdabilidad fue realizado, proporcionando la primera evidencia de que una proporción significativa de la variabilidad de los factores de riesgo cardiovascular (Colesterol, Presión arterial, Índice de Masa Corporal) se explica por factores genéticos. Las metodologías estadísticas utilizadas se enfocaron en la aplicación de modelos lineales mixtos. Ante estos resultados, se propusieron más estudios de asociación para identificar variantes genéticas específicas asociadas con estos importantes predictores de enfermedades cardiovasculares, (de Oliveira et al., 2008).

- Estudio de la influencia de factores genéticos y ambientales sobre el cronotipo (o preferencia diurna) el cual es un instrumento útil para estudiar la biología circadiana en humanos. La variable que fue estudiada, a través de modelos lineales mixtos, proviene de un cuestionario ya validado en el área (Morningness-Eveningness questionnaire, MEQ) y aplicado a los participantes. En esta investigación los resultados sugieren que los estudios de asociación de todo el genoma (GWAS) para el cronotipo de preferencia diurna son factibles en principio, pero también destacan una importante limitación en términos de la sorprendente variabilidad basada en el contexto de la puntuación, MEQ (von Schantz, et al. 2015).

- Baependi es una ciudad de Brasil que ofrece una ventana de oportunidades para estudiar la influencia de los patrones de sueño en una población rural muy mezclada con un estilo de vida conservador. El sueño está modulado por varios factores, incluidos el género, la edad y el cronotipo. Se ha planteado la hipótesis de que las poblaciones urbanas contemporáneas están bajo presión hacia una duración más corta del sueño y una calidad de sueño más deficiente. A través de varias medidas y con la aplicación de estadísticas descriptivas, de modelos de regresión lineal y logística se evaluaron las características asociadas al sueño. La naturaleza longitudinal del estudio de Baependi permitirá investigar si estos parámetros cambiarán en los próximos años y de qué manera (Beijamini et al.,2016).

- Un estudio relacionado es acerca del insomnio. El insomnio afecta significativamente la morbilidad de por vida y por lo tanto tiene costos socioeconómicos sustanciales. En los países desarrollados, la prevalencia del insomnio está aumentando, sin embargo, se sabe poco sobre el insomnio en poblaciones menos urbanizadas y de bajos ingresos. Baependi es una ciudad rural donde se ha demostrado que mantiene los

ciclos de sueño sincronizados con la luz natural, a pesar de la electricidad. El objetivo fue investigar los componentes del insomnio basada en datos de familia, utilizando el cuestionario Insomnia Severity Index (ISI).

Se obtuvo un valor de herdabilidad que indica la presencia de genes modulando esta variable y se procedió a un estudio de asociación genética (GWAS) mediante la utilización de datos de familia y plataformas de marcadores moleculares utilizando la metodología modelos lineales mixtos. Este estudio arrojó resultados relevantes en investigaciones genéticas encontrando cuatro asociaciones de marcadores moleculares en todo el genoma con el fenotipo ISI, a saber rs869481, rs62037617, rs3747579, que se encuentran en el gen CORO7, y rs3789038, ubicado en el gen HMOX2 en el cromosoma 16 (Ahmed et al., 2019).

- Otro frente de investigación, del proyecto en mención, que se ha venido desarrollando es en el área de Neuropsicología en que varios investigadores expertos han estado recolectando datos para este tipo de análisis, por ejemplo, investigar la correlación genética entre los síntomas de Depresión y Ansiedad teniendo en cuenta la estructura familiar. Se utilizó la Escala Hospitalaria de Ansiedad y Depresión (HADS) para evaluar los síntomas de estas variables. Las estimaciones de heredabilidad se obtuvieron mediante un modelo lineal mixto y la correlación genética a través de um modelo lineal mixto bivariado para obtener la partición de la covarianza de Depresión y Ansiedad en componentes genéticos y ambientales y para calcular la contribución genética que modula ambos conjuntos de síntomas.

  Los resultados proporcionaron pruebas sólidas de una superposición genética entre los síntomas de Depresión y Ansiedad, lo que tiene relevancia para la comprensión de la base biológica de estos constructos y podría explotarse en estudios de asociación de todo el genoma para encontrar determinantes genéticos que modulan estas enfermedades (Taporoski, et al., 2015).

- Otra investigación se enfoca en estudios de influencia genética en factores cognitivos (Taporoski, et al., 2015). En la actualidad, los investigadores de esta área están interesados en dar continuidad al estudio en el área de Cognición, específicamente sobre las Funciones Ejecutivas, que son las habilidades cognitivas involucradas en la formación, planificación, selección, mantenimiento, seguimiento e implementación

de metas que se mantienen temporalmente en la memoria de trabajo (Friedman & Miyake, 2017).

- Por último, se mencionará que recientemente algunos estudios han evaluado el papel de las medidas de adiposidad en la predicción del riesgo de hipertensión. El objetivo de este estudio fue comparar cuáles de los cuatro indicadores de adiposidad (circunferencia de cintura-WC, índice de masa corporal-IMC, índice de adiposidad corporal- BAI e índice de adiposidad visceral-VAI) se asocian mejor con la hipertensión en la población de Baependi.

  Para evaluar los modelos de desempeño, se construyeron curvas ROC y se utilizó el área bajo la curva (AUC) para medir el poder discriminatorio para la hipertensión en hombres y mujeres. Los valores de sensibilidad y especificidad para cada medida de adiposidad se determinaron mediante análisis de curvas ROC. Los puntos de corte óptimos para WC, IMC, BAI y VAI se establecieron en función de la combinación más alta de sensibilidad y especificidad.

  Como resultados se destaca que los indicadores de adiposidad WC, IMC, BAI y VAI fueron más altos en hipertensos en comparación con los no hipertensos. Además, la WC y el IMC se asociaron más fuertemente con la hipertensión en hombres y mujeres, respectivamente. Los indicadores de adiposidad WC e IMC se asociaron mejor con la hipertensión que BAI y VAI, en ambos sexos, y podrían ser herramientas útiles para el cribado de pacientes hipertensos (de Oliveira et al., 2017).

Para la presente aplicación se utilizará una muestra total de 1227 individuos que no presentan ausencia de información en las variables de interés, con el fin de evitar el uso de algoritmos de estimación de datos perdidos. Se utilizará un total de 14 variables cuantitativas como presión sanguínea, glucosa, entre otras; y 8 variables cualitativas que ayudan a caracterizar en los individuos aspectos tales como estilo de vida y aspectos sociales en general. Las variables bajo análisis están descritas en la Tabla 4.1. En la primera columna se describe el nombre de la variable. En la segunda columna, para las variables cualitativas se presentan las categorías y para las variables continuas se presentan las unidades de medida. En la tercera columna se presenta el nivel de análisis, que puede ser Nominal múltiple, Nominal Simple, Ordinal, el cual se especifica en detalle en el capítulo 3. Finalmente en la cuarta columna se hace una breve descripción de cada variable.

Tabla 4.1: Diccionario de variables

Nombre	Categorías/Unidades	Nivel de análisis	Descripción
Género	1: Hombre. 2: Mujer.	Nominal múltiple.	Indica el género biológico del individuo.
Raza	1: Blanca. 2: Negra. 3: Parda - Mulato. 4: Amarilla. 5: Indígena. 6: Otros - Mestizos.	Nominal múltiple.	Indica la raza a la que pertenece el individuo.
Edad	1: Entre 18 y 30 años. 2: Entre 31 y 43 años. 3: Entre 44 y 56 años. 4: Entre 57 y 100 años.	Nominal múltiple.	Indica el rango de edad al que pertenece el individuo.
Renta	1: Hasta un salario mínimo. 2: 1 a 5 salarios mínimos. 3: 5 a 10 salarios mínimos. 4: 10 a 20 salarios mínimos. 5: Más de 20 salarios mínimos. 6: No sabe. 7: No responde.	Nominal simple.	Indica la cantidad de salarios mínimos mensuales ganados por el individuo.
Estado	1: Casado o en consenso. 2: Soltero. 3: Separado. 4: Viudo.	Nominal simple.	Indica el estado civil del individuo.
Escolaridad	1: Analfabeto. 2: Sabe leer y escribir. 3: Primaria incompleta. 4: Primaria completa. 5: Nivel 1 incompleto. 6: Nivel 1 completo. 7: Nivel 2 incompleto. 8: Nivel 2 completo. 9: Técnico. 10: Superior incompleto. 11: Superior completo	Ordinal.	Indica el nivel de escolaridad del individuo.
Fuma	1: Sí, en el pasado. 2: Sí, todavía fumo. 3: No. 4: No responde.	Nominal simple.	Indica si el individuo fuma.
Licor	1: Bebe diariamente. 2: 1 a 3 veces por semana. 3: 4 a 6 veces por semana. 4: 1 a 3 veces por mes. 5: Menos de una vez por semana. 6: Se embriaga al menos una vez por mes. 7: No bebe. 8: No responde.	Nominal simple.	Indica la frecuencia con la que el individuo fuma.
Peso	Kilogramos.	Numérico.	
Altura	Centímetros.	Numérico.	
Abdomen	Centímetros.	Numérico.	Perímetro abdominal.
Cadera	Centímetros.	Numérico.	Circunferencia de cadera.
Glucosa	Miligramos/decilitros.	Numérico.	
Urea	Miligramos/decilitros.	Numérico.	
Creatinina	Miligramos/decilitros.	Numérico.	
Triglicéridos	Miligramos/decilitros.	Numérico.	
CHDL	Miligramos/decilitros.	Numérico.	Colesterol de lipoproteína de alta densidad.
CLDL	Miligramos/decilitros.	Numérico.	Colesterol de lipoproteína de baja densidad.
PAS	Milimetros de mercurio.	Numérico.	Presión arterial sistólica.
PAD	Milimetros de mercurio.	Numérico.	Presión arterial diastólica.
Depresión		Numérico.	
Ansiedad		Numérico.	

En primera instancia, en la sección 4.2, el análisis consistirá en aplicar la metodología de Componentes Principales o PCA lineal, presentado en el capítulo 2, con ayuda del paquete *Scikit-learn* del Programa Python. Según la teoría, en este análisis solo se tendrán en cuenta las variables cuantitativas y el objetivo es la discriminación de las agrupaciones, sabiendo que se tienen variables asociadas a enfermedades cardiovasculares, variables psicológicas y variables antropométricas. Al finalizar la sección se ilustrará la teoría de Cartas de Control.

Posteriormente, en la sección 4.3, se ilustrará la metodología de Componentes Principales No Lineales, NLPCA, (desarrollada en el Capítulo 3), por medio del módulo de reducción de dimensión del programa *SPSS*.

4.2. Análisis de Componentes Principales

En la práctica el PCA lineal se puede realizar ya sea sobre la matriz de covarianzas o sobre la matriz correlaciones de los datos. El criterio utilizado para elegir cual de las dos matrices conviene usar debe ir de la mano con el orden de magnitud de los datos utilizados en el análisis, ya que si el orden de magnitud de una de las variables es mucho mayor que las demás, su varianza será mayor y tendrá más influencia en las combinaciones lineales que forman a las componentes principales, de forma que es posible que una componente dependa netamente de una variable produciendo un sesgo en el análisis. En el caso de estudio de la población de Baependi se utilizaron variables cuyos valores oscilan desde el orden decimal, como la creatinina, hasta orden de las centenas, como la altura. Es por esta razón que se opta por utilizar la matriz de correlaciones cuyo gráfico se presenta en la Figura 4.1.

De acuerdo con la matriz de correlaciones, se observan variables con correlaciones bajas respecto a las demás como en el caso de la altura (segunda línea), a excepción de la correlación entre altura y peso que es de 0.37. Por otro lado, se observan correlaciones altas como es el caso de peso con abdomen, 0.79; abdomen con cadera, 0.80; presión arterial sistólica (PAS) con presion arterial diástólica (PAD), 0.74. Por lo general se considera un valor de correlación como muy alto para valores superiores a 0.90, los cuales pueden generar una serie de problemas en el análisis de componentes debido a la redundancia en la información, posiblemente generando valores propios iguales a cero que indica dependencia lineal entre dichas variables y por ello

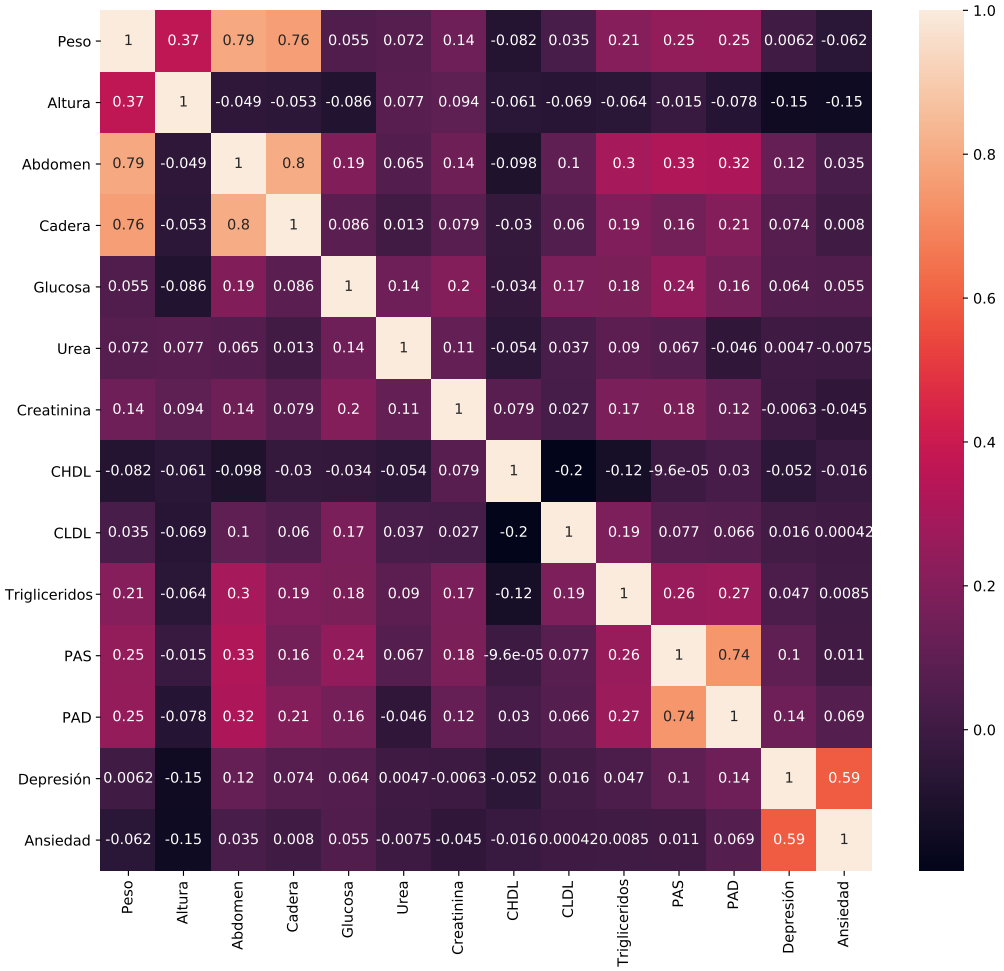

Figura 4.1: Matriz de correlaciones.

en el proceso de descomposición matricial generando valores propios iguales a cero. Ya que las 14 variables analizadas presentan un bajo nivel de correlación lineal se utilizan todas en el análisis.

Una vez definidas las variables a utilizar en el análisis, se procede a presentar los resultados obtenidos del PCA. En la Tabla 4.3 se presentan los valores propios de cada componente, la proporción de la varianza y la proporción de la varianza acumulada. Se observa que la primera componentes no tiene una explicación muy alta de la variabilidad de los datos, apenas el explica el 23 %. Por lo tanto, si se desea, por ejemplo, aproximadamente el 80 % de la variabilidad se necesitarían 7 componentes.

Tabla 4.2: Proporción de la varianza explicada por las Componentes Principales

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Valor propio	3.221	1.802	1.507	1.303	1.155	1.046	0.867
Proporción de la varianza	0.230	0.128	0.107	0.093	0.080	0.074	0.064
Proporción Acumulada	0.230	0.359	0.467	0.560	0.641	0.715	0.780
	PC8	PC9	PC10	PC11	PC12	PC13	PC14
Valor propio	0.794	0.730	0.658	0.402	0.250	0.170	0.093
Proporción de la varianza	0.055	0.052	0.047	0.028	0.017	0.012	0.006
Proporción Acumulada	0.835	0.887	0.934	0.963	0.981	0.993	1.000

Una herramienta útil a la hora de determinar un número apropiado de componentes principales a utilizar, es el *gráfico de sedimentación*, conocido en inglés como *scree plot*. El gráfico de sedimentación consiste en ubicar los valores propios de mayor a menor asociándolos con sus respectivas componentes. Permite seleccionar cuántas componentes se deben usar para el estudio, principalmente si el fin del análisis consiste en retirar tantas variables como información se considere irrelevante en los datos, vale resaltar que estas componentes pueden tener información valiosa para la identificación de valores atípicos, Sección 2.7. En la Figura 4.2, se presenta el gráfico de sedimentación, es posible identificar un cambio brusco en el comportamiento de los valores propios a partir de la décima componente, siendo este el punto de inflexión para el presente análisis.

Sin embargo, en ocasiones el interés del investigador se centra en conservar una determinada proporción de la variabilidad contenida en los datos, para este fin es necesario calcular la variabilidad porcentual para cada una de las componentes y definir por criterio de la investigación qué cantidad de la variabilidad porcentual acumulada se desea retener, Sección 2.6. En

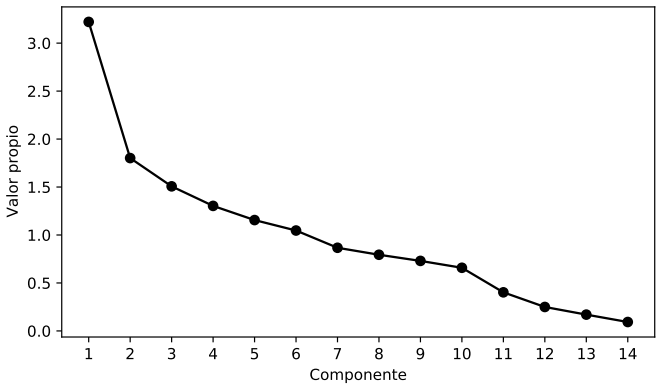

Figura 4.2: Gráfico de sedimentación.

este punto, se recomienda al lector adaptar el gráfico de sedimentación a los valores relativos de variabilidad conservada y graficar tanto una linea que indique la variabilidad acumulada como una línea constante en el valor de variabilidad porcentual que se desea conservar. La Figura 4.3 presenta el gráfico de sedimentación modificado. Se aprecia no solo el hecho de que el punto de inflexión se ubica sobre la décima componente y que conserva un 95 % de variabilidad, sino también, que el 80 % de la variabilidad de los datos está representada por aproximadamente las primeras 7 componentes principales.

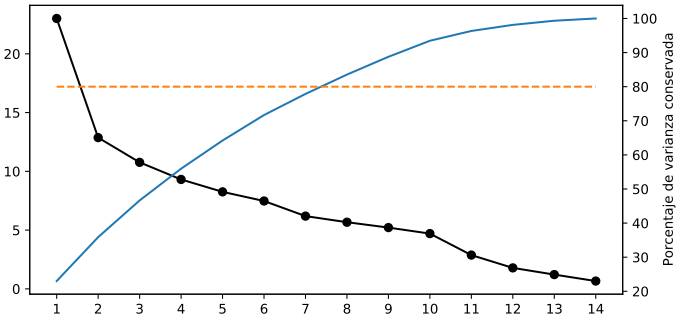

Figura 4.3: Gráfico de sedimentación modificado.

Además de utilizar el criterio de la varianza acumulada por cada componente con el fin de reducir dimensiones de los datos originales por medio de retirar información de ellos, es interesante para el investigador analizar las entradas de los vectores de cargas como puntos en un espacio bidimensional para identificar grupos o clusters de variables que tengan una alta relación. En la Figura 4.4 se presenta el gráfico de los dos primeros vectores, en el eje x vector 1 y en el eje y el vector 2. Se observan agrupaciones de bastante interés para el estudio como el de las enfermedades mentales (ansiedad y depresión) ubicado en al parte superior del gráfico, las medidas cardiovasculares en el centro (presión diastólica y sistólica, CLDL, CHDL, etc.) y en la región inferior las medidas antropométricas de los individuos (peso, altura, circunferencia abdominal y perímetro de cadera).

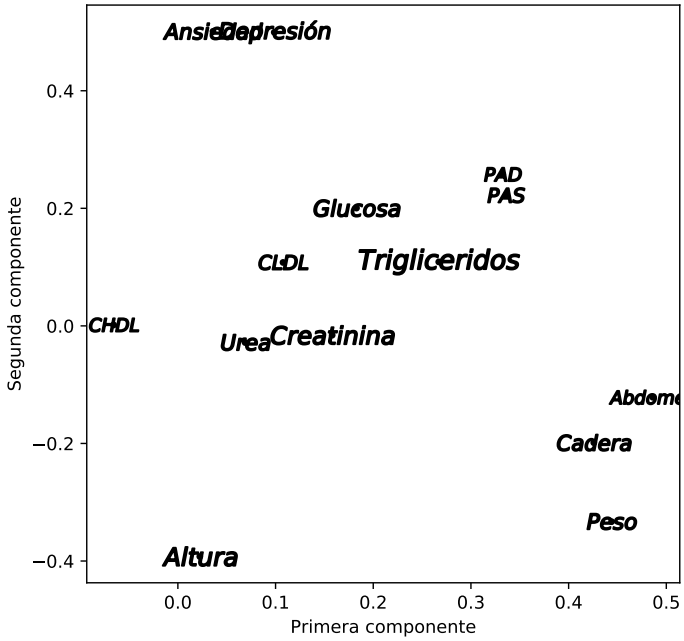

Figura 4.4: Gráfico de los primeros vectores para discriminar conjuntos de variables.

Una medida de mayor interés para este último grupo puede ser el índice de masa corporal ($IMC = peso/(altura)^2$), el cual en la Figura 4.5 muestra

una diferencia aún más notoria en la separación de las tres agrupaciones en las variables, como es esperado.

Para efectos de observar el poder de separación de las variables al aplicar cada uno de los métodos, se seguirá incluyendo en el análisis la variable altura.

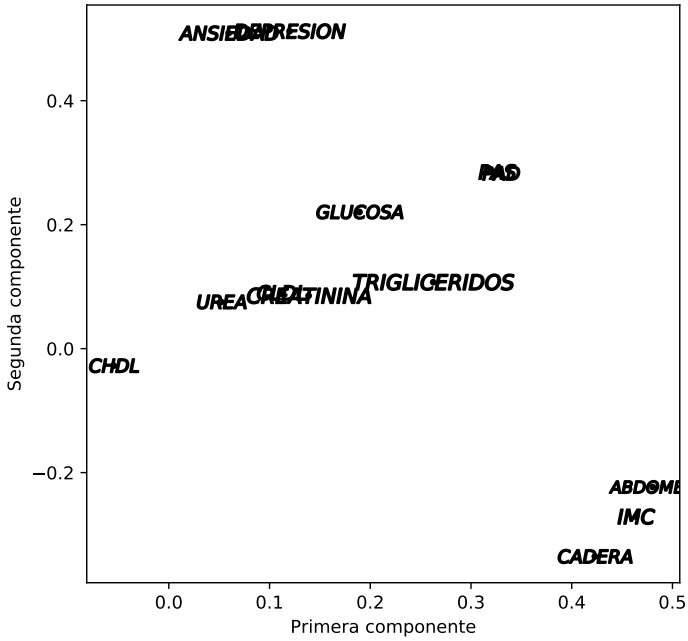

Figura 4.5: Gráfico de los primeros vectores para discriminar conjuntos de variables con la agrupación de IMC.

También es posible estudiar la influencia de cada una de las variables sobre las componentes analizando las entradas de cada vector de cargas, ya sea por medio de elevar cada elemento al cuadrado o calcular el respectivo coeficiente de correlación entre cada componente y variable, Sección 2.9. En las Figuras 4.6 y 4.7 se presenta el comportamiento de las cargas al cuadrado y los coeficientes de correlación de cada vector propio asociado a las dos primeras componentes, respectivamente, determinando si el aporte de cada variable es positivo o negativo respecto a las componentes.

Específicamente en la Figura 4.6, primera parte, se nota una alta influencia de las variables antropométricas (menos la altura), al igual que las variables que representan la presión presión arterial, PAS Y PAD y triglicéridos. La influencia positiva de todas las variables es importante resaltarla ya que una medición muy alta en cualquiera de las variables puede indicar problemas de salud frecuentemente relacionados con sobrepeso y problemas cardiovasculares. Observe que el colesterol (CHDL) está en contravía con el resto de variables, esto se explica ya que este colesterol de alta densidad, que es el colesterol bueno, ya que contribuye de manera positiva en problemas de orden vascular y renal; en otras palabras, subir los niveles de CHDL disminuye el riesgo de padecer trastornos vasculares y renales. Por lo tanto, es lógico que el CHDL aporte de forma negativa a ésta componente. En cuanto a la segunda componente (segunda parte de la Figura 4.6) se aprecia mayor preponderancia de las variables asociadas con enfermedades mentales (depresión y ansiedad), a las cuales las variables antropométricas les hacen un alto contrapeso. De esta manera es posible esperar que la primera componente funcione como una variable latente con la cual se puedan explicar la relación entre los problemas cardiacos y de sobrepeso mientras que en la segunda componente se puede estudiar la relación entre las enfermedades mentales y las medidas antropométricas.

Las conclusiones anteriores pueden extenderse a los resultados de los coeficientes de correlación entre las variables, cuyo gráfico se presenta en la Figura 4.7.

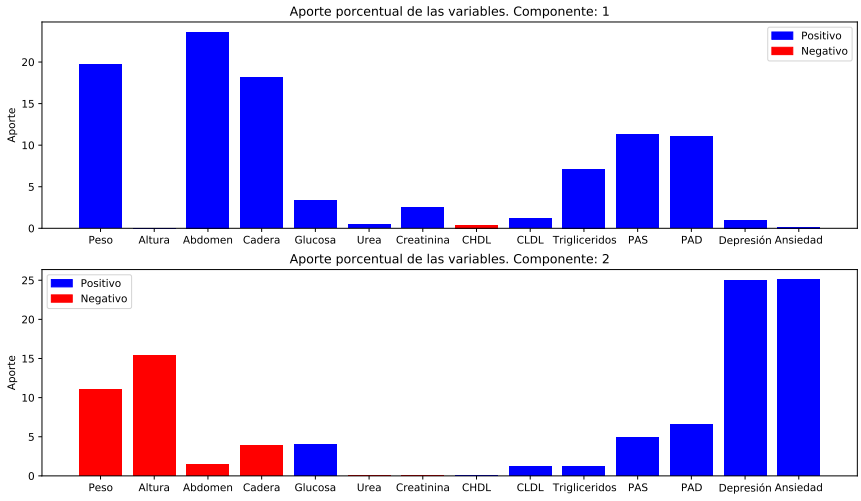

Figura 4.6: Elementos al cuadrado de los vectores de cargas.

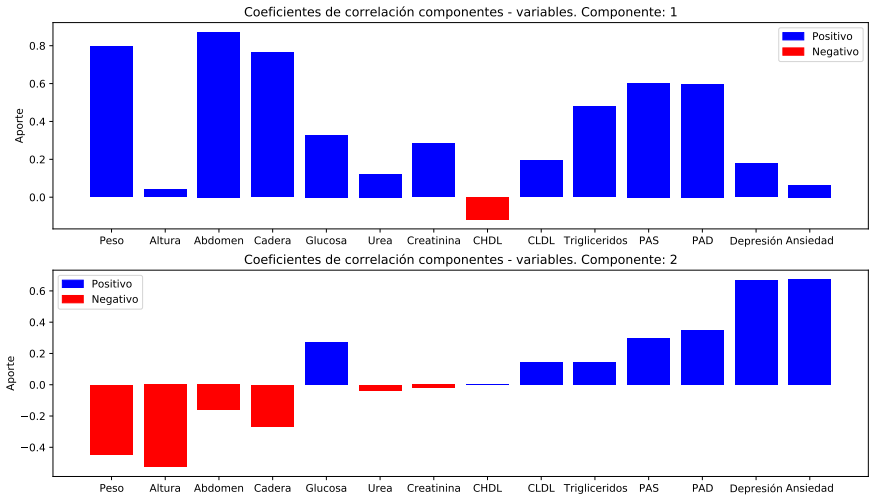

Figura 4.7: Coeficientes de correlación entre las variables y las componentes principales.

El PCA no solo permite estudiar el comportamiento de las variables en cada componente, también permite realizar una proyección de los datos en un nuevo espacio de menor dimensión, con el fin de obtener una visión general de los individuos en este caso. En la Figura 4.8 se presenta el gráfico de las dos primeras componentes principales, donde se observan las puntuaciones de los datos que corresponden a los individuos, no se discriminan patrones o grupos de individuos, apenas algunos individuos que se salen del comportamiento general de la población, pudiendo afirmar que este gráfico no aportan mucho en el análisis estadístico y discriminación de individuos. Sin embargo, este tipo de gráficos de dispersión son utilizados para apoyar lo gráficos *biplot*, sección 2.6.1, y realizar las *cartas de control*, sección 2.11, más que para la interpretación por sí solo.

Los biplots son gráficos construidos con la información de los vectores de cargas y las puntuaciones de las componentes, obteniendo una herramienta como la presentada en la Figura 4.9 en donde se puede apreciar la relación entre las variables y el comportamiento de los individuos. En términos generales un *biplot tradicional* consiste en graficar los vectores de correlación de los datos, que no son más que los coeficientes de correlación entre cada

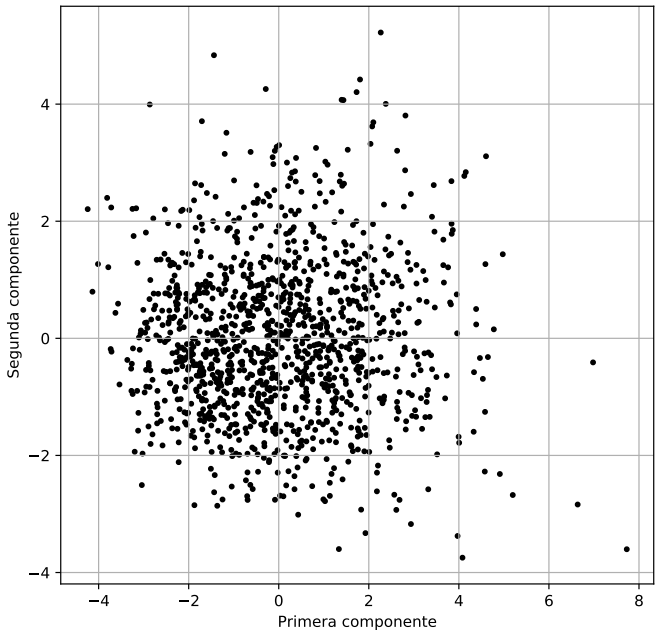

Figura 4.8: Gráfico de puntuaciones de los 1227 individuos de la muestra sobre la primera y segunda componente principal

variable y la componente respectiva, usados como dirección de los vectores superpuestos sobre un gráfico de dispersión de las puntuaciones de los individuos. Los círculos dibujados se llaman círculos de correlación e indican qué tan fuerte es la correlación de una variable con cada componente, el círculo interno indica una correlación de 0.5 y el círculo externo una correlación de 1.0. Los individuos están discriminados por códigos de color en las diferentes razas que conformaron el estudio, sin embargo, no es posible identificar patrones en los individuos que sean de interés para el investigador. En general, la información presentada en la gráfica no es clara respecto a los individuos y en cuanto a las variables se puede distorsionar su interpretación debido al exceso de elementos expuestos.

Con el fin de facilitar la interpretación de la información presentada en los biplots tradicionales se recomienda al investigador utilizar las medias o

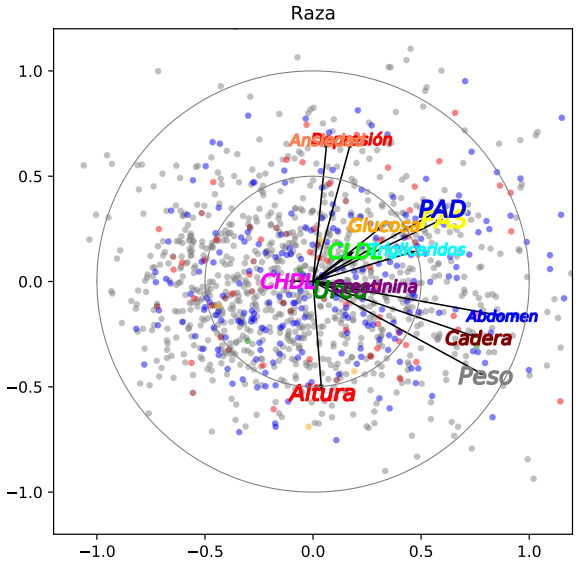

Figura 4.9: Biplot tradicional. Datos discriminados por raza.

centroides de los datos en lugar de su dispersión, siempre discriminando los individuos bajo la luz de una variable categórica. En este estudio se utilizarán las variables género, raza y edad como variables discriminatorias para describir de una forma más detallada el comportamiento de la población respecto a las mediciones consideradas. Cabe resaltar que el uso de centroides también trae una gran ventaja visual ya que da una vista más limpia de las relaciones entre las variables. Para la interpretación de las relaciones entre las variables es necesario resaltar el hecho de que un conjunto de vectores paralelos entre ellos indican un alto nivel de correlación lineal, lo que implica un alto nivel de dependencia lineal entre las variables. Por el contrario, un conjunto de vectores perpendiculares entre sí indican un bajo nivel de correlación lineal, es decir, son vectores linealmente independientes, para más información revisar el Apéndice A.1.

Antes de analizar el comportamiento de los individuos respecto a las variables es conveniente estudiar las relaciones presentadas entre las ellas. Al igual que en la gráfica 4.4 en la Figura 4.9 se observaron tres grupos claros de variables, uno relacionado con enfermedades mentales (depresión

y ansiedad) en la parte superior, otro relacionado con medidas renales y cardiovasculares y el último con medidas antropométricas. Las variables relacionadas con las mediciones cardiovasculares y antropométricas tienen una fuerte influencia sobre la primera componente, de manera que altos valores en esta dimensión puede implicar problemas asociados con sobrepeso, obesidad y enfermedades cardiovasculares. La segunda dimensión tiene una clara diferenciación en medidas antropométricas en el semieje negativo y medidas cardiovasculares y de salud mental. De manera que una alta puntuación negativa en esta dimensión (eje y) indica un individuo más asociado con medidas antropométricas, mientras que una alta puntuación positiva puede indicar problemas de salud mental.

Una vez analizado el comportamiento de las variables entre sí, se analiza como las variables explican el comportamiento promedio de los individuos gracias al uso de los biplots modificados. En el contraste de los individuos discriminados por género (**1**. Masculino, **2**. Femenino) con respecto a las variables, Figura 4.10, se aprecia que los hombres tiene mayor relación con las variables antropométricas, mientras que las mujeres presentan mayores índices de ansiedad y depresión y se ven mayormente influenciadas por las medidas cardiovasculares. De manera que es posible encontrar con más frecuencia a las mujeres asociadas con enfermedades mentales y a los hombres con sobrepeso, además, de que los hombres se caracterizan por una alta estatura. Cabe anotar que estos resultados están de acuerdo con estudios encontrados en la literatura, a saber que las mujeres presentan mayores índices de depresión y ansiedad y para algunas poblaciones las mujeres tienen más incidencia en relación a problemas cardiovasculares (McLean et al., 2011; Albert, 2015).

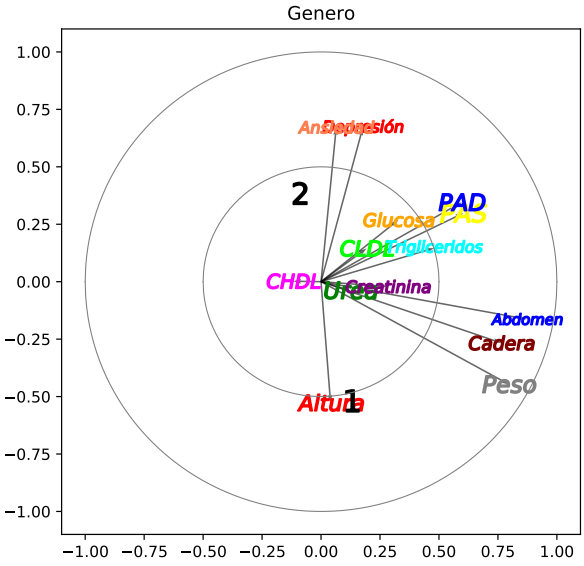

Figura 4.10: Biplot con centroides. Datos discriminados por género: **1**. Masculino, **2**. Femenino.

En la Figura 4.11 se presenta el biplot estratificado por edad. Se observa el comportamiento de los cuatro grupos de edad (**1**. 18 a 30, **2**. 31 a 43, **3**. 44 a 56, **4**. 57 a 100). Es necesario resaltar el hecho de que a medida que los individuos aumentan su edad obtienen un mayor nivel en las mediciones cardiovasculares. Ya que una alta puntuación en estas variables puede indicar problemas de salud cardiovascular, específicamente el grupo de mayor rango de edad presenta grandes afectaciones de estas características al igual que altas mediciones en los índices de depresión y ansiedad. Vale notar que los grupos más jóvenes presentan una mayor estatura y mediciones muy bajas en índices de depresión y ansiedad y medidas cardiovasculares. En conclusión a medida que la población aumenta de edad su calidad de salud se ve disminuida no solo físicamente sino también mentalmente. La literatura ha fundamentado a través de diversos estudios que la población adulta es particularmente susceptible a las enfermedades cardiovasculares. La edad es un factor de riesgo independiente de enfermedad cardiovascular (ECV) en adultos, pero estos riesgos son agravado por factores adicionales, que incluyen fragilidad, obesidad y diabetes. Estos factores son conocidos para complicar y potenciar los factores de riesgo cardíaco asociados con el inicio de la edad avanzada. El género es otro factor de riesgo potencial en los adultos mayores, dado que se informa que las mujeres mayores se encuentran en un mayor riesgo de ECV que los hombres de la misma edad. Sin embargo, tanto en hombres como en mujeres, los riesgos asociados con ECV aumentan con la edad (Rodgers et al., 2019).

En la Figura 4.12 se presenta el biplot de la variable raza (**1**. Blanca, **2**. Negra, **3**. Mulata, **4**. Amarilla, **5**. Indígena, **6**. Mestiza). En este caso se presentan comportamientos muy diversos para cada raza. En el caso de la raza blanca y mulata los comportamientos son promedios, es decir, ayudan como referentes en el análisis de las demás razas. A partir de este hecho, se puede resaltar que los individuos de raza negra tienen altas mediciones de variables asociadas con enfermedades cardiovasculares, lo cual está de acuerdo con los hallazgos de la literatura. También tiene asociación con los índices de enfermedades mentales. La raza mestiza por su parte presenta influencia en los individuos principalemente por las medidas antropométricas y medidas cardiovasculares, siendo vulnerables a enfermedades relacionadas con sobrepeso o problemas cardiacos. La raza amarilla (4) y la raza indígena (5) presentan un perfil más saludable que las demás razas. La raza amarilla presenta un comportamiento promedio en cuanto a los indicadores de enfermedades mentales, sin embargo, su mayor ventaja son los bajos niveles de puntuación en las medidas antropométricas, cardiovasculares y renales. La

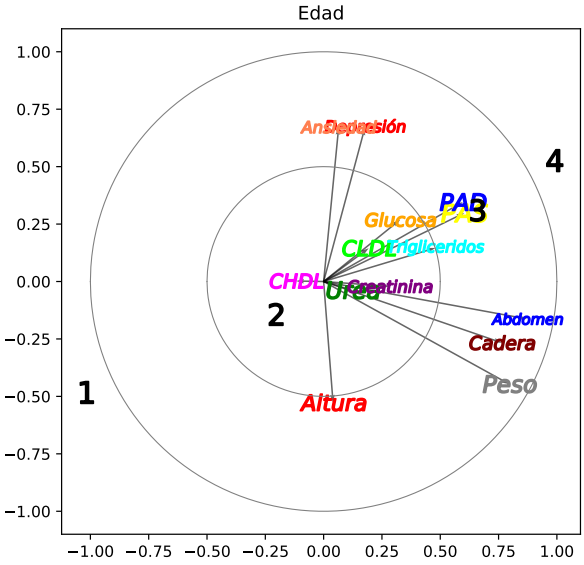

Figura 4.11: Biplot con centroides. Datos discriminados por rangos de Edad en años: **1**. 18 a 30, **2**. 31 a 43, **3**. 44 a 56, **4**. 57 a 100.

raza indígena por su parte presenta una gran estatura y mediciones bajas en los indicadores de depresión y ansiedad, no obstante, presenta asociados con las variables peso, abdomen y cadera, lo que indica señales de sobrepeso en la población. En relación a la altura de esta raza aborigen, específicamente la media fue de 173.25 cm, desviación estándar de 16.39 cm y la mediana fue de 179.50 cm, valores superiores a las otras razas de la población de Baependi. Se recomienda un análisis más específico ya que estas medidas no están de acuerdo con el estudio publicado en Oliveira (2014) donde se muestra que la altura media de indígenas del centro de Brasil fue de 165.7cm con una desviación estándar de 6.8 cm. Por otro lado, el mismo estudio presenta tendencia alta de enfermedades cardiovasculares entre los indios de esa región. Lackland (2014) afirma que la disparidad racial en la hipertensión y resultados relacionados (por ejemplo: diabetes, sobrepeso, triglicéridos, colesterol) se han reconocido durante décadas entre los afroamericanos con mayores riesgos que los caucásicos y otras razas como los asiáticos. Los niveles de presión arterial han sido consistentemente más altos para los afroamericanos con un inicio más temprano de hipertensión.

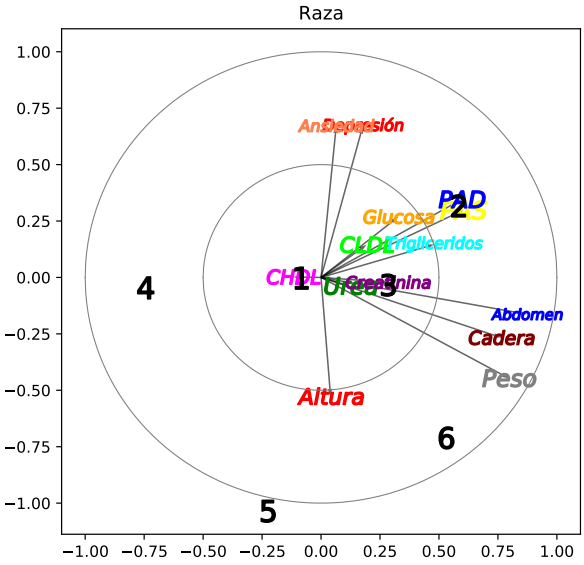

Figura 4.12: Biplot con centroides. Datos discriminados por raza: **1**. Blanca, **2**. Negra, **3**. Mulata, **4**. Amarilla, **5**. Indígena, **6**. Mestiza.

Para estudiar el comportamiento de los individuos en una muestra se le recomienda al investigador definir una región de confianza en la que puedan ser identificados comportamientos atípicos para la población. Esta región de confianza está definida por una elipse que es calculada a partir de la variabilidad de los datos y cuyo centro es su media. Todo dato que se encuentre por fuera de la elipse se considera fuera de control bajo la filosofía de la gestión de calidad. Sin embargo, en este análisis serán identificados como *outliers* y harán referencia a individuos con mediciones muy por encima del promedio. Con el fin de describir adecuadamente la variabilidad de los datos se suelen utilizar dos tipos de cartas de control con la metodología PCA. La primera es construida únicamente con las dos primeras componentes y se conoce como carta de formato elíptico o carta elipsoidal y la segunda se construye a partir de las demás componentes y se conoce como la carta T^2, sección 2.11. Se recomienda utilizar ambas cartas en conjunto ya que aunque la carta elipsoidal puede facilitar la visualización de los datos puede a su vez no contener una proporción de la variabilidad que sea lo suficientemente representativa como para describir el comportamiento de las observaciones

a plenitud. Es en este punto es necesario el uso de la carta de control T^2 donde se puede estudiar el restante de la variabilidad de los datos.

Con el fin de describir de forma más adecuada el comportamiento de las observaciones se replica tres veces la carta de control de formato elipsoidal discriminando los individuos nuevamente por género, edad y raza. Esta discriminación ayuda a entender un poco más el comportamiento de la muestra. Por ejemplo, en la Figura 4.13 se presenta la carta estratificada según el género. Se observa que la mayor proporción de individuos con datos atípicos son mujeres. En la Figura 4.14 se presenta la estratificación por edad. Se observa que los individuos por fuera de la elipse son en menor medida aquellos con edades entre los 31 y 43 años de edad (codificados como 2). Por último, en la Figura 4.15 se presenta la carta asociada con raza. Se muestra una alta variabilidad de los individuos pertenecientes a la raza blanca (1), seguidos de la raza mulata (3) y negra (2) y no presenta individuos de las razas amarilla, indígena y mestiza por fuera de la elipse, de manera que estas razas no presentan comportamientos atípicos respecto a la población.

Sin embargo, el especial interés de utilizar cartas de control como una herramienta descriptiva consiste en identificar comportamientos atípicos de los individuos. Por tal motivo, se decidió utilizar las codificaciones asignadas a cada individuo durante la recolección de los datos, de manera que se pueda estudiar a profundidad los motivos de sus comportamientos extremos. Para ilustrar el uso de la herramienta se estudiará a profundidad los individuos identificados por 9101, 77101 y 35801. Gracias al uso de las variables categóricas para la discriminación de los individuos se sabe de ante mano que el individuo 9101 es una mujer entre 44 a 56 años de raza blanca, la particularidad este individuo es que presenta un peso de 137.2 kg, una altura de 167 cm, un perímetro abdominal de 139 cm y radio de cadera de 160 cm, todos los indicadores apuntan a una alta obesidad en el individuo, el IMC de esta mujer es de 49.8 para su estatura un peso normal debería estar entre 51.6 kg y 69.4 kg. En el caso del individuo 77101 se trata de una mujer, blanca entre 31 y 43 años con valores igualmente altos en sus medidas antropométricas, un peso de 120 kg, altura de 165 cm, abdomen de 132 cm y cadera de 138 cm, presenta además una medida de triglicéridos de 302 mg/dL y una presión arterial sistólica de 139.7 mmHg. De igual forma el individuo 35801 se trata de una mujer en este caso de raza negra mayor de edad, entre 57 y 100 años, presenta mediciones cardiovasculares altas a diferencia de las dos mujeres anteriores con un colesterol de baja densidad de 185.2 mg/dL, triglicédiros de 285.3 mg/dL presión arterial sistólica de

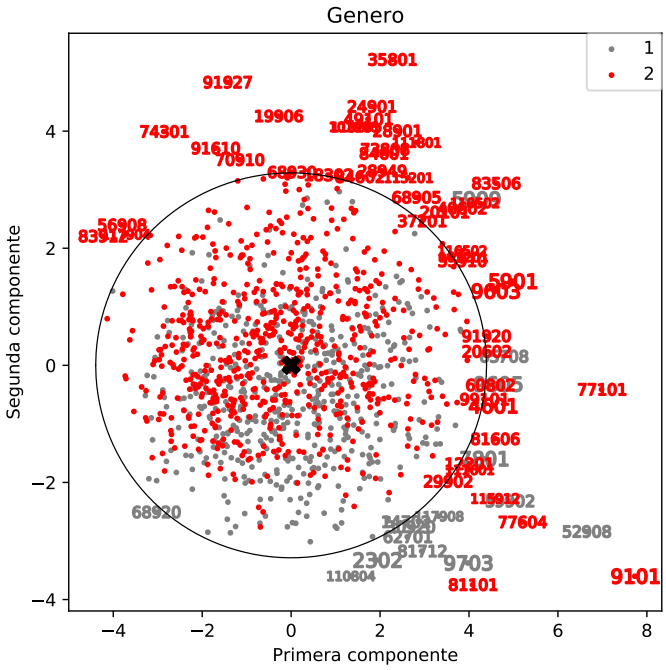

Figura 4.13: Gráfico de los individuos estratificados por género: 1. Masculino, 2. Femenino.

195.3 mmHG, además, presenta una alta puntuación en los indicadores de depresión y ansiedad con valores de 14 y 18 respectivamente. Estadísticamente es interesante observar que tanto el individuo 9101 y como el 77101 se ubican en el cuadrante correspondiente al semieje positivo de la primera componente y el semieje negativo de la segunda componente donde en los gráficos de biplots se identificaron preponderancias de las variables referentes a las medidas antropométricas. Por su parte, el individuo 35801 se encuentra en el primer cuadrante en donde las variables relacionadas con las enfermedades cardiovasculares, renales y de salud mental tienen mayor influencia.

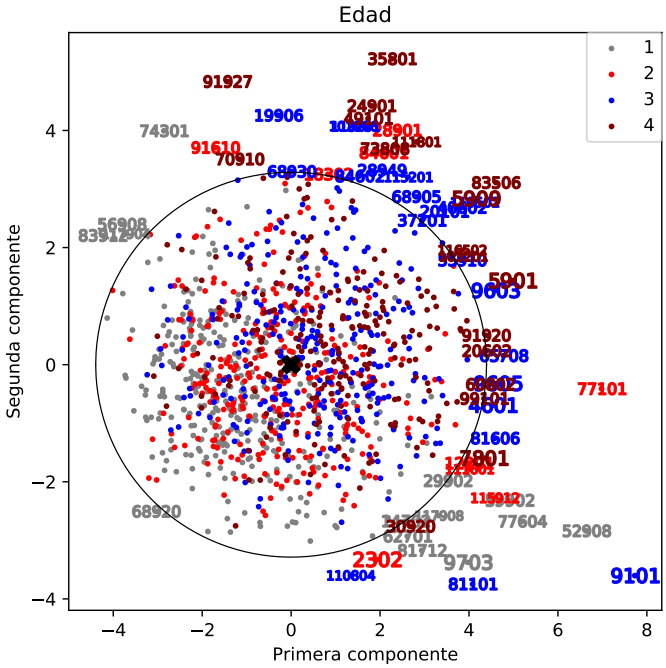

Figura 4.14: Gráfico de los individuos estratificados por rangos de Edad en años: **1**. 18 a 30, **2**. 31 a 43, **3**. 44 a 56, **4**. 57 a 100.

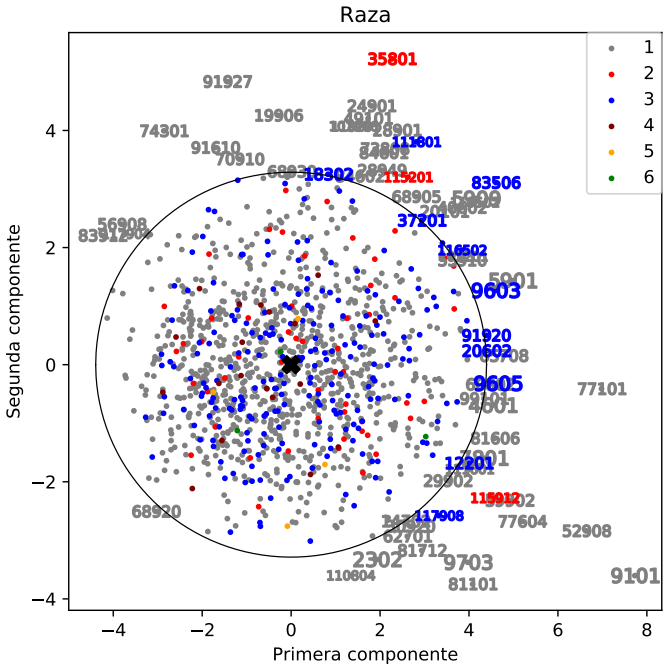

Figura 4.15: Gráfico de los individuos estratificados por raza: **1**. Blanca, **2**. Negra, **3**. Mulata, **4**. Amarilla, **5**. Indígena, **6**. Mestiza.

A pesar de que la carta elíptica logra identificar comportamientos atípicos en los datos, gran cantidad de información de éstos se encuentra en las componentes restantes y por medio del uso de la carta T^2 es posible estudiar esta variabilidad. En la Figura 4.16 se presentan cinco valores atípicos con un comportamiento realmente fuera de lo común a los cuales la carta elipsoidal no identificó fácilmente. Con finalidades de profundizar en el comportamiento de los datos se verán los individuos 68601 y 14601 en detalle. El individuo 68601 es un hombre blanco perteneciente al grupo de mayor edad, presenta peso de 68 kg con una altura de 171 cm, un perímetro abdominal de 88 cm, circunferencia de cadera de 102 cm, glucosa de 132.5 mg/dL, urea de 198 mg/dL y 219 mg/dL de triglicéridos. Su comportamiento altamente atípico se debe a su alto nivel de urea cuya media es 27.37 mg/dL en la muestra y una desviación estándar de 10.63 mg/dL, de forma que la medición de este individuo se ubica a 18 desviaciones estándar del promedio. Por su parte, el individuo 14601 es un hombre de raza blanca entre los 44 y 56 años con presencia de valores atípicos en mediciones como la glucosa con 196.8 mg/dL, urea de 143.4 mg/dL, creatinina 1.71 mg/dL y triglicéridos de 221 mg/dL. La variable cuya mayor variabilidad aporta es la creatinina que se encuentra a aproximadamente 5 desviaciones estándar de su promedio. Ya que la construcción del estadístico T^2 consiste en calcular que tan alejado se encuentra una observación de su promedio en términos de desviaciones estándar y luego realizar la suma al cuadrado de cada variable es factible darse una gran idea de cuál es la razón de este comportamiento analizando el residual de la medición para cada variable como se presentó anteriormente. Una descripción de la estadística T puede ser vista en el Apéndice, sección A.2.

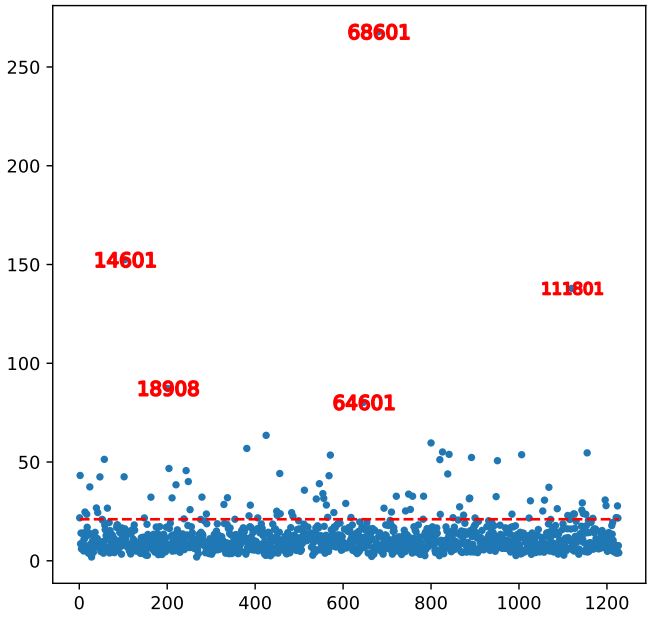

Figura 4.16: Carta de control T^2. La línea roja indica el límite de control definido para el estadístico T^2.

4.3. Análisis de Componentes Principales No Lineales

Para el PCA no lineal (NLPCA) es importante en extremo definir los niveles de análisis de cada una de las variables que se van a utilizar en el estudio. Los niveles de análisis pueden ser definidos por el investigador conociendo el origen y la naturaleza de las variables utilizadas o dejando que las variables hablen por si mismas, realizando una primera corrida exploratoria del algoritmo con el nivel de análisis menos restrictivo de las variables, es decir, nominal simple. Una vez realizado el análisis exploratorio, se ve cómo se comporta cada uno de los gráficos de cuantificación generados y se debe seleccionar el nivel de análisis que mejor se ajuste a cada variable. En la Figura 4.18 se presenta la ilustración de las cuantificaciones. Por ejemplo, donde a pesar que la edad fue seleccionada como nominal múltiple se ve claramente que su comportamiento es creciente y ordenado, características propias de una variable ordinal.

Los niveles de análisis en este caso son presentados en la Tabla 4.1. Por necesidad del estudio, los niveles de género, raza y rango de edad serán tratados como nominal múltiples ya que permiten resumir el comportamiento del conjunto de individuos en cada dimensión discriminados por cada categoría de las variables. De este modo, se pueden replicar todos los análisis realizados por medio de las variables discriminatorias utilizadas en el PCA lineal dentro de un mismo procedimiento y, además, éstas aportarán su variabilidad en la construcción de las Componentes Principales en el caso de NLPCA. Un detalle a tener en cuenta a la hora de utilizar variables nominales múltiples es que tendrán tantas dimensiones de cuantificación como categorías contengan las variables.

La cantidad máxima de dimensiones que se pueden obtener en el análisis será la suma de la cantidad de categorías que conforman cada una de las variables nominales múltiples más el resto de variables utilizadas en el análisis. En este caso, se tendrá un máximo de 31 dimensiones de las cuales 12 se deben a las variables nominales múltiples y 19 a las demás variables. El algoritmo le solicita al investigador introducir la cantidad de dimensiones a incluir en el análisis desde un principio, si se utiliza la totalidad de dimensiones es probable que todas las cuantificaciones converjan a un comportamiento “ordinal” inducido por un ajuste perfecto en la solución del problema de optimización, ya que por la naturaleza de las variables no es

Tabla 4.3: Proporción de la varianza explicada por las Componentes Principales No Lineales

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Varianza	3.696	2.494	1.835	1.504	1.413	1.297	1.244	1.120
Proporción de la varianza	0.132	0.089	0.065	0.054	0.050	0.046	0.044	0.040
Proporción Acumulada	0.132	0.221	0.286	0.340	0.390	0.436	0.480	0.520
	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16
Varianza	1.101	1.040	1.018	1.006	1.002	0.908	0.892	0.849
Proporción de la varianza	0.039	0.037	0.036	0.036	0.036	0.032	0.032	0.030
Proporción Acumulada	0.559	0.596	0.632	0.668	0.704	0.736	0.768	0.798
	PC17	PC18	PC19	PC20	PC21	PC22	PC23	PC24
Varianza	0.814	0.757	0.708	0.685	0.573	0.560	0.452	0.399
Proporción de la varianza	0.029	0.027	0.025	0.024	0.020	0.020	0.016	0.014
Proporción Acumulada	0.827	0.854	0.879	0.903	0.923	0.943	0.959	0.973
	PC25	PC26	PC27	PC28				
Varianza	0.232	0.220	0.151	0.112				
Proporción de la varianza	0.008	0.008	0.005	0.004				
Proporción Acumulada	0.981	0.989	0.994	0.998				

lógico que todas conserven un comportamiento ordinal, se recomienda trabajar con una cantidad de dimensiones que conserven una proporción de variabilidad que sea interesante para el investigador, entre ambos análisis habrá una clara diferencia respecto a la cuantificación de las variables, principalmente porque el algoritmo no es anidado. La selección de las variables a trabajar se puede definir corriendo la totalidad de las dimensiones una vez para seleccionar cuántas de ellas conservan la cantidad de variabilidad deseada y posteriormente utilizar el número de dimensiones correspondientes para generar un nuevo análisis. De esta manera para el presente caso se seleccionó 16 dimensiones que conservan aproximadamente el 80 % de la variabilidad de los datos sobre un total de 28 obtenidas por el programa. El criterio de selección de número de componentes puede ser realizado en término de dimensiones, componentes a conservar, o información contenida, en este caso la medida de información es la variabilidad. En la literatura no existe un criterio robusto de la variabilidad a conservar pero a menudo se recomienda que sea entre un 80 % y 90 % de ella (Johnson & Wichern, 2007; Diaz & Morales, 2012), observar la Figura 4.17.

Una vez corrido el análisis, cada categoría de cada variable corresponderá con una cuantificación que permitirá el tratamiento de las variables cualitativas como numéricas. En la Figura 4.18 se presenta la ilustración de las cuantificaciones, se puede ver que variables como edad y escolaridad presentan comportamientos ascendentes en todo instante, este tipo de cuantificación es una cuantificación monotónica, en la que se respeta el orden de jerarquía

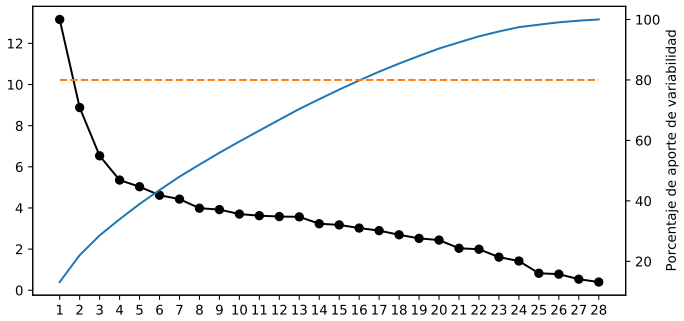

Figura 4.17: Gráfico de sedimentación para la totalidad de dimensiones.

de los valores iniciales. Para el caso de variables como raza, estado civil (codificada como Estado), licor (consumo) y renta, las gráficas presentan un comportamiento no monotónico, en donde no se respeta el ordenamiento original de las categorías. La ventaja de utilizar un nivel de cuantificación nominal es que las variables se cuantificarán de la forma más adecuada ya que no tienen mayor cantidad de restricciones a diferencia de las variables ordinales y, por lo tanto, retendrán la mayor cantidad de variabilidad posible.

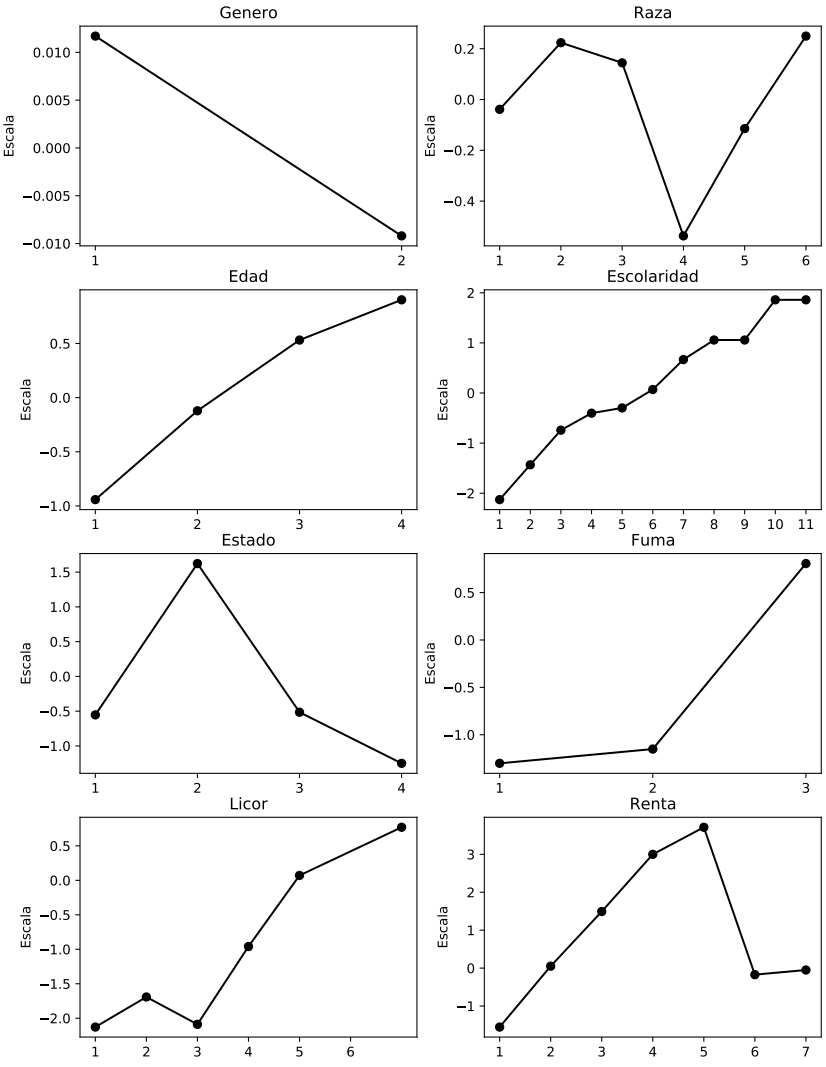

Figura 4.18: Cuantificaciones de las variables cualitativas. En el eje x se presentan las categorías de cada variable y en el eje y el valor de su cuantificación. Las variables Estado y Fuma presentan un comportamiento monotónico al ser de naturaleza ordinal.

Ahora se estudiará la relación entre las variables y los individuos discriminados para cada categoría de las variables género, edad y raza como se realizó en el PCA lineal. Para tal finalidad se seleccionó el nivel de análisis nominal múltiple para estas tres variables. En este punto se presenta una gran ventaja del NLPCA sobre el PCA y consiste en que en un solo análisis se pueden obtener todos los centroides de las variables nominales múltiples que son utilizadas para discriminar el comportamiento de la población en los diferentes grupos de interés. Para comenzar con el análisis el foco de atención serán las relaciones presentes entre las diferentes variables por medio del uso de Biplots modificados. Si el investigador desea se podría generar un solo Biplot con los centroides de todos las diferentes variables, pero para este caso se presentará un Biplot para cada una de las variables nominales múltiples, Figuras 4.19, 4.20 y 4.21.

Recordando el hecho de que dos vectores paralelos entre sí indican una alta correlación entre las variables y dos vectores perpendiculares una correlación nula, se puede concluir que variables antropométricas como la medida del abdomen, la cadera, las presiones sistólica y diastólica, los triglicéridos y el colesterol de alta densidad se encuentran altamente correlacionados y que variables como la renta va en sentido opuesto al consumo de licor y la ansiedad y la depresión. Por otro lado, el peso y el consumo de tabaco se encuentran altamente relacionados de manera inversa, pero estos a su vez tienen una baja relación con la escolaridad y el estado civil. En general, sobre el semieje positivo de la primera componente se encuentran todas aquellas variables relacionadas con las medidas antropométricas y cardiovasculares en las cuales al obtener una puntuación alta el individuo es susceptible a sufrir de problemas cardiacos y de sobrepeso. Sobre el semieje negativo de la segunda componente se encuentran variables relacionadas con la presencia de enfermedades mentales y la adquisición de adicciones.

En la Figura 4.19 se presenta el biplot estratificado por género. Se aprecia que los hombres presentan una relación con las variables altura, renta, estado civil y escolaridad. Por otro lado, llama la atención que las mujeres están más asociadas con el consumo de licor, el hábito de fumar y las enfermedades mentales de depresión y ansiedad. Resaltando que el comportamiento de los hombres es altamente contrastado por el de las mujeres y que en cuanto a enfermedades de tipo cardiovascular ambos grupos están igualmente influenciados por estas variables.

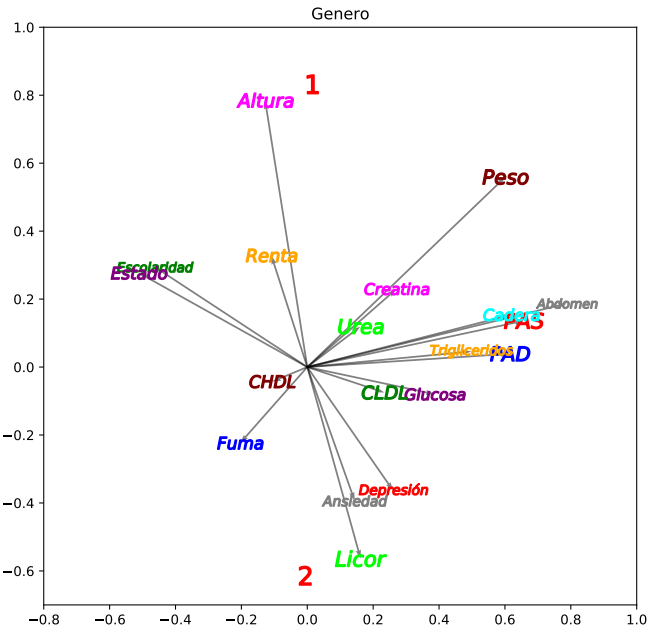

Figura 4.19: Biplot NLPCA estratificado por Género: 1. Hombres, 2. Mujeres.

En cuanto a la variable raza, la estratificación se presenta en la Figura 4.20. Las razas blanca, negra y mulata se encuentran dentro del comportamiento promedio de la población aunque vale resaltar que se ven altamente influenciados por el peso, la creatinina, la urea y la frecuencia con la que se fuma. Por otro lado, la raza amarilla presenta grandes niveles de escolaridad al igual que de depresión y ansiedad con alta tendencia al consumo de licor y tabaco. Por último, los indígenas y los mestizos presentan un gran valor en la altura, sin embargo, los segundos presentan mayores valores en características antropométricas como el peso y el perímetro abdominal, en variables cardiovasculares como la presión diastólica y sistólica y glucosa, creatinina, urea, tiglicéridos y colesterol de baja densidad. De esta manera las razas indígenas y mestizas presentan tendencias a enfermedades de varios tipos; la blanca, amarilla e indígena mejores niveles de salud mientras que la negra, mulata y mestiza mayores implicaciones de salud cardiovascular.

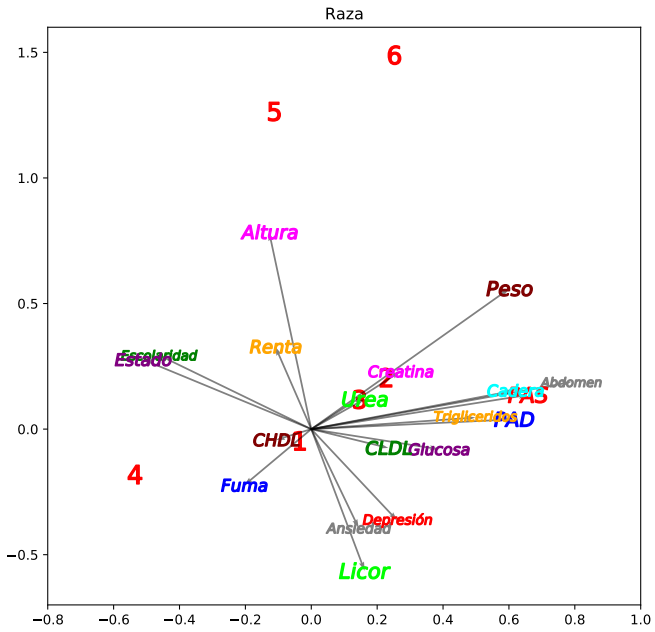

Figura 4.20: Biplot NLPCA estratificado por la variable Raza: **1**. Blanca, **2**. Negra, **3**. Mulata, **4**. Amarilla, **5**. Indígena, **6**. Mestiza.

En la Figura 4.21 es presentada la estratificación por edad. El grupo de personas más jóvenes presentan mayor nivel de escolaridad, altura y peso y se ve altamente influenciado por la renta. En cuanto el segundo grupo de personas presentan un comportamiento cercano al promedio de la población, con buenos niveles de CHDL o colesterol bueno. El tercer grupo se caracteriza por presentar altos niveles de medidas antropométricas y de mediciones cardiovasculares lo que implica que es un grupo de alta susceptibilidad a problemas de sobrepeso y cardiovasculares. Por último, el cuarto grupo, conformado por las personas de mayoría de edad en la población, pierden un poco de influencia de las medidas antropométricas pero se ven mayormente afectadas por el consumo de licor y tabaco y son más susceptibles a enfermedades mentales como la depresión y la ansiedad, siendo este el segmento de la población con mayor riesgo y menor nivel de educación.

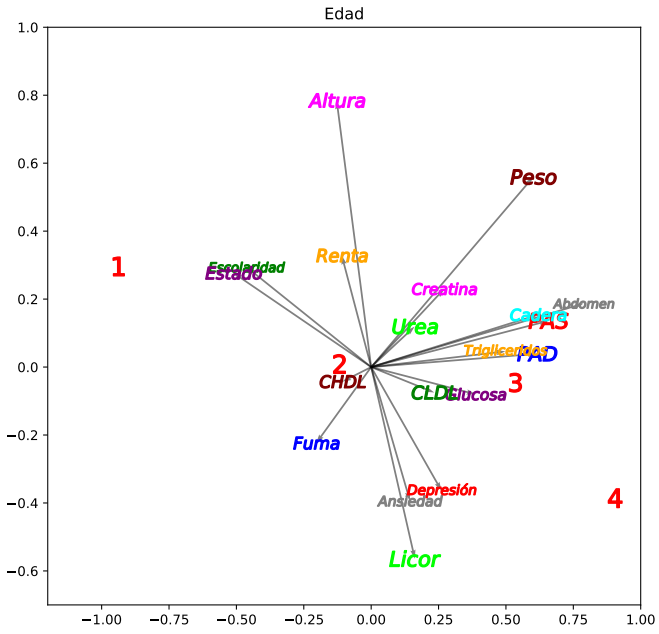

Figura 4.21: Biplot NLPCA estratificado por los rangos de la variable edad en años: **1**. 18 a 30, **2**. 31 a 43, **3**. 44 a 56, **4**. 57 a 100.

Si se desea ahondar un poco más en el comportamiento de los individuos, las cartas de control son una herramienta adecuada ya que ayudan a definir un límite en la variabilidad de los datos sobre el cuál se puede identificar de forma más explícita comportamientos atípicos. A partir del NLPCA es posible construir cartas de control tanto para las variables cuantificadas óptimamente como para las componentes obtenidas. El primer tipo de cartas de control son presentadas en la sección 3.6 y permite identificar patrones en grupos de individuos ya que las variables cuantificadas están muy limitadas a la hora de asignar valores numéricos a las categorías de cada variable. Por otro lado, las cartas de control basadas en componentes, sección 3.7, pueden ayudar a identificar sucesos fuera de lo común en las observaciones individuales gracias a que se pueden involucrar variables numéricas sin inconveniente alguno en el análisis.

En el presente estudio se analizará el comportamiento de las variables rango de edad, cantidad de renta obtenida por mes, nivel de escolaridad y frecuencia de consumo de tabaco por medio de una carta de control elipsoidal. Para su análisis se crearán tres cartas de control que asocien a dos características de interés. En este caso serán las relaciones: Escolaridad - Edad; Escolaridad - Renta; Fuma - Edad. Para apoyar la interpretación de las gráficas se construye la Tabla 4.4 en donde se relacionan las cuantificaciones y las categorías de cada variable. Vale resaltar que cada punto en las cartas de control corresponderá a un conjunto de individuos que pertenecen a una categoría en cada variable, funcionando las cuantificaciones como coordenadas. En este caso la elipse funciona como límite de control y permite identificar comportamientos atípicos en los grupos de individuos.

Tabla 4.4: Cuantificaciones de las variables: Edad, Renta, Escolaridad y Fuma.

Categoría	Edad	Renta	Escolaridad	Fuma
1	-0.9415	-1.5542	-2.1274	-1.3016
2	-0.1213	0.0513	-1.4329	-1.1510
3	0.5316	1.4917	-0.7416	0.8064
4	0.9031	2.9994	-0.4023	
5		3.7130	-0.2978	
6		-0.1726	0.0698	
7		-0.0499	0.6677	
8			1.0577	
9			1.0577	
10			1.8597	
11			1.8597	

La carta de control presentada en la Figura 4.22, presenta cuatro grupos de individuos por fuera de la elipse de control. Los grupos presentes en la esquina superior derecha, ubicados en el exterior de la elipse, tienen una escolaridad superior completa e incompleta con un rango de edad entre 44 y 100 años de edad. Por otro lado, los grupos de la esquina inferior izquierda están asociados con individuos entre los 18 a los 30 años con un nivel de escolaridad mínimo relacionado con las competencias básicas de escritura y lectura o incluso con el analfabetismo. El comportamiento de estos conjuntos de datos es atípico en ambos casos ya que las generaciones más recientes presentan altos niveles de escolaridad.

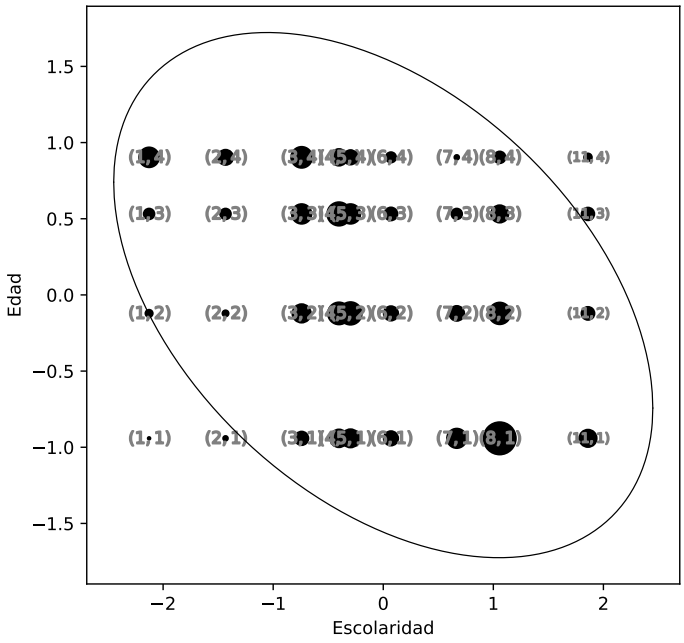

Figura 4.22: Carta de control Escolaridad - Edad.

La carta de la Figura 4.23 presenta la relación entre las variables escolaridad y renta donde los ingresos superiores a los diez salarios mínimos tienen comportamientos atípicos sin importar el nivel de escolaridad, al igual que se aprecia que son pocos los individuos que tienen este nivel de ingresos. De la misma forma, hay dos pequeños grupos de individuos que tienen ingresos entre cinco y diez salarios mínimos con un bajo nivel de escolaridad y otro que por el contrario presenta el mínimo nivel de ingreso para un nivel de escolaridad alto. Vale resaltar que el máximo nivel de ingreso presentado para una persona analfabeta es de cinco a diez salarios y que los máximos niveles de ingreso presentados se relacionan con personas con un nivel de escolaridad técnico o superior.

Las variables relacionadas con la frecuencia de consumo de tabaco y rango de edad no presentan comportamientos de especial interés en cuanto a variabilidad de los datos, ver la Figura 4.24. Sin embargo, se debe resaltar que

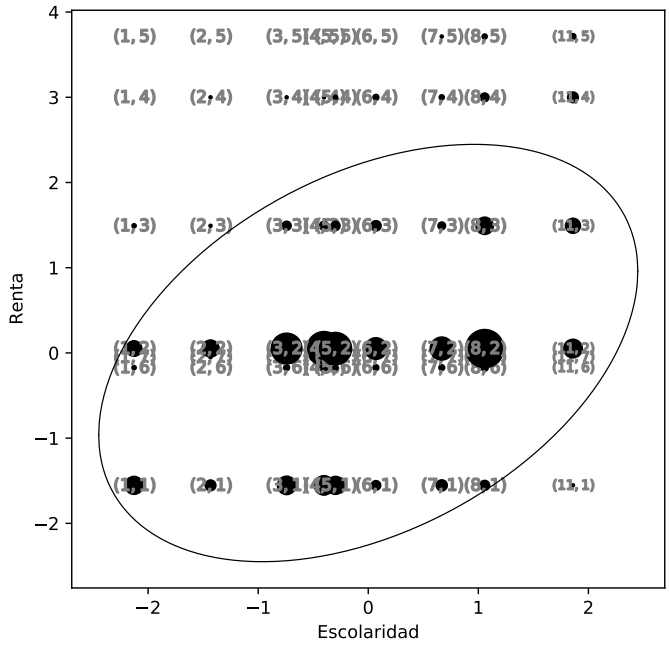

Figura 4.23: Carta de control Escolaridad - Renta.

en toda la población es más frecuente el no consumo de tabaco en especial entre las personas más jóvenes y que el rango de edad entre 44 y 56 años presenta mayor consumo respecto a los demás grupos.

La carta de control T^2 para todas las variables óptimamente escaladas contendrían en teoría la misma información que las cartas de control T^2 para todas las dimensiones de las componentes principales generadas, siempre y cuando no se trabajen niveles de análisis nominales múltiples. Esto se debe a que bajo estas condiciones la varianza total de las variables cuantificadas será igual a la cantidad de variables utilizadas en el análisis y por ende al número de componentes generadas.

De forma que para el estudio de la variabilidad de los individuos se optará por hacer uso de las cartas presentadas en la sección 3.7, basadas en la metodología NLPCA. Para evitar generar más dimensiones que las contenidas

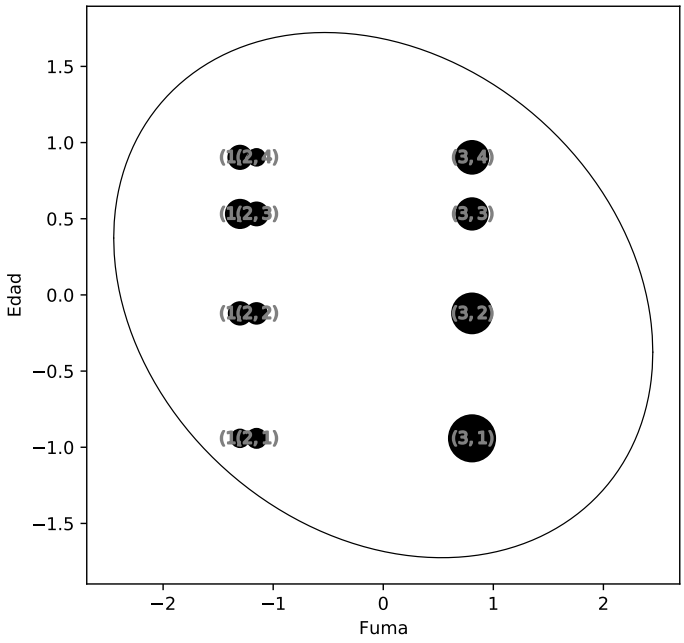

Figura 4.24: Carta de control Fuma - Edad.

por el conjunto inicial de datos se propone evitar trabajar con variables nominales múltiples, de esta forma la variable edad tomará un comportamiento ordinal mientras las variables género y raza tendrán niveles nominales simples. La distribución de la variabilidad para cada componente se presentan en la Figura 4.25 donde se presenta el gráfico de sedimentación. Como se ve en el gráfico las dos primeras componentes solo conservan el 28 % de la variabilidad, por lo tanto, no basta con la construcción de una carta de control elipsoidal para la identificación de datos atípicos, es necesario complementar el análisis con una carta de control T^2 en donde se analice la influencia de las 20 componentes restantes.

Las cartas de control basadas en NLPCA capturan aún más variabilidad que las cartas de control basadas en PCA gracias a la inclusión de las variables óptimamente escaladas. Sin embargo, si la variabilidad presente en las variables cuantitativas es bastante alta es de esperarse que algunos de los

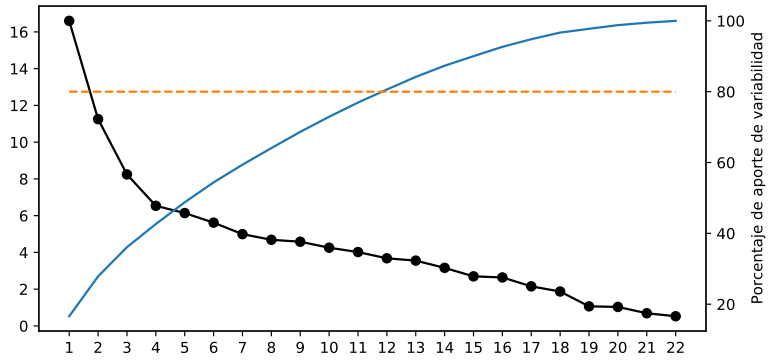

Figura 4.25: Gráfico de sedimentación para la totalidad de dimensiones.

datos identificados como atípicos por las cartas de la sección 4.2. Por ejemplo, en la carta elipsoidal presentada en la Figura 4.26 aparecen nuevamente los individuos 9101 y 35801 quienes fueron identificados como atípicos en la gráfica 4.14. Respecto a la carta presentada en la Figura 4.27, los individuos 9101, 18908 y 6401 resaltan nuevamente por su alta variabilidad.

El individuo 9101 presenta un alto valor en peso, circunferencia abdominal y perímetro de cadera separados por una distancia de 5.43, 4.28 y 6.28 desviaciones estándar respecto al valor promedio, de manera que es su gran talla lo que lo lleva a sobresalir nuevamente sin tener una gran influencia de las variables categóricas. Por su parte, el individuo 35801 resalta por ser una mujer perteneciente al cuarto grupo etario de raza negra, la raza negra presenta una alta variabilidad respecto a las demás variables, lo que ubica al individuo a 2.05 desviaciones de la media; en cuanto a las variables cuantitativas su alto nivel de urea resalta por 16.06 desviaciones por encima del promedio. La alta variabilidad del individuo 18908 en este caso viene dada por sus bajas medidas de perímetro de cadera, 51 cm, y niveles de creatinina 0.04 mg/dL, que corresponden a -4.72 y -2.23 distancias en términos de desviaciones, y por su alto nivel de renta, de 10 a 20 salarios mínimos, que lo ubica a 3 desviaciones estándar sobre la media.

La carta presentada en la Figura 4.26 identifica entre varias observaciones al individuo 16703 como dato de interés. Se trata de un hombre de raza blanca entre los 18 y 30 años con una renta de 10 a 20 salarios mínimos con

una una escolaridad superior incompleto, soltero, quien bebe licor de 2 a 3 veces por semana cuyos mayores aportes a su variabilidad provienen de las variables edad, escolaridad, estado civil, consumo de licor y principalmente renta. En cuanto a las variables cualitativas los mayores aportes provienen de la estatura, con una medida 180 cm, y los puntuaciones de ansiedad y depresión con 1 en cada uno. Sin embargo, la única medición que supera una distancia de 3 desviaciones estándar respecto al centroide es la renta, el resto de variables tienen un valor entre 1 y 2 desviaciones. El individuo 64601 se trata de una mujer blanca, con edad entre 31 y 43 años y un ingreso mínimo, con escolaridad de nivel técnico 1, quien fumaba en el pasado. La mayor variabilidad corresponde a la renta y el consumo de tabaco con unas distancia del -1.55 y -1.30 respecto a la media. En cuanto a las variables cuantitativas las medidas de abdomen y cadera sobresalen con valores de 57 cm y 55 cm, ubicando al individuo 2.46 y 4.31 desviaciones por debajo de la media.

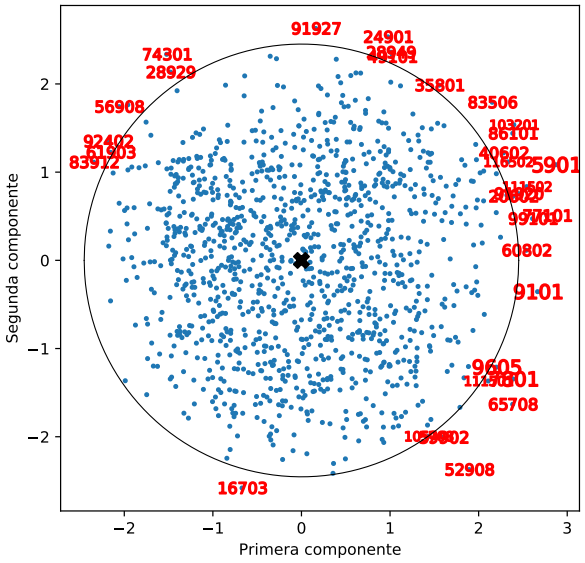

Figura 4.26: Carta de control elipsoidal.

La carta presentada en la Figura 4.27 estudia la variabilidad restante en los datos y, por lo tanto, permite identificar valores atípicos diferentes a

los de la carta elipsoidal, los cuales serán ubicados por encima del nivel de confiabilidad marcado por el estadístico chi cuadrado. Uno de los individuos sobresalientes corresponde a la observación 18908 la cual se trata de un hombre blanco entre los 21 y 43 años con una escolaridad superior incompleta y una renta entre los 10 y 20 salarios mínimos. Tiene una altura de 159 cm con una circunferencia abdominal de 63 cm, un perímetro de cadera de 51 cm y un bajo valor de creatinina de 0.04 mg/dL. El individuo 64601 es una mujer blanca perteneciente al segundo grupo etário, con ingresos mínimos y nivel de educación 1 completo quien fumó en el pasado. Cuenta con una altura de 166 cm, una medida de abdomen de 57 cm y cadera de 55 cm. Las variables que aportan en mayor medida a su variabilidad son el abdomen, la cadera, la renta y la frecuencia de consumo de tabaco con distancias de 2.46, 4.31, 1.55 y 1.30 desviaciones por debajo de la media.

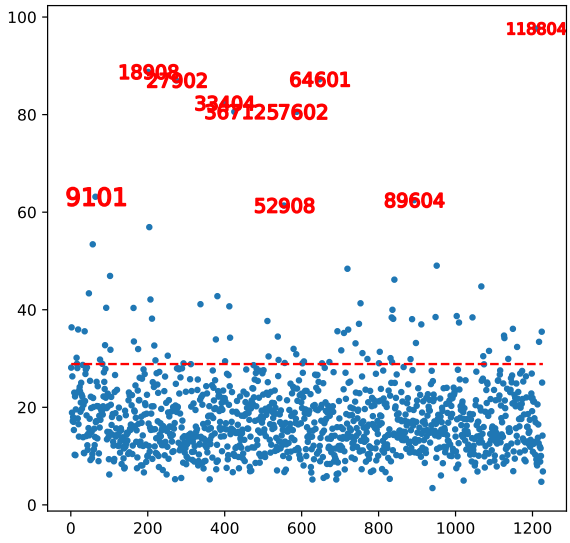

Figura 4.27: Carta de control T^2. La línea roja indica el límite de control definido para el estadístico T^2.

Como conclusiones generales de la aplicación de las metodologías al con-

junto de datos reales, se puede afirmar que los resultados obtenidos a lo largo de este estudio son consistentes con los encontrados en la literatura, resumiendo lo expuesto anteriormente, las mujeres presentan mayores índices de depresión y ansiedad así como mayor incidencia a problemas cardiovasculares y que estas enfermedades tanto en hombres como en mujeres aumentan con la edad (McLean et al., 2011; Albert, 2015; Rodgers et al., 2019). Por otro lado se ha reconocido que los afroamericanos tienen mayores tendencias a problemas de hipertensión cuando comprados con otras razas (Lackland, 2014).

Se citan algunos resultados y estudios que pueden ayudar a entender mejor el alcance de la investigación. Cabe mencionar que cifras elevadas de glicemia, PAS, PAD, CLDL, triglicéridos, urea y creatinina son factores de riesgo, y combinadas entre sí, son potencialmente perjudiciales para el organismo. Por ejemplo, la coexistencia de la Diabetes Mellitus e Hipertensión Arterial en el individuo es un factor de alto riesgo para enfermedades macro y microvasculares, con el consecuente aumento en el riesgo de enfermedades coronarias, cardiovasculares, insuficiencia cardíaca, enfermedades cerebrovasculares, periféricas e incluso la muerte. Para más información consultar Escribano et al. (2010), Lahera & de las Eras (2009), Cachofeiro (2009), Revista de Endocrinología y Nutrición (2004).

Capítulo 5

Conclusiones

En este trabajo de tesis de maestría se presenta la teoría de Análisis de Componentes Principales PCA desde el punto de vista de optimización. También se logró la formalización del análisis de Componentes Principales no Lineales NLPCA estableciendo un enlace desde varios puntos de vista de las dos teorías y además formulando un resultado asociado a Cartas de Control en el contexto del NLPCA. Las aplicaciones de esta teoría fueron realizadas a un conjunto de datos con variables mixtas relacionadas a enfermedades cardiovasculares del Proyecto "Corazones de Baependi" en que los resultados encontrados son relevantes y están de acuerdo con algunos hallazgos presentados en la literatura.

En el primer Capítulo se presentó el siguiente objetivo general:
"Establecer una relación directa entre la teoría de reducción de dimensión mixta NLPCA y su homónimo cuantitativo PCA".
Para lograrlo se plantearon 3 objetivos específicos que a continuación se describe cómo fueron alcanzados.

- En el Capítulo 2 del documento se dio una amplia y certera visión de los fundamentos matemáticos y estadísticos de la metodología del Análisis de Componentes Principales PCA desde el punto de vista de optimización, en donde también se incluyó la teoría asociada al Biplot tradicional. Por otro lado, se presentó la teoría de Cartas de Control asociadas al PCA y en el Apéndice, las secciones A.2. y A.3., se presentó el desarrollo teórico fundamental para la construcción de las cartas de control. Donde se resalta la presentación de la formalización estadística multivariada asociada con la distancia de Mahalanobis y la Estadística T^2 de Hotelling.

En el Capítulo 3, teniendo en cuenta el marco de referencia del PCA, se presentó el desarrollo teórico del Análisis de Componentes Principales no Lineales NLPCA. A diferencia del PCA, que se fundamenta en la descomposición espectral de la matriz de distancia de los datos, el NLPCA implica elementos extra como el uso de las metodologías OS, HOMALS y ALS que se encuentran presentes en el proceso de optimización, esto con el fin de alternar un paso de escalamiento óptimo que permitan a los datos cualitativos ajustarse a una escala cuantitativa y un paso de Descomposición Espectral que permitan la proyección de los datos en el nuevo subespacio. Se resalta que en las secciones 3.6 y 3.7 se presentan las Cartas de Control asociadas al NLPCA en donde se propone una metodología nueva para variables cualitativas, la cual permite la toma de decisiones por medio del control de su variabilidad y también ayuda a identificar patrones atípicos en los datos. Al final del Capítulo se plantea una propuesta para futuros trabajos basados en este resultado en donde se replantea el problema de optimización del PCA no lineal con las restricciones utilizadas en el PCA lineal.

- Al contrastar las metodologías PCA y NLPCA, se puede concluir que la aplicación metodológica del NLPCA varía en un número de características específicas respecto al PCA, como se ha evidenciado en los resultados de la aplicación presentados en el Capítulo 4. Resaltando el hecho de que el NLPCA tiene la posibilidad de crear un subespacio de mayor cantidad de dimensiones a diferencia del PCA que solo puede crear un subespacio de máximo la misma cantidad de dimensiones de los datos originales. Lo anterior es gracias al uso de la teoría de pérdida de reunión que se basa en la metodología HOMALS (Capítulo 3, sección 5.2) en donde los datos tienen la capacidad de ajustarse a cuantas dimensiones permita el rango columna de la matriz de codificaciones.

  Otro hecho importante es que en el PCA la matriz de cargas está restringida a ser ortonormal mientras que en el NLPCA es la matriz de puntuaciones a la que se le impone esta restricción. Esto implica que las medidas de variabilidad para cada componente será representada por una matriz distinta en cada técnica. En el análisis lineal la norma al cuadrado de cada componente representa el valor propio asociado a la misma, mientras que en el no lineal corresponde a la norma al cuadrado de los vectores de cargas.

- En la sección de la aplicación de este estudio, se utilizó el paquete Scikit Learn de Python y el módulo de reducción de dimensión de SPSS,

respectivamente y se obtuvieron los siguientes resultados principales:
1- Los resultados obtenidos a lo largo de este estudio son consistentes con los encontrados en la literatura en relación a la discriminación de los individuos según el género, la raza y el rango de edad teniendo en cuenta varias variables asociadas a depresión y ansiedad, variables asociadas a enfermedades cardiovasculares y también asociadas a medidas antropométricas.
2- Las enfermedades cardiovasculares y renales tienen una mayor prevalencia en los Afrodescendientes, mientras que Asiáticos e Indios presentan una menor prevalencia de las mismas.
3- La Depresión y Ansiedad tiene una mayor prevalencia en el sexo femenino y en general en individuos de mayor edad.
4- Se comparó el desempeño de los métodos PCA y NLPCA. Si bien en ambos casos los resultados concuerdan con los hallazgos de la literatura respecto a enfermedades cardiovasculares y mentales estratificadas según raza, género y edad y escolaridad, se observa una mejor discriminación y robustez del método NLPCA. Ambos métodos lograron discriminar a la raza Afrodescendiente que tienen factores de riesgo de enfermedades cardiovasculares y renales, pero el segundo método logró identificar de forma más contundente los individuos en las diversas estratificaciones. Por otro lado, ambos métodos identificaron que las mujeres tienen más tendencia a la depresión y a la ansiedad.
5- Finalmente, se resalta el uso de las cartas de control como una herramienta poderosa en la discriminación de individuos. En este trabajo se propuso la extensión de las cartas de control en el método NLPCA donde se vió la utilidad para analizar variables cualitativas según las puntuaciones dadas por el método.

Capítulo 6

Trabajos futuros

Durante el desarrollo teórico del análisis de componentes principales no lineales se ha dejado una cantidad de caminos abiertos para profundizar más en el desarrollo de esta técnica y sus posibles usos en diversas áreas de las ciencias y la ingeniería. Los más relevantes e inmediatos se consideran a continuación.

- Es necesario recordar en este punto que las transformaciones obtenidas del escalamiento óptimo por medio de regresiones monotónicas y no monotónicas simples pueden ser suavizadas si se utilizan en su lugar splines. La suavidad en las curvas de transformación puede ayudar a la comprensión de los resultados obtenidos durante el procesamiento de los datos. Es por esto que se ve necesario profundizar aún más en las implicaciones, no solo estadísticas sino también algorítmicas, del uso de Splines en el escalamiento óptimo de datos cualitativos.

- En el presente trabajo se decidió no abordar el procesamiento de bases de datos con datos faltantes en la metodología NLPCA, principalmente por la necesidad de fijar un marco teórico sencillo que permita hallar similitudes de forma directa con el PCA. Autores como Linting & van der Kooij (2012) y Mair (2018) presentan soluciones al problema de tratar con datos ausentes desde el marco del análisis de homogeneidad. Sin embargo, resulta interesante estudiar el problema por medio del uso de técnicas de imputación de datos clásicas en el algoritmo de mínimos cuadrados alternantes y su rendimiento respecto a la metodología tratada tradicionalmente.

- Se ve necesario profundizar aún más en las ideas inicialmente planteadas en torno al desarrollo e implementación de cartas de control

basadas en OS y el NLPCA. Principalmente, poniéndolas a prueba como herramienta de apoyo en problemas centrados en el control de calidad de bienes o servicios.

- Por último, gracias al reciente impacto que han tenido las tecnologías de la información en la actividad socio-económica de todas las naciones y por el creciente interés del uso de la ciencia de datos en diferentes contextos productivos e investigativos, es necesario estudiar el rendimiento del NLPCA como una herramienta más para el procesamiento de información en el contexto de la inteligencia artificial y la minería de datos.

Apéndice A

Estadística descriptiva e inferencial

A.1. Estadísticos descriptivos

Dado un vector aleatorio $\mathbf{x}_{(i)}$, su valor esperado $E(\mathbf{x}_{(i)})$, es el vector μ conformado por cada una de las medias de las variables aleatorias μ_j.

$$\boldsymbol{\mu} = E(\mathbf{x}_{(i)}) = \begin{bmatrix} E(x_{i1}) \\ E(x_{i2}) \\ \vdots \\ E(x_{ip}) \end{bmatrix} = \frac{1}{N} \begin{bmatrix} \sum_{i=1}^{N} x_{i1} \\ \sum_{i=1}^{N} x_{i2} \\ \vdots \\ \sum_{i=1}^{N} x_{ip} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix},$$

y se puede calcular a través de la matriz de datos por medio de

$$\boldsymbol{\mu} = \frac{1}{N}\mathbf{X}'\mathbf{1}_p,$$

donde N es el tamaño de la población y $\mathbf{1}_p$ es un vector columna de tamaño p con 1 en cada una de sus entradas. Ya que pocas veces conoceremos los datos de la población es necesario hacer uso de su estimador $\bar{\mathbf{x}}' = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_p \end{bmatrix}$, calculado por medio de

$$\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X}'\mathbf{1}_p.$$

La matriz de covarianzas de la variable aleatoria $\mathbf{x}_{(i)}$ es

$$\mathbf{\Sigma} = Cov(\mathbf{x}_{(i)}) = E\{(\mathbf{x}_{(i)} - \mu)(\mathbf{x}_{(i)} - \mu)\} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \dots & \sigma_{pp} \end{bmatrix},$$

cuyo cálculo se puede realizar de la forma

$$\mathbf{\Sigma} = \frac{1}{N}\mathbf{X}_c'\mathbf{X}_c,$$

donde $\mathbf{X}_c$ es la matriz de datos centrada en el origen

$$\mathbf{X}_c = \left(\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N'\right)\mathbf{X},$$

con $\mathbf{I}_n$ la matriz identidad de tamaño n.

La varianza de una variable $\mathbf{x}_j$ se puede conocer por

$$\sigma_{jj} = \sigma_j^2 = \frac{\sum_{i=1_p}^{N}(x_{ij} - \mu_j)^2}{N},$$

cuya raíz cuadrada se llama desviación estándar o típica $\sigma_j = \sqrt{\sigma_{jj}}$ y es de particular importancia como estadístico descriptivo ya que sus unidades de medida son iguales a las de la variable original. Por su parte, la covarianza entre dos variables $\mathbf{x}_j$ y $\mathbf{x}_l$ se puede calcular por medio de

$$\sigma_{jl} = \frac{\sum_{i=1}^{N}(x_{ij} - \mu_j)(x_{il} - \mu_l)}{N}.$$

El mejor estimador de la matriz de covarianza poblacional es la matriz de covarianza muestral $\mathbf{S}$,

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}_c'\mathbf{X}_c = \frac{1}{n-1}\mathbf{X}'\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\right)\mathbf{X} = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{12} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1p} & s_{2p} & \dots & s_{pp} \end{bmatrix},$$

donde $n-1$ son los grados de libertad de la muestra. La varianza muestral de una variable $\mathbf{x}_j$ es

$$s_{jj} = s_j^2 = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}{n-1},$$

y la covarianza muestral entre dos variables $\mathbf{x}_j$ y $\mathbf{x}_l$ es

$$s_{jl} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{il} - \bar{x}_l)}{n-1}.$$

Una cualidad especial de las matrices de covarianza es el hecho de que en ambos casos son matrices simétricas semidefinidas positivas. Esto se debe a que las varianzas σ_{jj}, s_{jj} ubicadas en sus diagonales son siempre mayores o iguales a cero, σ_{jj}, $s_{jj} \geq 0$, y las covarianzas entre las variables j y l serán iguales a las covarianzas entre l y j. Es decir, $\sigma_{jl} = \sigma_{lj}$, $s_{jl} = s_{lj}$ con $j \neq l$ donde $j, l = 1, 2, ..., p$.

Si es necesario definir la varianza del conjunto de datos en un solo número se puede hacer uso tanto de la varianza generalizada $VG = |\mathbf{\Sigma}|$ como de la varianza total $VT = tr(\mathbf{\Sigma}) = \sum_{j=1}^{p} \sigma_{jj}$. La raíz cuadrada de la varianza generalizada es conocida como desviación típica generalizada. Ambas varianzas son utilizadas en métodos de análisis de varianza multivariado, como componentes principales, análisis factorial, regresión múltiple, MANOVA, entre otros (Johnson et al., 2007; Diaz et al., 2012).

A partir de la matriz de covarianza es posible construir la matriz de correlación, para esto se define una matriz auxiliar $\mathbf{D} = diag(\mathbf{\Sigma})$ con $j = 1, 2, ...p$, la cual contiene en su diagonal las covarianzas de cada variable. De manera que la matriz de covarianza poblacional ρ se obtiene por

$$\boldsymbol{\rho} = Corr(\mathbf{x}_{(i)}) = \mathbf{D}^{-\frac{1}{2}}\mathbf{\Sigma}\mathbf{D}^{-\frac{1}{2}} = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & \rho_{22} & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & \rho_{pp} \end{bmatrix},$$

donde ρ_{jl} es el cociente entre la covarianza de las variables $\mathbf{x}_j$ y $\mathbf{x}_l$ y la división de las respectivas desviaciones estándar

$$\rho_{ij} = \frac{\sigma_{jl}}{\sqrt{\sigma_{jj}}\sqrt{\sigma_{ll}}}.$$

Para el caso muestral la matriz $\mathbf{D}$ se compondrá de los elementos de la diagonal de la matriz $\mathbf{S}$, en este caso la matriz de correlaciones $\mathbf{R}$ se debe calcular por

$$\mathbf{R} = \begin{bmatrix} s_{11}^{-1/2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & s_{pp}^{-1/2} \end{bmatrix} \begin{bmatrix} s_{11} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{1p} & \cdots & s_{pp} \end{bmatrix} \begin{bmatrix} s_{11}^{-1/2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & s_{pp}^{-1/2} \end{bmatrix},$$

es decir

$$\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2} = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1p} \\ r_{12} & r_{22} & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \dots & r_{pp} \end{bmatrix}.$$

Por razones similares a las presentadas para las matrices de covarianzas, las matrices de correlaciones son matrices simétricas semi definidas positivas con elementos $-1 \leq \rho_{jl}, r_{jl} \leq 1$ y $\rho_{jj}, r_{jj} = 1$ donde $j, l = 1, 2, ..., p$ para todo $j \neq l$. Un valor de correlación negativo indica una relación lineal inversa entre variables y un valor positivo indica una relación lineal directa entre ellas, en el caso de una correlación cercana a 1 se tiene una alta relación lineal y en el caso de una correlación cercana a 0 se dice que no existe relación entre las variables.

Un puto de vista más geométrico del coeficiente de correlación proviene de analizar la proyección de la variable $\mathbf{x}_1$ sobre una variable $\mathbf{x}_2$ donde los vectores contienen los N residuales $x_{ij} - \mu_j$ de las características 1 y 2 respectivamente. La proyección del vector $\mathbf{x}_1$ sobre el vector $\mathbf{x}_2$ es un nuevo vector $P_{\mathbf{x}_2}\mathbf{x}_1$ que tendrá la misma dirección del vector $\mathbf{x}_2$ pero magnitud $||\mathbf{x}_1|| \cos\theta$, de manera que

$$P_{\mathbf{x}_2}\mathbf{x}_1 = ||\mathbf{x}_1|| \cos\theta \frac{\mathbf{x}_2}{||\mathbf{x}_2||},$$

esta expresión se puede reescribir en términos del producto punto de la forma

$$P_{\mathbf{x}_2}\mathbf{x}_1 = \frac{\mathbf{x}_1 \cdot \mathbf{x}_2}{||\mathbf{x}_2||^2}\mathbf{x}_2,$$

la cual contiene toda la información sobre la correlación entre ambos vectores. Para ver esto, es necesario recordar que el producto punto entre los vectores $\mathbf{x}_1$ y $\mathbf{x}_2$ es

$$\mathbf{x}_1 \cdot \mathbf{x}_2 = \mathbf{x}_1'\mathbf{x}_2 = ||\mathbf{x}_1||\ ||\mathbf{x}_2|| \cos\theta,$$

en donde al despejar $\cos\theta$ se obtiene

$$\cos\theta = \frac{\mathbf{x}_1 \cdot \mathbf{x}_2}{||\mathbf{x}_1||\ ||\mathbf{x}_2||}. \tag{A.1}$$

Como el vector $\mathbf{x}_1$ y el vector $\mathbf{x}_2$ están conformados por los residuales de sus mediciones, son de la forma

$$\mathbf{x}_j' = \begin{bmatrix} x_{1j} - \mu_j & x_{2j} - \mu_j & \dots & x_{Nj} - \mu_j \end{bmatrix},$$

y su norma puede ser escrita como

$$||\mathbf{x}_j||^2 = \sum_{i=1}^{N}(x_{ij} - \mu_j)^2,$$

con producto punto

$$\mathbf{x}_1 \cdot \mathbf{x}_2 = \sum_{i=1}^{N}(x_{i1} - \mu_1)(x_{i2} - \mu_2).$$

Reemplazando estas expresiones en A.1 se obtiene

$$\cos\theta = \frac{\sum_{i=1}^{N}(x_{i1} - \mu_1)(x_{i2} - \mu_2)}{\sqrt{\sum_{i=1}^{N}(x_{i1} - \mu_1)^2}\sqrt{\sum_{i=1}^{N}(x_{i2} - \mu_2)^2}},$$

al dividir por N arriba y abajo es posible definir el coseno del ángulo de los vectores en términos de medidas estadísticas

$$\cos\theta = \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}},$$

de forma que $\rho_{12} = \cos\theta$.

De esta manera, el coeficiente de correlación indica el grado de dependencia lineal entre las variables o características medidas. Cuando el coeficiente de correlación es igual a 1 o -1 indica que los vectores forman un ángulo de $0°$ o $180°$ respectivamente, por lo tanto, son dos vectores paralelos con direcciones iguales u opuestas, es decir, son vectores linealmente dependientes. Si el coeficiente de correlación es igual a 0 el ángulo formado por las dos variables es de $90°$ lo que indica independencia lineal entre los vectores.

A.2. Inferencia Estadística Multivariada

Una de las herramientas más importantes en la inferencia estadística multivariada es la distancia de Mahalanobis la cual tiene interesantes propiedades gracias a su distribución como una chi-cuadrado. Para esto es necesario definir la distribución T^2 de Hotelling usando la distirbución Wishart y sus propiedades.

A.2.1. Distancia de Mahalanobis y su distribución

Para iniciar se partirá de la distribución normal univariada y se realizará la respectiva extensión al caso multivariado.

Si $x \sim N(\mu, \sigma^2)$ entonces la función de densidad está dada por

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{A.2}$$

$$= \frac{(\sigma^2)^{-1/2}}{\sqrt{2\pi}} \exp\left(\frac{1}{2}(x-\mu)(\sigma^2)^{-1}(x-\mu)\right), \tag{A.3}$$

Por teorema

$$z = \frac{x-\mu}{\sigma} \sim N(0,1),$$

y por tanto se puede escribir

$$x = \mu + z\sigma, \tag{A.4}$$

donde

$$E(x) = \mu,$$
$$Var(x) = \sigma^2.$$

Generalizando la distribución normal univariada dada en A.2 para $\mathbf{x}' = \begin{bmatrix} x_1 & x_2 & \cdots & x_p \end{bmatrix}'$, se tiene la función de densidad normal multivariada dada por

$$f(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{|\boldsymbol{\Sigma}|^{-1/2}}{(\sqrt{2\pi})^p} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right),$$

en donde:

$$E(\mathbf{x}) = \boldsymbol{\mu},$$
$$Cov(\mathbf{x}) = \boldsymbol{\Sigma}.$$

Tomando el vector $\mathbf{x}' = \begin{bmatrix} x_1 & x_2 & \cdots & x_p \end{bmatrix}'$, por A.4 cada una de las componentes se puede escribir como una combinación lineal de p variables normales estándar, a saber

$$x_j = \mu_j + a_{j1}z_1 + a_{j2}z_2 + ... + a_{jp}z_p. \tag{A.5}$$

De acuerdo con la ecuación anterior se pueden escribir las p variables de forma matricial de la forma

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} + \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_p \end{bmatrix}, \tag{A.6}$$

tomando $\mathbf{A}\mathbf{A}' = \Sigma$ y $\mathbf{z}' = \begin{bmatrix} z_1 & z_2 & \cdots & z_p \end{bmatrix}'$. De la forma matricial explicita dada en A.6 se obtiene la ecuación matricial general

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{A}\mathbf{z}, \tag{A.7}$$

y premultiplicando esta expresión por la matriz $\mathbf{A}^{-1}$ se encuentra una ecuación matricial para $\mathbf{z}$

$$\mathbf{A}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{A}^{-1}(\mathbf{A}\mathbf{z}) \tag{A.8}$$

$$\mathbf{z} = \mathbf{A}^{-1}(\mathbf{x} - \boldsymbol{\mu}). \tag{A.9}$$

Recordando que $\mathbf{A}\mathbf{A}' = \boldsymbol{\Sigma}$ y así, $\boldsymbol{\Sigma}^{-1} = (\mathbf{A}\mathbf{A}')^{-1} = (\mathbf{A}')^{-1}\mathbf{A}^{-1}$.

Ahora, realizando una transformación de variables con el objetivo de encontrar una expresión para la distribución conjunta de $\mathbf{x}$ se tiene que

$$\mathbf{z} = \mathbf{A}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \rightarrow \left| \frac{\partial \mathbf{z}}{\partial \mathbf{x}} \right| = |\mathbf{A}^{-1}|,$$

con $z_j \sim N(0,1)$ e independientes. Por lo anterior es posible escribir

$$\begin{aligned} f_x(x_1, \cdots, x_p) &= f_z(z_1(\mathbf{x}), \cdots, z_p(\mathbf{x})) \left| \frac{\partial \mathbf{z}}{\partial \mathbf{x}} \right|, \\ &= f_z(z_1(\mathbf{x})) \cdots f_z(z_p(\mathbf{x})) \left| \frac{\partial \mathbf{z}}{\partial \mathbf{x}} \right|, \\ &= \frac{1}{\sqrt{2\pi}} e^{-z_1^2/2} \cdots \frac{1}{\sqrt{2\pi})} e^{-z_p^2/2} |\mathbf{A}^{-1}|, \\ &= \frac{1}{(\sqrt{2\pi})^p} \exp\left(-\frac{1}{2} \sum_{j=1}^{p} z_j^2 \right) |\mathbf{A}^{-1}|, \end{aligned}$$

donde el interés se centra en la sumatoria a la que está elevada la función es exponencial, de esta manera $(\sum_{j=1}^{p} z_j^2) \sim \chi_p^2$ claramente.

Por otro lado se sabe que

$$\sum_{j=1}^{p} z_j^2 = z_1^2 + z_2^2 + \cdots + z_p^2,$$

$$\mathbf{z}'\mathbf{z} = \begin{bmatrix} z_1 & z_2 & \cdots & z_p \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_p \end{bmatrix} = \sum_{j=1}^{p} z_j^2,$$

pero, $\mathbf{z} = \mathbf{A}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ (ver A.9) por lo que $\mathbf{z}'\mathbf{z}$ es dado por

$$\begin{aligned} \mathbf{z}'\mathbf{z} &= [\mathbf{A}^{-1}(\mathbf{x} - \boldsymbol{\mu})]'[\mathbf{A}^{-1}(\mathbf{x} - \boldsymbol{\mu})], \\ &= (\mathbf{x} - \boldsymbol{\mu})'(\mathbf{A}^{-1})'(\mathbf{A}^{-1})(\mathbf{x} - \boldsymbol{\mu}), \\ \mathbf{z}'\mathbf{z} &= (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}). \end{aligned}$$

De esta manera se concluye que la distancia de Mahalanobis tiene una distribución chi-cuadrado con p grados de libertad, a saber

$$\sum_{j=1}^{p} z_j^2 = \mathbf{z}'\mathbf{z} = \overbrace{(\mathbf{x} - \boldsymbol{\mu})'}^{a'} \boldsymbol{\Sigma}^{-1} \overbrace{(\mathbf{x} - \boldsymbol{\mu})}^{a} \sim \chi^2, z_j \sim N(0,1), z_j^2 \sim \chi^2.$$

Finalmente,

$$\begin{aligned} E(\mathbf{x}) &= \boldsymbol{\mu}, \\ Var(\mathbf{x}) &= Var(\boldsymbol{\mu}) + Var(\mathbf{A}\mathbf{z}), \\ &= \mathbf{0} + \mathbf{A}\ \mathrm{var}(\mathbf{z})\mathbf{A}' \\ &= \mathbf{A}\mathbf{A}', \end{aligned}$$

también por definición de varianza

$$\begin{aligned} Var(\mathbf{x}) &= E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'] \\ &= E[(\mathbf{A}\mathbf{z})(\mathbf{A}\mathbf{z})'] \\ &= E[(\mathbf{A}\mathbf{z})(\mathbf{z}'\mathbf{A}')] \\ &= \mathbf{A}E[\mathbf{z}\mathbf{z}']\mathbf{A}' \\ &= \mathbf{A}\mathbf{I}_p\mathbf{A}' \\ &= \mathbf{A}\mathbf{A}' = \boldsymbol{\Sigma}, \end{aligned}$$

lo cual es consistente con la definición de la matriz $\boldsymbol{\Sigma}$.

A.2.2. Demostración por descomposición espectral

Sea $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ con $|\boldsymbol{\Sigma}| > 0$ entonces

$$(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \sim \chi^2_p.$$

En el transcurrir de la demostración se hará uso de los siguientes teoremas bastante conocidos:

Si $x \sim N(\mu, \sigma^2)$ entonces se cumple que

$$z = \frac{x - \mu}{\sigma} \sim N(0, 1),$$
$$z^2 = \left(\frac{x - \mu}{\sigma} \right)^2 \sim \chi^2,$$
$$\sum_{j=1}^{p} z_j^2 = z_1^2 + z_2^2 + \cdots + z_p^2 \sim \chi^2_p.$$

Descomposición espectral

Siendo $\boldsymbol{\Sigma}$ una matriz simétrica de tamaño $p \times p$ según el teorema de la descomposición espectral esta matriz puede ser escrita en función de sus valores propios (λ_j) y vectores propios ($\mathbf{v}_j$) de la siguiente forma

$$\boldsymbol{\Sigma} = \lambda_1 \mathbf{v}_1 \mathbf{v}_1' + \lambda_2 \mathbf{v}_2 \mathbf{v}_2' + \cdots + \lambda_p \mathbf{v}_p \mathbf{v}_p',$$
$$\boldsymbol{\Sigma} = \sum_{j=1}^{p} \lambda_j \mathbf{v}_j \mathbf{v}_j' = \mathbf{V} \Lambda \mathbf{V}',$$

con $\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p \end{bmatrix}$ y $\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_p \end{bmatrix}$ la matriz diagonal de los valores propios y la matriz de los vectores propios, respectivamente.

Ahora, a través de la función característica se tiene el siguiente resultado (Johnson & Wichern, 2007):

Resultado

$$\boldsymbol{\Sigma} \mathbf{v} = \lambda \mathbf{v} \rightarrow \boldsymbol{\Sigma}^{-1} \mathbf{v} = \frac{1}{\lambda} \mathbf{v} = \lambda^{-1} \mathbf{v},$$

con $\mathbf{v}\mathbf{v}' = 1$.

Nota: Para $\mathbf{\Sigma}$ se tiene $(\lambda, \mathbf{v})$.

Prueba

Como $\mathbf{\Sigma}$ semidefinida positiva, $\mathbf{v} \neq \mathbf{0}$ y $0 < \mathbf{v'\Sigma v}$

$$\begin{aligned}\mathbf{v'\Sigma v} &= \mathbf{v'}(\mathbf{\Sigma v}),\\ &= \mathbf{v'}(\lambda \mathbf{v}),\ \text{por hipótesis}\\ &= \lambda \mathbf{v'v} = \lambda,\end{aligned}$$

y así, $\mathbf{v'\Sigma v} = \lambda$. Ahora es posible escribir $\mathbf{v}$ como

$$\mathbf{v} = \mathbf{\Sigma}^{-1}\mathbf{\Sigma v} = \mathbf{\Sigma}^{-1}(\lambda \mathbf{v}) = \lambda \mathbf{\Sigma}^{-1}\mathbf{v}, \tag{A.10}$$

de donde se tiene que $\mathbf{v} = \lambda \mathbf{\Sigma}^{-1}\mathbf{v}$. Se puede dividir por λ a los dos lados de la igualdad dada en A.10 y se obtiene

$$\frac{1}{\lambda}\mathbf{v} = \mathbf{\Sigma}^{-1}\mathbf{v}, \tag{A.11}$$

es decir, se le asocia el par valor-vector propio $(\frac{1}{\lambda}, \mathbf{v})$ a la matriz $\mathbf{\Sigma}^{-1}$.

Con los resultados obtenidos en A.10 y A.11 y además teniendo en cuenta que $z_j = \dfrac{\mathbf{x}_j - \mu_j}{\sigma_{jj}} \sim N(0,1)$ y $z_j^2 \sim \chi_1^2$, se tiene

$$\sum_{j=1}^{p} z_j^2 = z_2^1 + z_2^2 + \cdots z_p^2 \sim \chi_p^2,$$

$$\mathbf{\Sigma} = \lambda_1 \mathbf{v}_1 \mathbf{v}_1' + \cdots + \lambda_p \mathbf{v}_p \mathbf{v}_p' = \sum_{j=1}^{p} \lambda_j \mathbf{v}_j \mathbf{v}_j',$$

$$\mathbf{\Sigma}^{-1} = \frac{1}{\lambda_1}\mathbf{v}_1 \mathbf{v}_1' + \cdots + \frac{1}{\lambda_p}\mathbf{v}_p \mathbf{v}_p' = \sum_{j=1}^{p} \frac{1}{\lambda_j}\mathbf{v}_j \mathbf{v}_j'.$$

En consecuencia, sustituyendo la matriz $\mathbf{\Sigma}^{-1}$ por su equivalente y realizando

algunas operaciones, se obtiene

$$\begin{aligned}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) &= (\mathbf{x}-\boldsymbol{\mu})'\sum_{i=1}^{p}\frac{1}{\lambda_j}\mathbf{v}_j\mathbf{v}_j'(\mathbf{x}-\boldsymbol{\mu}),\\ &= \sum_{j=1}^{p}\frac{1}{\lambda_i}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{v}_j\mathbf{v}_j'(\mathbf{x}-\boldsymbol{\mu}),\\ &= \sum_{j=1}^{p}\frac{1}{\lambda_j}[\mathbf{v}_j'(\mathbf{x}-\boldsymbol{\mu})]^2,\\ &= \sum_{j=1}^{p}\left[\frac{1}{\sqrt{\lambda_i}}\mathbf{v}_j'(\mathbf{x}-\boldsymbol{\mu})\right]^2,\\ &= \sum_{j=1}^{p}z_j^2 \sim \chi_p^2.\end{aligned}$$

Resumiendo,

$$\begin{aligned}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) &= \sum_{j=1}^{p}\left[\frac{1}{\sqrt{\lambda_j}}\mathbf{v}_j'(\mathbf{x}-\boldsymbol{\mu})\right]^2,\\ &= \sum_{j=1}^{p}[\mathbf{A}(\mathbf{x}-\boldsymbol{\mu})]^2.\end{aligned}$$

Demostrando que $\mathbf{A}(\mathbf{x}-\boldsymbol{\mu})$ tiene una distribución normal multivariada estándar se tiene el resultado anterior y se completa la demostración. En este sentido, explícitamente el vector $\mathbf{z}$ y la matriz $\mathbf{A}$ son dados por

$$\mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_p \end{pmatrix} \mathbf{A} = \begin{pmatrix} \frac{1}{\sqrt{\lambda_1}}\mathbf{v}_1' \\ \frac{1}{\sqrt{\lambda_2}}\mathbf{v}_2' \\ \vdots \\ \frac{1}{\sqrt{\lambda_p}}\mathbf{v}_p' \end{pmatrix}.$$

También a partir de la distribución normal multivariada se conocen los siguientes dos resultados

$$\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \rightarrow (\mathbf{x}-\boldsymbol{\mu}) \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}), \tag{A.12}$$

$$\mathbf{z} = \mathbf{A}(\mathbf{x}-\boldsymbol{\mu}) \sim N(\mathbf{0}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'), \tag{A.13}$$

además, la varianza de $\mathbf{z}$ es dada por

$$Var(\mathbf{z}) = \mathbf{A}(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{A},$$

$$Var(\mathbf{x} - \boldsymbol{\mu})\mathbf{A}' = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}',$$

que es igual a la matriz identidad, como se verá a continuación. Con ese objetivo, en $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$ se sustituye cada matriz por su equivalente

$$\begin{aligned}
\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' &= \begin{bmatrix} \frac{1}{\sqrt{\lambda_1}}\mathbf{v}_1' \\ \frac{1}{\sqrt{\lambda_2}}\mathbf{v}_2' \\ \vdots \\ \frac{1}{\sqrt{\lambda_p}}\mathbf{v}_p', \end{bmatrix} \left(\sum_{j=1}^{p} \lambda_j \mathbf{v}_j' \mathbf{v}_j \right) \begin{bmatrix} \frac{1}{\sqrt{\lambda_1}}\mathbf{v}_1' & \frac{1}{\sqrt{\lambda_2}}\mathbf{v}_2' & \cdots & \frac{1}{\sqrt{\lambda_p}}\mathbf{v}_p' \end{bmatrix}, \\
&= \left(\sum_{j=1}^{p} \lambda_j \right) \begin{bmatrix} \frac{1}{\sqrt{\lambda_1}}\mathbf{v}_1' \\ \frac{1}{\sqrt{\lambda_2}}\mathbf{v}_2' \\ \vdots \\ \frac{1}{\sqrt{\lambda_p}}\mathbf{v}_p', \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{\lambda_1}}\mathbf{v}_1' & \frac{1}{\sqrt{\lambda_2}}\mathbf{v}_2' & \cdots & \frac{1}{\sqrt{\lambda_p}}\mathbf{v}_p' \end{bmatrix}, \\
&= \mathbf{I}_p,
\end{aligned}$$

y por tanto teniendo en cuenta A.13 se tiene que $\mathbf{z} = \mathbf{A}(\mathbf{x} - \boldsymbol{\mu}) \sim N_p(\mathbf{0}, \mathbf{I}_p)$, que era lo que se quería probar.

Intervalo de confianza

Realizando la analogía del caso univariado, y teniendo en cuenta la demostración anterior, simplemente se reemplaza la distancia de Mahalanobis en el lugar del estadístico que tiene una distribución chi - cuadrado, como se ilustra a continuación.

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1},$$

$$P(\chi^2 \leq \chi^2_\alpha) = 1 - \alpha,$$

$$P\left(\frac{(n-1)s^2}{\sigma^2} \leq \chi^2_\alpha \right) = 1 - \alpha,$$

$$P\left((\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \leq \chi^2_\alpha \right) = 1 - \alpha.$$

A.2.3. Distribución T^2 de Hotelling

La distribución T^2 de Hotelling es una distribución utilizada comúnmente para la inferencia estadística multivariada gracias a que permite el uso de la matriz de covarianzas estimada cumpliendo un papel similar al de la distribución t de Student en el caso univariado. De manera que se abordará primero la distribución t de Student de forma específica y luego se generalizará para el caso multivariado obteniendo así la distribución T^2 de Hotelling.

Distribución t de student

Un resultado muy conocido es el hecho de que el estadístico que envuelve la desviación estándar muestral tiende a una distribución t de student con $n-1$ grados de libertad, específicamente,

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}.$$

De manera que cuando se necesita probar las hipótesis

$$\begin{cases} H_0 = \mu = \mu_0, \\ H_1 = \mu \neq \mu_0, \end{cases}$$

bajo H_0 se tiene que

$$t_0 = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}},$$

Donde se rechaza H_0 cuando $t_0 > t_{\alpha/2}$ o $t_o < t_{\alpha/2}$ lo que es equivalente a $|t_0| > t_{\alpha/2}$. Rechazar H_0 cuando $|t_0|$ es grande es equivalente a rechazar H_0 si el cuadrado t_0^2 es grande, donde t_0^2 está dado por

$$t_0^2 = \left(\frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \right)^2 = n(\bar{x} - \mu_0)(s^2)^{-1}(\bar{x} - \mu_0).$$

De manera que el intervalo de confianza para μ_0 es dado por

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}.$$

Vale resaltar que el intervalo de confianza consiste de todos aquellos valores μ_0 que no se rechazan por la prueba con hipótesis nula $H_0 : \mu = \mu_0$.

Caso Multivariado

Para el caso multivariado se considera el problema de determinar si $\boldsymbol{\mu}_0$ es un vector pausible para el vector de medias de una distribución normal multivariada. Entonces, se define la distribución T^2 de Hotelling por medio del estadístico

$$T^2 = (\bar{\mathbf{x}} - \boldsymbol{\mu}_0)' \left(\frac{\mathbf{S}}{n}\right)^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}_0),$$

$$T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)' \mathbf{S}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}_0),$$

con matriz de covarianza muestral

$$\mathbf{S} = \frac{1}{n-1} \sum_{j=1}^{p} (x_j - \bar{X})(x_j - \bar{X})'.$$

Así pues, el estadístico T^2 se distribuye como una Fisher de $p/(n-p)$ grados de libertad. De manera que

$$T^2 \sim \frac{(n-1)p}{n-p} F_{p,n-p},$$

donde

$$F = \frac{\frac{\chi^2_p}{p}}{\frac{\chi^2_{n-p}}{n-p}}.$$

Distribución T^2 de Hotelling y Whishart

Antes de demostrar a donde tiende la distribución T^2 de Hotelling, se definirá la distribución de Whishart.

Se tiene $\mathbf{X}$ de tamaño $m \times p$ con distribución normal multivariada con $\boldsymbol{\mu} = \mathbf{0}$ y matriz de covarianza $\boldsymbol{\Sigma}$, a saber, $\mathbf{X} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$.

Se define la matriz $\mathbf{M}$ con distribución de Wishart, específicamente,

$$\mathbf{M} = \mathbf{X}'\mathbf{X} \sim Wishart(\Sigma, m). \tag{A.14}$$

Considerando solamente un vector con distribución normal univariada de media 0 y varianza σ^2, específicamente

$$\begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{bmatrix} \sim N(0, \sigma^2),$$

fácilmente se ve que

$$\mathbf{M} = \mathbf{x}'\mathbf{x} \sim Wishart(\sigma^2, m). \tag{A.15}$$

Por otro lado, teniendo en cuenta que $\mathbf{x}$ tiene media 0,

$$\begin{aligned} z &= \frac{\bar{x} - \mu}{\sigma} \sim N(0,1), \\ z &= \frac{\bar{x}}{\sigma} \sim N(0,1), \\ z^2 &= \frac{\mathbf{x}'\mathbf{x}}{\sigma^2} \sim \chi^2_m, \\ \mathbf{M} &= \mathbf{x}'\mathbf{x} \sim \sigma^2 \chi^2_m. \end{aligned}$$

Luego considerando A.15, $Wishart(\sigma^2, m) \sim \sigma^2 \chi^2_m$; con lo que se concluye que

$$Wishart(1, m) \sim \chi^2_m. \tag{A.16}$$

De A.14, para el caso multivariado

$$\mathbf{M} \sim Wishart(\mathbf{\Sigma}, m),$$

y por propiedades de la distribución Wishart

$$\mathbf{a}'\mathbf{M}\mathbf{a} \sim Wishart(\mathbf{a}'\mathbf{\Sigma}\mathbf{a}, m), \tag{A.17}$$

y por tanto

$$\frac{\mathbf{a}'\mathbf{M}\mathbf{a}}{\mathbf{a}\mathbf{\Sigma}\mathbf{a}} \sim Wishart(1, m) \sim \chi^2_m. \tag{A.18}$$

Ahora con el objetivo de encontrar la distribución de T^2 de Hotelling, específicamente de $T^2 = m(\mathbf{x} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ se toma $\mathbf{M} = \mathbf{S}$ y $\mathbf{a} = \mathbf{x} - \boldsymbol{\mu}$ se multiplica y se divide al lado derecho de la igualdad anterior por $(\mathbf{x} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ y se encuentra que T^2 queda como el cociente de dos distribuciones chi-cuadrado que por definición al dividirse por sus respectivos grados de libertad se obtiene la distribución Fisher. Esto es

$$T^2 = m(\mathbf{x}-\boldsymbol{\mu})'\mathbf{S}^{-1}(\mathbf{x}-\boldsymbol{\mu}),$$

$$T^2 = \frac{m(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{M}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}),$$

teniendo en cuenta que,

$$F = \frac{\frac{\chi^2_u}{u}}{\frac{\chi^2_v}{v}} \sim F_{u,v},$$

$$T^2 = \frac{m(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{\frac{(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{(\mathbf{x}-\boldsymbol{\mu})'\mathbf{M}^{-1}(\mathbf{x}-\boldsymbol{\mu})}},$$

para el denominador tener en cuenta A.18 y organizando los grados de libertad, se obtiene la expresión para T^2,

$$T^2 = \frac{\frac{mp(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{p}}{\frac{(m-p+1)\frac{(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{(\mathbf{x}-\boldsymbol{\mu})'\mathbf{M}^{-1}(\mathbf{x}-\boldsymbol{\mu})}}{m-p+1}},$$

$$T^2 = \frac{mp}{m-p+1}\frac{\frac{\chi^2_p}{p}}{\frac{\chi^2_{m-p+1}}{m-p+1}} \sim \frac{mp}{m-p+1}F_{p,m-p}.$$

Tomando $m = n-1$ se obtiene la expresión deseada para la distribución de Hotelling, a saber

$$T^2 = m(\mathbf{x}-\boldsymbol{\mu})'\mathbf{S}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \sim \frac{(n-1)p}{n-p}F_{p,n-p}.$$

A.2.4. Una aproximación geométrica a la distancia de Mahalanobis

La distancia de Mahalanobis entre dos individuos $\mathbf{x}_{(i)}$ y $\mathbf{x}_{(k)}$ seleccionados aleatoriamente de una población con matriz de covarianza $\boldsymbol{\Sigma}$ está definida por la forma cuadrática

$$d^2_{ik} = (\mathbf{x}_{(k)}-\mathbf{x}_{(i)})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_{(k)}-\mathbf{x}_{(i)}).$$

Sus propiedades como medida de distancia se desprenden del hecho de que la matriz $\mathbf{\Sigma}^{-1}$ es simétrica definida positiva, es decir, cumple con la forma cuadrática $\mathbf{x}'\mathbf{A}\mathbf{x} > 0$, donde $\mathbf{A}$ es una matriz cuadrada definida positiva de tamaño $p \times p$. Para una mayor descipción de esta propiedad consultar Johnson et al. (2007) e Ipsen (2009).

La distancia de Mahalanobis puede ser usada para conocer que tan alejado se encuentra un individuo $\mathbf{x}_{(i)}$ del centro de masa μ. De modo que

$$d_i^2 = (\mathbf{x}_{(i)} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}_{(i)} - \boldsymbol{\mu}),$$

mide la distancia de cada observación $\mathbf{x}_{(i)}$ a la media $\boldsymbol{\mu}$ corregida por la variabilidad de la muestra.

Una propiedad interesante de la distancia Mahalanobis se presenta cuando se trabaja bajo el supuesto de multinormalidad en los datos. Si $\mathbf{x}_{(i)}$ se distribuye como $\mathbf{x}_{(i)} \sim N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, entonces d_i^2 se distribuye como una ji-cuadrado con p grados de libertad, o sea, $d_i^2 \sim \chi_p^2$.

Para una aproximación sencilla, se supone que las variables $\mathbf{x}_j$ no se encuentran correlacionadas de forma que $\sigma_{jl} = 0$ para todo $j, l = 1, 2, \ldots, p$ con $j \neq l$. Reescribiendo la distancia de Mahalanobis se tiene

$$d_i^2 = (\mathbf{x}_{(i)} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}_{(i)} - \boldsymbol{\mu}) = \sum_{j=1}^{p} \frac{(x_{ij} - \mu_j)^2}{\sigma_{jj}} = \sum_{j=1}^{p} z_{ij}^2,$$

bajo el supuesto de normalidad $z_{ij} \sim N(0, 1)$ de forma que $z_{ij}^2 \sim \chi^2$ y, por lo tanto, $d_i^2 \sim \chi_p^2$ al ser una sumatoria de variables normales estándar al cuadrado. Así que

$$(\mathbf{x}_{(i)} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}_{(i)} - \boldsymbol{\mu}) \sim \chi_p^2. \tag{A.19}$$

Bajo el supuesto de normalidad, al igualar la distancia de Mahalanobis a un valor de $\chi_p^2(\alpha)$ estamos generando un contorno de densidad de probabilidad que describe un hiper elipsoide que contiene el $(1 - \alpha)\,\%$ de los datos, obteniendo así una herramienta útil en la identificación de datos atípicos.

Para ilustrar esto, se supone un vector aleatorio de dos dimensiones $\mathbf{x}_{(i)} = \begin{bmatrix} x_{i1} & x_{i2} \end{bmatrix}$ con vector de medias $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 & \mu_2 \end{bmatrix}$ y matriz de covarianzas $\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$, para hacer uso de la distancia de Mahalanobis es necesario que

la varianza generalizada sea diferente de cero para evitar la singularidad en la matriz de covarianzas, $VG = \sigma_{11}\sigma_{22} - \sigma_{12}^2 \neq 0$. De forma que

$$d_i^2 = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{bmatrix} x_{i1} - \mu_1 & x_{i2} - \mu_2 \end{bmatrix} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix} \begin{bmatrix} x_{i1} - \mu_1 \\ x_{i2} - \mu_2 \end{bmatrix},$$

$$d_i^2 = \frac{(x_{i1} - \mu_1)^2\sigma_{22} - 2(x_{i1} - \mu_1)(x_{i2} - \mu_2)\sigma_{12} + (x_{i2} - \mu_2)^2\sigma_{11}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}.$$

Si se fija la distancia d_i^2 en un valor $\chi^2(\alpha)$ se llega a que la expresión anterior define el contorno de una elipse rotada un ángulo θ respecto al eje $\mathbf{x}_1$, es de esperar que esta elipse contenga al $(1 - \alpha)\,\%$ de los datos.

Los ejes de la elipsoide indican las direcciones de máxima variabilidad de los datos. Como se muestra en el capítulo 1 la longitud de cada uno de los ejes del hiperelipsoide está dada por $\sqrt{\lambda_j^2\chi_p^2(\alpha)}$ donde λ_j es el valor propio relacionado con la j-ésima componente principal. Sin embargo, si las variables no se encuentran correlacionadas la longitud de los ejes del hiperelipsoide estarán dadas por $\sqrt{\sigma_{jj}\chi_p^2(\alpha)}$.

Si para el caso de dos dimensiones se tiene $\rho_{12} = 0$ de forma que $\sigma_{12} = 0$, al reemplazar en la expresión anterior y fijar la distancia a $\chi_p^2(\alpha)$ se obtiene

$$\frac{(x_{i1} - \mu_1)^2}{\sigma_{11}} + \frac{(x_{i2} - \mu_2)^2}{\sigma_{22}} = \chi_p^2(\alpha),$$

llevando la expresión a su forma canónica

$$\frac{(x_{i1} - \mu_1)^2}{\sigma_{11}\chi_p^2(\alpha)} + \frac{(x_{i2} - \mu_2)^2}{\sigma_{22}\chi_p^2(\alpha)} = 1,$$

donde la distancia desde el centroide al contorno sobre el semieje $\mathbf{x}_1$ es $\sqrt{\sigma_{11}\chi_p^2(\alpha)}$ y al semieje $\mathbf{x}_2$ es $\sqrt{\sigma_{22}\chi_p^2(\alpha)}$.

A.3. Cartas de control

Una de las herramientas estadísticas comúnmente utilizadas en la industria son las cartas de control cuyo fin consiste en diagnosticar causas especiales de variabilidad presente en los procesos productivos. Sin embargo, también pueden funcionar como criterio en la identificación de datos atípicos en una muestra. Las dos cartas multivariadas más comunes son la

carta en formato elipse y la carta T^2, estas cartas son utilizadas para monitorear la estabilidad de una muestra de observaciones multivariadas y así definir regiones de control para observaciones futuras. Para la construcción de las cartas de control multivariadas inicialmente se considerará los procedimientos para una muestra de observaciones individuales y posteriormente para grupos de medias muestrales, con base en lo presentado en Jhonson et al. (2007) y Montgomery (2012).

A.3.1. Cartas para el análisis de estabilidad de una muestra de observaciones

Se asume una muestra aleatoria de gran tamaño $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(n)}$ con distribución $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ gracias al teorema del límite central. El interés es identificar aquellos datos que presentan una variabilidad inusual para el proceso o que son atípicos para la muestra, es decir, que se encuentran fuera de la región de control. Para esto, es necesario analizar los residuales o la distancia de cada una de las observaciones $\mathbf{x}_{(i)}$ al centroide de la muestra $\bar{\mathbf{x}}$. De manera que

$$\mathbf{x}_{(i)} - \bar{\mathbf{x}} = \left(1 - \frac{1}{n}\right)\mathbf{x}_{(i)} - \frac{1}{n}\mathbf{x}_{(1)} - \cdots - \frac{1}{n}\mathbf{x}_{(i-1)} - \frac{1}{n}\mathbf{x}_{(i+1)} - \cdots - \frac{1}{n}\mathbf{x}_{(n)},$$

en donde el valor esperado es

$$E(\mathbf{x}_{(i)} - \bar{\mathbf{x}}) = E(\mathbf{x}_{(i)}) - E(\bar{\mathbf{x}}) = \boldsymbol{\mu} - \boldsymbol{\mu} = \mathbf{0},$$

y la covarianza

$$Cov(\mathbf{x}_{(i)} - \bar{\mathbf{x}}) = \left(1 - \frac{1}{n}\right)^2 \boldsymbol{\Sigma} + (n-1)n^{-2}\boldsymbol{\Sigma} = \frac{(n-1)}{n}\boldsymbol{\Sigma},$$

En donde cada residual $\mathbf{x}_{(i)} - \bar{\mathbf{x}}$ tiene una distribución normal $N_p\left(\mathbf{0}, \frac{(n-1)}{n}\boldsymbol{\Sigma}\right)$. Sin embargo, para fijar los límites de control es posible utilizar el hecho de que la distancia de Mahalanobis estimada se distribuye bajo una chi cuadrada de p grados de libertad cuando las muestras son de gran tamaño gracias al teorema del límite central, obteniendo

$$(\mathbf{x}_{(i)} - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_{(i)} - \bar{\mathbf{x}}) \sim \chi^2_p.$$

A partir de este hecho es posible construir tanto la carta de formato elipse que está pensada para dos variables como la carta de formato T^2 que está pensada para más de dos variables.

Cartas de formato elipse

La carta de formato elipse es una carta de control bivariada. Para obtener la carta es necesario ubicar las dos características de la i-ésima unidad (x_{i1}, x_{i2}) en un gráfico de dispersión con los límites definidos por la región de control del 95 % de confianza que consiste de todos los $\mathbf{x}_{(i)}$ que satisfacen

$$(\mathbf{x}_{(i)} - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_{(i)} - \bar{\mathbf{x}}) \leq \chi^2_2(0{,}05).$$

Cuando un punto se encuentra fuera de la región de control, es necesario analizar el comportamiento de cada dimensión independientemente. De forma que se construyen dos cartas $\bar{x}$ cuyos límites de están dados por los siguientes límites de control superior (LCS) e inferior (LCI):

$$LCS = \bar{x}_j + 3\sqrt{s_{jj}},$$

$$LCI = \bar{x}_j - 3\sqrt{s_{jj}}.$$

Donde el valor central del intervalo corresponde a $\bar{x}_j$. Si los datos son no negativos el límite inferior se define en cero en caso tal que sea menor que cero.

Cartas de control $\mathbf{T}^2$

Una carta T^2 por su parte puede ser aplicada a un gran número de características a diferencia del formato elipsoidal. Además, los datos son presentados en orden temporal, lo que permite identificar tendencias y patrones en los datos.

Para el i-ésimo punto se calcula el estadístico T^2

$$T_i^2 = (\mathbf{x}_{(i)} - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_{(i)} - \bar{\mathbf{x}}),$$

una vez obtenidos los valores T_i^2 se grafican los individuos en un eje temporal donde el límite de control inferior es cero y el límite de control superior está definido por un estadístico chi cuadrado

$$LCS = \chi^2_p(0{,}05),$$

$$LCI = 0,$$

para una carta de control del 95 % de confiabilidad. De forma que la carta T^2 no tiene un valor o línea central.

Cuando la carta muestra que la i-ésima unidad se encuentra fuera de control, es necesario determinar cuales variables son responsables. Una región modificada con base en los intervalos de Bonferroni es comúnmente utilizada para este propósito. La j-ésima variable está fuera de control si x_{ij} no pertenece al intervalo

$$(\bar{x}_j - t_{n-1}(0{,}005/p)\sqrt{s_{jj}},\ \bar{x}_j + t_{n-1}(0{,}005/p)\sqrt{s_{jj}}),$$

donde p es el total de variables medidas.

A.3.2. Regiones de control para observaciones futuras individuales

La meta ahora es usar los datos $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(n)}$, recolectados cuando el proceso es estable para fijar una región de control, llamada región de predicción, para una observación futura $\mathbf{x}$. El objetivo de este tipo de cartas consiste en identificar posibles defectos de calidad en un proceso que ya se encuentra bajo control o comportamientos atípicos en nuevas observaciones pertenecientes a una población cuyo comportamiento se ha descrito con anterioridad. De manera que el proceso debe ser estable antes de que los datos puedan ser usados para determinar regiones de control para las nuevas observaciones.

Sean $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(n)}$ distribuidas independientemente como $N_p(\mu, \Sigma)$ y sea $\mathbf{x}$ una futura observación de la misma distribución. Entonces el estadístico T^2 tiene la forma

$$T^2 = \frac{n}{n+1}(\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}),$$

y se distribuye como

$$T^2 \sim \frac{(n-1)p}{n-p}F_{p,n-p}.$$

Y una elipsoide de predicción p dimensional del $100(1-\alpha)\,\%$ está dada por todos los $\mathbf{x}$ que satisfacen

$$(\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) \leq \frac{(n^2-1)p}{n(n-p)}F_{p,n-p}(\alpha).$$

Para esto, primero se debe resaltar que $\mathbf{x} - \bar{\mathbf{x}}$ tiene media $\mathbf{0}$, ya que

$$E(\mathbf{x} - \bar{\mathbf{x}}) = E(\mathbf{x}) - E(\bar{\mathbf{x}}) = \boldsymbol{\mu} - \boldsymbol{\mu} = \mathbf{0},$$

y como $\mathbf{x}$ es una observación futura $\mathbf{x}$ y $\bar{\mathbf{x}}$ son independientes, entonces

$$Cov(\mathbf{x} - \bar{\mathbf{x}}) = Cov(\mathbf{x}) + Cov(\bar{\mathbf{x}}) = \mathbf{\Sigma} + \frac{1}{n}\mathbf{\Sigma},$$

$$Cov(\mathbf{x} - \bar{\mathbf{x}}) = \frac{(n+1)}{n}\mathbf{\Sigma},$$

de forma que

$$\sqrt{\frac{n}{n+1}}(\mathbf{x} - \bar{\mathbf{x}}) \sim N_p(\mathbf{0}, \mathbf{\Sigma}).$$

Ahora,

$$\sqrt{\frac{n}{n+1}}(\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}\sqrt{\frac{n}{n+1}}(\mathbf{x} - \bar{\mathbf{x}}),$$

lo cual combina un vector aleatorio normal multivariado $N_p(\mathbf{0}, \mathbf{\Sigma})$ y una matriz aleatoria Wishart $W_{p,n-1}$ y por lo tanto se distribuye como una Fitcher modificada por sus grados de libertad.

Cartas de control elipsoidales para futuras observaciones

Con $p = 2$, el 95 % la elipse de predicción se puede definir como

$$(\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) \leq \frac{2(n^2-1)}{n(n-2)}F_{2,n-2}(0,05),$$

de forma que cualquier observación futura $\mathbf{x}$ se declara fuera de control si se ubica fuera de la elipse.

Cartas de control T^2 para futuras observaciones

Para cada observación $\mathbf{x}$ se debe ubicar el punto

$$T^2 = \frac{n}{n+1}(\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}),$$

en orden temporal definiendo el límite de control inferior como cero y el límite de control superior como

$$LCS = \frac{(n-1)p}{(n-p)}F_{p,n-p}(0{,}05).$$

Los puntos por encima del límite de control superior representan causas potenciales de variación y sugieren que el proceso debería ser examinado para determinar si es necesario tomar una decisión correctiva o que un individuo presenta características altamente atípicas respecto a la población a la que pertenece.

A.3.3. Cartas de control basadas en medias submuestrales

Cuando la obtención de las mediciones se realiza por un muestreo aleatorio de $m > 1$ unidades seleccionadas al tiempo es necesario trabajar sobre las medias de esas muestras. Es decir, la muestra i determinará la media muestral $\bar{\mathbf{x}}_{(i)}$ y la matriz de covarianza muestral $\mathbf{S}_i$ donde $i = 1, 2, ..., n$. Por lo general, se asume que cada vector de observaciones aleatorio se distribuye independientemente como $N_p(\mathbf{0}, \mathbf{\Sigma})$, supuesto bajo el cual tanto las medias muestrales como las matrices de covarianza son independientes.

Para un submuestreo general de medias $\bar{\mathbf{x}}_{(i)}$, el residual $\bar{\mathbf{x}}_{(i)} - \bar{\bar{\mathbf{x}}}$ tiene una distribución normal con media $\mathbf{0}$ y

$$Cov(\bar{\mathbf{x}}_{(i)} - \bar{\bar{\mathbf{x}}}) = \left(1 - \frac{1}{n}\right)^2 Cov(\bar{\mathbf{x}}_{(i)}) + \left(\frac{n-1}{n^2}\right) Cov(\bar{\mathbf{x}}_{(1)}) = \frac{(n-1)}{nm}\mathbf{\Sigma},$$

donde

$$\bar{\bar{\mathbf{x}}} = \frac{1}{n}\sum_{i=1}^{n} \bar{\mathbf{x}}_{(i)}.$$

Por otra parte, las covarianzas muestrales pueden ser combinadas para obtener un estimador. Este estimador es

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{S}_i.$$

Aquí $(nm - n)\mathbf{S}$ es independiente de cada $\bar{\mathbf{x}}_{(i)}$ y, por lo tanto, de su media $\bar{\bar{\mathbf{x}}}$. Además, $(nm - n)\mathbf{S}$ se distribuye como una matriz aleatoria Wishart con $nm - n$ grados de libertad. Vale notar que se estima $\mathbf{\Sigma}$ de los datos recolectados en cada periodo. Al combinar estos estimadores se obtiene un estimador con un mayor número de grados de libertad. Consecuentemente,

$$T^2 = \frac{nm}{n-1}(\bar{\mathbf{x}}_{(i)} - \bar{\bar{\mathbf{x}}})'\mathbf{S}^{-1}(\bar{\mathbf{x}}_{(i)} - \bar{\bar{\mathbf{x}}}),$$

la cual se distribuye como

$$T^2 \sim \frac{(nm-n)p}{(nm-n-p+1)} F_{p,nm-n-p+1}.$$

Cartas formato elipse

En analogía con la discusión sobre observaciones multivariadas individuales, la carta de control elipsoidal para medias submuestrales es

$$(\bar{\mathbf{x}}_{(i)} - \bar{\bar{\mathbf{x}}})'\mathbf{S}^{-1}(\bar{\mathbf{x}}_{(i)} - \bar{\bar{\mathbf{x}}}) \leq \frac{2(n-1)(m-1)}{m(nm-n-1)} F_{2,nm-n-1}(0{,}05),$$

aunque el lado derecho se suele aproximar a $\chi^2_2(0{,}05)/m$. Las submuestras correspondientes a los puntos fuera de las elipses de control deben ser revisadas cuidadosamente para analizar el cambio de las características de calidad medidas.

Cartas $\mathbf{T}^2$

Para construir una carta T^2 con datos submuestrales y p características se grafica la expresión

$$T_i^2 = m(\bar{\mathbf{x}}_{(i)} - \bar{\bar{\mathbf{x}}})'\mathbf{S}^{-1}(\bar{\mathbf{x}}_{(i)} - \bar{\bar{\mathbf{x}}}),$$

para cada $i = 1, 2, \ldots, n$ donde el límite de control superior esta dado por

$$LCS = \frac{(n-1)(m-1)p}{(nm-n-p+1)} F_{p,nm-n-p+1}(0{,}05).$$

El LCS es a menudo aproximado a $\chi^2_p(0{,}05)$ cuando n es grande. Los valores T_i^2 que exceden el LCS corresponden a las potenciales causas de variación que pueden sacar de control al proceso.

A.3.4. Regiones de control para futuras observaciones submuestrales

Una vez recolectados los datos del proceso estable pueden ser usados para fijar los límites de control para futuras observaciones de medias submuestrales.

Si $\bar{\mathbf{x}}$ es una media submuestral futura, entonces $\bar{\mathbf{x}} - \bar{\bar{\mathbf{x}}}$ tiene una distribución normal multivariada con media $\mathbf{0}$ y

$$Cov(\bar{\mathbf{x}} - \bar{\bar{\mathbf{x}}}) = Cov(\bar{\mathbf{x}}) + \frac{1}{n} Cov(\bar{\bar{\mathbf{x}}}) = \frac{(n+1)}{nm}\mathbf{\Sigma},$$

consecuentemente.

$$T^2 = \frac{nm}{n+1}(\bar{\mathbf{x}} - \bar{\bar{\mathbf{x}}})'\mathbf{S}^{-1}(\bar{\mathbf{x}} - \bar{\bar{\mathbf{x}}}),$$

la cual se distribuye como

$$T^2 \sim \frac{(nm-n)p}{(nm-n-p+1)} F_{p,nm-n-p+1}.$$

Carta formato elipse

La elipse de predicción para futuras medias submuestrales de $p = 2$ características se define por el conjunto de todas las $\bar{\mathbf{x}}$ tales que

$$(\bar{\mathbf{x}} - \bar{\bar{\mathbf{x}}})' \mathbf{S}^{-1} (\bar{\mathbf{x}} - \bar{\bar{\mathbf{x}}}) \leq \frac{2(n+1)(m-1)}{m(nm-n-1)} F_{2,nm-n-1}(0{,}05),$$

donde, nuevamente, el lado derecho es aproximado usualmente a $\chi^2_2(0{,}05)/m$.

Carta $\mathbf{T}^2$

Como antes, se pasa $n/(n+1)$ al límite de control y se grafica

$$T^2 = m(\bar{\mathbf{x}} - \bar{\bar{\mathbf{x}}})' \mathbf{S}^{-1} (\bar{\mathbf{x}} - \bar{\bar{\mathbf{x}}}),$$

para las futuras medias muestrales en orden cronológico. El límite superior es definido por

$$LCS = \frac{(n+1)(m-1)p}{(nm-n-p+1)} F_{p,nm-n-p+1}(0{,}05),$$

el cual puede ser aproximado a $\chi^2_p(0{,}05)$ cuando n es grande.

Los puntos fuera de la elipse de predicción o cerca al LCS sugieren que los valores actuales de las características de calidad son diferentes de alguna forma a aquellas del proceso estable previo. Esto puede ser bueno o malo, lo único certero es la necesidad de buscar cuidadosamente las razones de este cambio.

Bibliografía

[1] **Abdi, H., Valentin, D. (2007).** Multiple correspondence analysis. Encyclopedia of measurement and statistics, 2(4), 651-657.

[2] **Abdi, H., & Williams, L. J. (2010).** Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4), 433-459.

[3] **Ahmed, S., Taporoski, T., Gómez, L., Ruiz, F., Beijamini, F., Horimoto, A, et al. (2019).** Data from the brazilian baependi heart study cohort yield new insights into the genetic epidemiology of insomnia. BMJ Open Respiratory Research. Vol. 6.

[4] **Ahsan, M., Mashuri, M., Kuswanto, H., Prastyo, D. D., Khusna, H. (2018).** Multivariate control chart based on PCA mix for variable and attribute quality characteristics. Production Manufacturing Research, 6(1), 364-384.

[5] **Albert, P. R. (2015).** Why is depression more prevalent in women?. J Psychiatry Neurosci;40(4).

[6] **Almalki, S. (2016).** Integrating Quantitative and Qualitative Data in Mixed Methods Research–Challenges and Benefits. Journal of education and learning, 5(3), 288-296.

[7] **Arévalo-Avecillas, D., Nájera-Acuña, S., & Piñero, E. A. (2018).** La Influencia de la Implementación de las Tecnologías de Información en la Productividad de Empresas de Servicios. Información tecnológica, 29(6), 199-212.

[8] **Balasubramanian, M., Schwartz, E. L., Tenenbaum, J. B., de Silva, V., Langford, J. C. (2002).** The isomap algorithm and topological stability. Science, 295(5552), 7-7.

[9] **Beijamini, F. et al. (2016).** Timing and quality of sleep in a rural Brazilian family-based cohort, the Baependi Heart Study. Sci. Rep. 6, 39283, doi: 10.1038/srep39283.

[10] **Belkin, M., Niyogi, P. (2001).** Laplacian eigenmaps and spectral techniques for embedding and clustering. In Nips (Vol. 14, No. 14, pp. 585-591).

[11] **Belkin, M., Niyogi, P. (2003).** Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation, 15(6), 1373-1396.

[12] **Benzécri, J. P. (1973).** L'analyse des données (Vol. 2, p. l). Paris: Dunod.

[13] **Bersimis, S., Psarakis, S., Panaretos, J. (2007).** Multivariate statistical process control charts: an overview. Quality and Reliability engineering international, 23(5), 517-543.

[14] **Bezdek, J. C., Keller, J., Krisnapuram, R., & Pal, N. (1999).** Fuzzy models and algorithms for pattern recognition and image processing (Vol. 4). Springer Science Business Media.

[15] **Bishop, C. M. (2006).** Pattern recognition and machine learning. springer.

[16] **Cachofeiro, V. (2009).** Alteraciones del colesterol y enfermedad cardiovascular. Lopez Farré A., Macaya Miguel C. et al Libro de la salud cardiovascular. 1ª ed. Bilbao: Fundación BBVA, 131-139.

[17] **Carreira-Perpinán, M. A., Lu, Z. (2007, March).** The laplacian eigenmaps latent variable model. In Artificial Intelligence and Statistics (pp. 59-66). PMLR.

[18] **Chen, J., Liu, Y. (2011).** Locally linear embedding: a survey. Artificial Intelligence Review, 36(1), 29-48.

[19] **Choong, A. C. H., Lee, N. K. (2017, November).** Evaluation of convolutionary neural networks modeling of DNA sequences using ordinal versus one-hot encoding method. In 2017 International Conference on Computer and Drone Applications (IConDA) (pp. 60-65). IEEE.

[20] **Costantini, P., Linting, M., & Porzio, G. C. (2010).** Mining performance data through nonlinear PCA with optimal scaling. Applied Stochastic Models in Business and Industry, 26(1), 85-101.

[21] **Cox, M. A., Cox, T. F. (2008).** Multidimensional scaling. In Handbook of data visualization (pp. 315-347). Springer, Berlin, Heidelberg.

[22] **Crawford, J., Hughes, C. E., & Lykoudis, S. (2014).** Alternative least squares methods for determining the meteoric water line, demonstrated using GNIP data. Journal of Hydrology, 519, 2331-2340.

[23] **Datta, A., Ghosh, S., Ghosh, A. (2018).** PCA, kernel PCA and dimensionality reduction in hyperspectral images. In Advances in Principal Component Analysis (pp. 19-46). Springer, Singapore.

[24] **Diaz, L. G., & Morales, M. A. (2012).** Análisis estadístico de datos multivariados. Universidad Nacional de Colombia.

[25] **Di Franco, G. (2016).** Multiple correspondence analysis: one only or several techniques?. Quality Quantity, 50(3), 1299-1315.

[26] **Doersch, C. (2016).** Tutorial on variational autoencoders. arXiv preprint arXiv:1606.05908.

[27] **Du, T. Y. (2019).** Dimensionality reduction techniques for visualizing morphometric data: comparing principal component analysis to nonlinear methods. Evolutionary Biology, 46(1), 106-121.

[28] **Duarte, N., Giolo, S., & de Andrade, M. (2015).** On the equivalence of methods for population stratification and their application in genetic association studies. Rev. Bras. Biom., São Paulo, v.33, n.4, 494-507.

[29] **Dunteman, G. H. (1989).** Principal components analysis (No. 69). Sage.

[30] **Egan, K. J., Von Schantz, M., Negrão, A. B., Santos, H. C., Horimoto, A. R., Duarte, N. E., ... & Pereira, A. C. (2016).** Cohort profile: the Baependi Heart Study—a family-based, highly admixed cohort study in a rural Brazilian town. BMJ open, 6(10).

[31] **Escribano Hernández, A., Vega Alonso, A. T., Lozano Alonso, J. E., Álamo Sanz, R., Castrodeza Sanz, J. J., & Lleras Muñoz, S. (2010).** Dislipidemias y riesgo cardiovascular en la población adulta de Castilla y León. Gaceta Sanitaria, 24, 282-287.

[32] **E. P˛ekalska, D. de Ridder, R.P.W. Duin, and M.A. Kraaijveld (1999).** A new method of generalizing Sammon mapping with

application to algorithm speed-up. ASCI'99, 5th Annual Conference of the Advanced School for Computing and Imaging, pages 221–228.

[33] **Feng, X., Xie, Y., Song, M., Yu, W., Tang, J. (2018, November).** Fast randomized PCA for sparse data. In Asian conference on machine learning (pp. 710-725). PMLR.

[34] **Friedman, N. P. & Miyake, A. (2017).** Unity and diversity of executive functions: Individual differences as a window on cognitive structure. Cortex, 86, 186-204.

[35] **Gabriel, K. R. (1971).** The biplot graphic display of matrices with application to principal component analysis. Biometrika 58, 453-467.

[36] **Géron, A. (2019).** Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media.

[37] **Ghojogh, B., Ghodsi, A., Karray, F., Crowley, M. (2020).** Multidimensional scaling, Sammon mapping, and Isomap: Tutorial and survey. arXiv preprint arXiv:2009.08136.

[38] **Gifi, A. (1985).** Princals. Department of Data Theory.

[39] **Gifi, A. (1989).** Algorithm Descriptions for Anacor, Homal, Princals and Overals. Department of Data Theory, University of Leiden.

[40] **Gifi, A. (1990).** Nonlinear multivariate analysis. Wiley.

[41] **Gower, J. C. (1966).** Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika, v.53, 325-338.

[42] **Gower, J. C., Le Roux, N. J., & Gardner-Lubbe, S. (2015).** Biplots: quantitative data. WIREs Comput Stat, 7:42–62.

[43] **Gower, J. C., Le Roux, N. J., & Gardner-Lubbe, S. (2016).** Biplots: qualititative data. WIREs Comput Stat, 8:82–111.

[44] **Greenacre, M. J. (1984).** Theory and application of Correspondence Analysis. London: Academic Press.

[45] **Greenacre, M., Blasius, J. (2006).** Multiple correspondence analysis and related methods. Chapman and Hall/CRC.

[46] **Guttman, L. (1941).** The quantification of a class of attributes: A theory and method of scale construction. The Prediciton of Personal Adjustment.

[47] **Hinton, G. E. (1986).** Proceedings of the eighth annual conference of the cognitive science society.

[48] **Hoffmann, H. (2007).** Kernel PCA for novelty detection. Pattern recognition, 40(3), 863-874.

[49] **Hotelling, H. (1933).** Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24.

[50] **Hout, M. C., Papesh, M. H., Goldinger, S. D. (2013).** Multidimensional scaling. Wiley Interdisciplinary Reviews: Cognitive Science, 4(1), 93-103.

[51] **Ipsen, Ilse C. F. (2009).** Numerical Matrix Analysis: Linear Systems and Least Squares. SIAM (Society for Industrial and Applied Mathematics). Philadelphia.

[52] **Jenssen, R. (2009).** Kernel entropy component analysis. IEEE transactions on pattern analysis and machine intelligence, 32(5), 847-860.

[53] **Jin, J., Loosveldt, G. (2019).** Assessing response quality by using multivariate control charts for numerical and categorical response quality indicators. Journal of Survey Statistics and Methodology.

[54] **Jhonson, R. A., & Wichern, D. W. (2007).** Applied multivariate statistical analysis. New Jersey: Prentice Hall, 794p.

[55] **Khemakhem, I., Kingma, D., Monti, R., Hyvarinen, A. (2020, June).** Variational autoencoders and nonlinear ica: A unifying framework. In International Conference on Artificial Intelligence and Statistics (pp. 2207-2217). PMLR.

[56] **Kiers, H. A. (2002).** Setting up alternating least squares and iterative majorization algorithms for solving various matrix optimization problems. Computational statistics data analysis, 41(1), 157-170.

[57] **Kingma, D. P., Welling, M. (2019).** An introduction to variational autoencoders. arXiv preprint arXiv:1906.02691.

[58] **Kruskal, J. B. (1964).** Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika, 29(1), 1-27.

[59] **Kruskal, J. B. (1965).** Analysis of factorial experiments by estimating monotone transformations of the data. Journal of the Royal Statistical Society: Series B (Methodological), 27(2), 251-263.

[60] **Kruskal, J. B., & Shepard, R. N. (1974).** A nonmetric variety of linear factor analysis. Psychometrika, 39(2), 123-157.

[61] **Kuhfeld, W. F. (1990).** SAS Technical Report R-108: Algorithms for the PRINQUAL and TRANSREG Procedures. Cary NC: SAS Institute Inc.

[62] **Kuroda, M., Mori, Y., Iizuka, M. (2020).** Initial value selection for the alternating least squares algorithm. In Advanced Studies in Classification and Data Science (pp. 227-239). Springer, Singapore.

[63] **Lahera, V., & de las Eras, N. (2009).** Libro de la salud cardiovascular. Diabetes y riesgo cardiovascular. Capítulo 11.

[64] **Lackland, D. T. (2014).** Racial Differences in Hypertension: Implications for High Blood Pressure Management. Am J Med Sci; 348(2), 135–138. doi:10.1097/MAJ.0000000000000308.

[65] **Laub, J. H., & Sampson, R. J. (1998).** Integrating quantitative and qualitative data. Methods of life course research: Qualitative and quantitative approaches, 213-230.

[66] **Lawrence, N. D. (2003, December)**. Gaussian process latent variable models for visualisation of high dimensional data. In Nips (Vol. 2, p. 5).

[67] **Lawrence, N., Hyvärinen, A. (2005).** Probabilistic non-linear principal component analysis with Gaussian process latent variable models. Journal of machine learning research, 6(11).

[68] **Lee, J. A., Verleysen, M. (2007).** Nonlinear dimensionality reduction. Springer Science Business Media.

[69] **De Leeuw, J., Young, F. W., & Takane, Y. (1976).** Additive structure in qualitative data: An alternating least squares method with optimal scaling features. Psychometrika, 41(4), 471-503.

[70] **De Leeuw, J., & Van Rijckevorsel, J. (1980).** HOMALS and PRINCALS—Some generalizations of principal components analysis. Data analysis and informatics, 2, 231-42.

[71] **De Leeuw, J., Heiser, W. (1982).** 13 Theory of multidimensional scaling. Handbook of statistics, 2, 285-316.

[72] **De Leeuw, J. (1984).** The Gifi system of nonlinear multivariate analysis. Data analysis and informatics III, 415-424.

[73] **Li, B., Li, Y. R., Zhang, X. L. (2019).** A survey on Laplacian eigenmaps based manifold learning methods. Neurocomputing, 335, 336-351.

[74] **Li, P., Chen, S. (2016).** A review on Gaussian process latent variable models. CAAI Transactions on Intelligence Technology, 1(4), 366-376.

[75] **Linting, M., Meulman, J. J., Groenen, P. J., & van der Koojj, A. J. (2007).** Nonlinear principal components analysis: introduction and application. Psychological methods, 12(3), 336.

[76] **Linting, M., & van der Kooij, A. (2012).** Nonlinear principal components analysis with CATPCA: a tutorial. Journal of personality assessment, 94(1), 12-25.

[77] **Mair, P. (2018).** Gifi Methods. In Modern Psychometrics with R (pp. 231-256). Springer, Cham.

[78] **McLean, C. P., Asnaani, A., Litz, B. T., Hofmann, S. G. (2011).** Gender Differences in Anxiety Disorders: Prevalence, Course of Illness, Comorbidity and Burden of Illness. J Psychiatr; 45(8), 1027–1035. doi:10.1016/j.jpsychires.2011.03.006.

[79] **Meulman, J. J. (1998).** Optimal scaling methods for multivariate categorical data analysis. SPSS White Paper: Chicago.

[80] **Meulman, J. J., Van der Kooij, A. J., & Heiser, W. J. (2004).** Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. The Sage handbook of quantitative methodology for the social sciences, 49-72.

[81] **Minsky, M., Papert, S. A. (2017).** Perceptrons: An introduction to computational geometry. MIT press.

[82] **Michailidis, G., & De Leeuw, J. (1998).** The Gifi system of descriptive multivariate analysis. Statistical Science, 307-336.

[83] **Miljković, D. (2017, May).** Brief review of self-organizing maps. In 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) (pp. 1061-1066). IEEE.

[84] **Montgomery, D. C. (2012).** Statistical quality control. Wiley Global Education.

[85] **Nasiriany, S., Thomas, G., Wang, W., Yang, A., Listgarten, J., & Sahai, A. (2019).** A Comprehensive Guide to Machine Learning. Department of Electrical Engineering and Computer Sciences. University of California, Berkeley. http://snasiriany.me/files/ml-book.pdf.

[86] **Ng, A. (2011).** Sparse autoencoder. CS294A Lecture notes, 72(2011), 1-19.

[87] **de Oliveira, C. M., Pereira, A. C., De Andrade, M., Soler, J. M., & Krieger, J. E. (2008).** Heritability of cardiovascular risk factors in a Brazilian population: Baependi Heart Study. BMC medical genetics, 9(1), 1-8.

[88] **de Oliveira, C. M., Ulbrich, A. Z., Neves, F. S., Dias, F.A. L., Horimoto, A.R.V.R., Krieger, J. E., et al. (2017).** Association between anthropometric indicators of adiposity and hypertension in a Brazilian population: Baependi Heart Study. PLoS ONE 12 (10): e0185225. https://doi.org/10.1371/journal.pone.0185225.

[89] **Oliveira, G. F., Oliveira, T. R., Ikejiri, A. T., Andraus, M. P., Galvao, T. F., et al. (2014).** Prevalence of Hypertension and Associated Factors in an Indigenous Community of Central Brazil: A Population-Based Study. 6, 19; doi:10.3390/jcdd6020019.

[90] **Özdemir, V., & Hekim, N. (2018).** Birth of industry 5.0: Making sense of big data with artificial intelligence,"the internet of things" and next-generation technology policy. Omics: a journal of integrative biology, 22(1), 65-76.

[91] **Potdar, K., Pardawala, T. S., Pai, C. D. (2017).** A comparative study of categorical variable encoding techniques for neural network classifiers. International journal of computer applications, 175(4), 7-9.

[92] **Pearson, K. (1901).** On lines and planes of closest fit to systems of points in space. Philosophical Magazine, 6, 559 - 572.

[93] **Peña, D. (2002).** Análisis de Datos Multivariantes. https://www.researchgate.net/publication/40944325.

[94] **Perreault Jr, W. D., & Young, F. W. (1980).** Alternating least squares optimal scaling: Analysis of nonmetric data in marketing research. Journal of Marketing Research, 17(1), 1-13.

[95] **Quiroga-Parra, D. J., Torrent-Sellens, J., & Murcia-Zorrilla, C. P. (2017).** Las tecnologías de la información en América Latina, su incidencia en la productividad: Un análisis comparado con países desarrollados. Dyna, 84(200), 281-290.

[96] **Rencher, A. C. (1998).** Multivariate statistical inference and applications (p. 559). New York: Wiley.

[97] **Rencher, A. C. (2002).** Methods of Multivariate Analysis. A Wiley-Interscience publication. ISBN 0-471-41889-7.

[98] **Revista de Endocrinología y Nutrición (2004**); Complicaciones macrovasculares en la diabetes mellitus tipo 2. Vol. 12, No. 2 Supl.1. pp S23-S30.

[99] **Rodgers, J. L., Jones, J., Bolleddu, S. I., Vanthenapalli, S., Rodgers, L. E., Shah, K., Karia, K. & Panguluri, S. K. (2019).** Cardiovascular Risks Associated with Gender and Aging. Journal of Cardiovascular Development and Disease. 6, 19; doi:10.3390/jcdd6020019.

[100] **Rodríguez, H. E. D. (2017).** Tecnologías de la información y comunicación y crecimiento económico. Economía Informa, 405, 30-45.

[101] **Roweis, S. T., Saul, L. K. (2000).** Nonlinear dimensionality reduction by locally linear embedding. science, 290(5500), 2323-2326.

[102] **Rumelhart, D. E., Hinton, G. E., Williams, R. J. (1985).** Learning internal representations by error propagation. California Univ San Diego La Jolla Inst for Cognitive Science.

[103] **Rumelhart, D. E., Hinton, G. E., Williams, R. J. (1986).** Learning representations by back-propagating errors. nature, 323(6088), 533-536.

[104] **Sammon, J. W. (1969).** A nonlinear mapping for data structure analysis. IEEE Transactions on computers, 100(5), 401-409.

[105] **Sands, R., & Young, F. W. (1980).** Component models for three-way data: An alternating least squares algorithm with optimal scaling features. Psychometrika, 45(1), 39-67.

[106] **Saul, L. K., Roweis, S. T. (2003).** Think globally, fit locally: unsupervised learning of low dimensional manifolds. Departmental Papers (CIS), 12.

[107] **Santos, M. Y., e Sá, J. O., Costa, C., Galvão, J., Andrade, C., Martinho, B., & Costa, E. (2017).** A big data analytics architecture for industry 4.0. In World Conference on Information Systems and Technologies (pp. 175-184). Springer, Cham.

[108] **Schölkopf, B., Smola, A., Müller, K. R. (1998).** Nonlinear component analysis as a kernel eigenvalue problem. Neural computation, 10(5), 1299-1319.

[109] **Scholz, M. (2012).** Validation of nonlinear PCA. Neural processing letters, 36(1), 21-30.

[110] **Shawe-Taylor, J., Williams, C. K., Cristianini, N., Kandola, J. (2005).** On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA. IEEE Transactions on Information Theory, 51(7), 2510-2522.

[111] **Shepard, R. N. (1962).** The analysis of proximities: multidimensional scaling with an unknown distance function. I. Psychometrika, 27(2), 125-140.

[112] **Shepard, R. N. (1966).** Metric structures in ordinal data. Journal of Mathematical Psychology, 3(2), 287-315.

[113] **Sriperumbudur, B., Sterge, N. (2017).** Approximate kernel PCA using random features: Computational vs. statistical trade-off. arXiv preprint arXiv:1706.06296.

[114] **Stewart, G. W. (1993).** On the Early History of the Singular Value Decomposition. SIAM REVIEW. Vol 35(4), 551 - 566.

[115] **Takane, Y., Young, F. W., & De Leeuw, J. (1977).** Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. Psychometrika, 42(1), 7-67.

[116] **Taporoski, T. P., Negrão, A. B., Horimoto, A. R. V. R., Duarte, N. .E, Alvim, R. O., de Oliveira, C. M., et al. (2015).** Shared Genetic Factors of Anxiety and Depression Symptoms in a Brazilian Family-Based Cohort, the Baependi Heart Study. PLoS ONE 10(12): e0144255, 1-10. doi:10.1371/journal.pone.0144255

[117] **Tenenbaum, J. B. (1998).** Mapping a manifold of perceptual observations. Advances in neural information processing systems, 10, 682-688.

[118] **Torgerson, W. S. (1952).** Multidimensional scaling: I. Theory and method. Psychometrika, 17(4), 401-419.

[119] **Tenenbaum, J. B., De Silva, V., Langford, J. C. (2000).** A global geometric framework for nonlinear dimensionality reduction. science, 290(5500), 2319-2323.

[120] **Tuerhong, G., Kim, S. B. (2014).** Gower distance-based multivariate control charts for a mixture of continuous and categorical variables. Expert systems with applications, 41(4), 1701-1707.

[121] **Tschannen, M., Bachem, O., Lucic, M. (2018).** Recent advances in autoencoder-based representation learning. arXiv preprint arXiv:1812.05069.

[122] **Van der Maaten, L., Hinton, G. (2008).** Visualizing data using t-SNE. Journal of machine learning research, 9(11).

[123] **Van der Burg, E., De Leeuw, J., & Verdegaal, R. (1988).** Homogeneity analysis withk sets of variables: An alternating least squares method with optimal scaling features. Psychometrika, 53(2), 177-197.

[124] **Venna, J., Peltonen, J., Nybo, K., Aidos, H., Kaski, S. (2010).** Information retrieval perspective to nonlinear dimensionality reduction for data visualization. Journal of Machine Learning Research, 11(2).

[125] **Von der Malsburg, C. (1973).** Self-organization of orientation sensitive cells in the striate cortex. Kybernetik, 14(2), 85-100.

[126] **Von Schantz, M. et al. (2015).** Distribution and heritability of diurnal preference (chronotype) in a rural Brazilian family-based cohort, the Baependi study. Sci. Rep., 5:9214, 1-6. DOI:10.1038/srep09214

[127] **Waagen, D., Hulsey, D., Godwin, J., Gray, D., Barton, J., Farmer, B. (2021, April).** t-SNE or not t-SNE, that is the question. In Automatic Target Recognition XXXI (Vol. 11729, p. 117290B). International Society for Optics and Photonics.

[128] **Wattenberg, M., Viégas, F., Johnson, I. (2016).** How to use t-SNE effectively. Distill, 1(10), e2.

[129] **Weinberger, K. Q., Saul, L. K. (2006, July).** An introduction to nonlinear dimensionality reduction by maximum variance unfolding. In AAAI (Vol. 6, pp. 1683-1686).

[130] **Werbos, P. J. (2008).** Foreword: ADP-The Key Direction for Future Research in Intelligent Control and Understanding Brain Intelligence. IEEE Trans. Syst. Man Cybern. Part B, 38(4), 898-900.

[131] **Wetzel, S. J. (2017).** Unsupervised learning of phase transitions: From principal component analysis to variational autoencoders. Physical Review E, 96(2), 022140.

[132] **Williams, C. K. (2001).** On a connection between kernel PCA and metric multidimensional scaling. In Advances in neural information processing systems (pp. 675-681).

[133] **Williams, C. K. (2002).** On a connection between kernel PCA and metric multidimensional scaling. Machine Learning, 46(1), 11-19.

[134] **Wold, S., Esbensen, K., & Geladi, P. (1987).** Principal component analysis. Chemometrics and intelligent laboratory systems, 2(1-3), 37-52.

[135] **Yang, L. (2004, August).** Sammon's nonlinear mapping using geodesic distances. In Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. (Vol. 2, pp. 303-306). IEEE.

[136] **Young, F. W., De Leeuw, J., & Takane, Y. (1976).** Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. Psychometrika, 41(4), 505-529.

[137] **Young, F. W., Takane, Y., & de Leeuw, J. (1978).** The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. Psychometrika, 43(2), 279-281.

[138] **Young, G., Householder, A. S. (1938).** Discussion of a set of points in terms of their mutual distances. Psychometrika, 3(1), 19-22.

[139] **Zhang, Z., Chow, T. W., Zhao, M. (2012).** M-Isomap: Orthogonal constrained marginal isomap for nonlinear dimensionality reduction. IEEE transactions on cybernetics, 43(1), 180-191.