Benemérita Universidad Autónoma de Puebla

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS
POSGRADO EN MATEMÁTICAS

Análisis de Correlación Canónica lineal y no lineal

TESIS

que para obtener el grado de

Doctora en Ciencias (Matemáticas)

Presenta

Brenda Catalina Matías Castillo

Director de tesis

Dra. Gladys Linares Fleites
Dra. Hortensia Josefina Reyes Cervantes

Puebla, Pue. Julio 2017

ÍNDICE GENERAL

Introducción

El Análisis de Correlación Canónica fue desarrollado por Hotelling [11] en 1936 como un procedimiento para evaluar la relación lineal entre dos conjuntos de variables aleatorias. Dentro del campo de la Estadística Multivariada, el Análisis de Correlación Canónica (ACC) se presenta como un método exploratorio de datos multivariados, y se basa en resultados del álgebra matricial [7],[13],[19], [32]. Su propósito es la exploración de las correlaciones entre dos conjuntos de variables cuantitativas observadas sobre el mismo conjunto de individuos, a través de combinaciones lineales de las variables iniciales, lo que permite reducir la dimensionalidad. El ACC no puede ser llevado a cabo cuando el número de individuos es menor al número de variables. Una manera para tratar con este problema es incluir un paso de regularización en el cálculo del ACC, obteniendo un método llamado Análisis de Correlación Canónica Regularizada (ACCR)[8]. Por otro lado, los tipos de análisis antes mencionados, son utilizados cuando se tienen dos grupos de variables. El Análisis de Correlación Canónica Regularizada Generalizada (ACCRG) es aplicado a tres o más conjuntos de variables, observados en el mismo conjunto de individuos [41]. Todas estas técnicas se han aplicado ampliamente y algunos de esos estudios se encuentran en [10], [18] y [40].

No es hasta finales de los años 90's del siglo pasado que se desarrollaron procedimientos no lineales como generalizaciones de las técnicas clásicas de Análisis Multivariado. A partir de estos desarrollos se introdujeron los métodos kernel [1] que han tenido una influencia considerable en la Estadística Multivariada y el Aprendizaje Automático. Se han propuesto varios kernel en la literatura, que han motivado investigaciones para determinar de manera adecuada (en un sentido matemático) el kernel para el problema bajo estudio, ya sea en la exploración de datos, en la reducción de dimensión o en problemas de clasificación. La propuesta de Schölkopf *et al.* en [34], [35], [36] y [37], es ampliamente utilizada en muchos trabajos, ya que estos autores introducen una nueva clase de algoritmos para técnicas del Análisis Multivariado. En el caso del ACC, estos algoritmos se basan en la idea de transformar los datos mediante una función no lineal, hacia un espacio de mayor dimensión en el que se encuentran los datos y realizar el Análisis Multivariado en los datos transformados. Estos nuevos enfoques se conocen como Análisis de Correlación Canónica con Kernel (ACCK). El ACCK ofrece una solución alternativa por medio de la proyección de los datos en un espacio de características de alta dimensión. Una propuesta de solución

es la presentada por Hardoon [9] que requiere solucionar el ACCK mediante un problema de valores y vectores propios generalizado. Sin embargo, el ACCK carece de interpretabilidad y robustez frente a características irrelevantes. Chang *et al.* [5] introdujeron dos extensiones del ACC no lineales que dependen del Criterio de Independencia de Hilbert-Schmidt y la Alineación Objetivo del Kernel Centrado (HSIC y KTA respectivamene por sus siglas en inglés), que ayudan a demostrar que el uso de proyecciones lineales permite la eliminación de características irrelevantes, mientras extrae combinaciones de características fuertemente asociadas. Como puede apreciarse el ACC, por su importancia en las aplicaciones, ha generado y sigue generando mucho trabajo de investigación.
Esta tesis tiene como objetivo el estudio del ACC en un sentido amplio. Para llevar a cabo este estudio, se revisó desde el ACC que se presenta en la literatura general del Análisis Multivariado, así como, algunas extensiones, haciendo hincapié en los métodos de optimización que se emplean clásicamente en la obtención de las correlaciones canónicas. Posteriormente se estudió el ACCK así como algunas soluciones no lineales del ACC, presentadas en trabajos actuales, como lo es el método HSIC y KTA, ya que cada uno de estos métodos es visto como un problema de maximización y la técnica que se utiliza para dar solución a cada uno de ellos es el método del gradiente descendente. Sin embargo, existen métodos como los Algoritmos Genéticos (AG) que son métodos adaptativos que pueden usarse para resolver problemas de búsqueda y optimización global. Estos métodos están basados en el proceso genético de los organismos vivos. En este trabajo se utilizan como un método de solución directa al ACC y al ACC no lineal.
A continuación se presenta el contenido de la tesis, en donde, dada la diversidad de procedimientos de correlaciones canónicas desarrollados, se ha decidido presentar los resultados teóricos y de aplicación alcanzados en la tesis en el capítulo correspondiente al procedimiento descrito.
En el Capítulo 1 se define al ACC, se presentan sus objetivos, diferentes métodos de solución encontradas en la literatura, así como la relación entre ellos, una descripción geométrica del problema y su interpretación. Se definen también extensiones del ACC mencionadas anteriormente como lo son el ACCR, ACCG y ACCRG. Finalmente, se presenta una aplicación al estudio de los manglares del sistema Lagunar de Chacahua-Pastorias en Oaxaca [4], usando cada uno de los métodos mencionados y usando la paquetería del software R [33].
El Capítulo 2 se inicia definiendo a las funciones kernel, las caracterizaciones a estas funciones, algunas propiedades entre ellas y algunos ejemplos de funciones kernel. Se define además el ACCK y se presenta un método de solución dado por Hardoon [9]. Después se definen dos métodos de solución al ACC no lineales que "mejoran"la solución propuesta para el ACCK, que son HSIC y KTA. Al igual que en el primer capítulo se presentan aplicaciones de estos métodos al estudio de interacciones medicamentosas y a un estudio de educación.
En el Capítulo 3 se definen los Algoritmos Genéticos y se presenta un amplio panorama sobre como estos son trabajados usualmente. Principalmente se hace un enfoque en su uso para dar solución al ACC mediante valores y vectores propios y además, su solución de forma directa a partir de la definición de

ACC. También se explica su uso para resolver el ACC no lineal de forma directa mediante el método KTA. Se incluye el programa que se realizó para realizar esta solución.
En el Capítulo 4 se llevan a cabo varios estudios comparativos utilizando las bases de datos de los problemas antes mencionados y otros, además de hacer comparaciones en datos reales, se hicieron estudios de simulación. Cabe destacar que se presentan los resultados de las aplicaciones del ACC no lineal mediante el uso del programa de Algoritmos Genéticos, que fue desarrollado en esta tesis.
En el Capítulo 5 se presentan conclusiones a este trabajo.
Se incluye, además un Apéndice dividido en tres partes: la primera sección sobre las herramientas matemáticas usadas, en la segunda sección se presentan las funciones usadas mediante el software R para las aplicaciones y en la tercera sección se incluyen los programas elaborados en esta tesis para algoritmos de ACC y ACCK-KTA.

Capítulo 1

Análisis de Correlación Canónica y sus extensiones

El Análisis de Correlaciones Canónicas (ACC) es un método de la Estadística Multivariada que permite investigar acerca de la existencia de correlaciones entre dos o más grupos de variables.
El método consiste en encontrar pares de nuevas variables (no observables), formadas por una combinación lineal de las variables originales (observadas) de cada grupo. Se determina la primera pareja de combinaciones lineales tal que tenga la correlación más grande; posteriormente, se obtiene la pareja de combinaciones lineales tal que tengan la segunda correlación más grande, y así sucesivamente. Las parejas de combinaciones lineales son llamadas variables canónicas y las correlaciones obtenidas mediante estas combinaciones son llamadas correlaciones canónicas.
En ocasiones es necesario incluir un paso previo de regularización en el cálculo del ACC; éste método es llamado Análisis de Correlación Canónica Regularizado (ACCR). También se presentan problemas en los que se requiere determinar el grado de relación que hay entre tres o más grupos de variables y el método que lo resuelve es llamado Análisis de Correlación Canónica Generalizada (ACCG). Así mismo, a veces es necesario utilizar la combinación entre el ACCR y ACCG mediante el método llamado Análisis de Correlación Canónica Regularizada Generalizada (ACCRG), que es aplicado a tres o más conjuntos de variables observados en el mismo conjunto de individuos, pero con el número de individuos menor que el número de variables. En este capítulo se formulan y se desarrollan los aspectos matemáticos de cada una de esos métodos y se brindan aplicaciones a problemas reales.

1.1. Análisis de Correlación Canónica Clásico

El ACC es generalmente utilizado para investigar la presencia de cualquier patrón de cambio que ocurra de forma simultánea en dos conjuntos de variables

por separado, para este objetivo se determina la correlación existente entre ellos [10]. En general, el ACC aborda los siguientes aspectos:

1. Calcula la magnitud de las relaciones lineales existentes entre dos conjuntos de variables observadas sobre los mismos individuos.

2. Deriva los pesos o ponderaciones para cada uno de los conjuntos de variables, de manera que las combinaciones lineales de cada conjunto estén máximamente correlacionados. Otras funciones lineales que maximizan la correlación restante son independientes del conjunto o conjuntos anteriores de combinaciones lineales.

3. Explica la naturaleza de las relaciones que existen entre los dos conjuntos de variables, por lo general mediante la medición de la contribución relativa de cada variable con las funciones canónicas (relaciones) que son extraídas.

A continuación se expondrán brevemente los aspectos matemáticos esenciales de este método multivariado.
En el ACC se está interesado en las medidas de asociación entre dos grupos de variables. El primer grupo, de p variables, está representado por el vector aleatorio $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p)$ de orden $(p \times 1)$ y el segundo grupo, de q variables, está representado por el vector aleatorio $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_q)$ de orden $(q \times 1)$, cada variable tomada sobre un número n de individuos. Se asume que, $p \leq q$. Los vectores $\mathbf{X}$ y $\mathbf{Y}$ están representadas como natrices de datos como a continuación

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \mathbf{Y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1q} \\ y_{21} & y_{22} & \cdots & y_{2q} \\ \vdots & \vdots & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nq} \end{pmatrix}$$

Para los vectores $\mathbf{X}$ y $\mathbf{Y}$ se calculan:

$$\begin{aligned} E[\mathbf{X}] = \mu_X, Var(\mathbf{X}) = \Sigma_{XX} \\ E[\mathbf{Y}] = \mu_Y, Var(\mathbf{Y}) = \Sigma_{YY} \\ Cov(\mathbf{X}, \mathbf{Y}) = \Sigma_{XY} = \Sigma'_{YX}. \end{aligned}$$

Se definen también U y V, que son combinaciones lineales que proporcionan medidas de resumen simples de los conjuntos de variables.

$$\begin{aligned} U &= \mathbf{a}'\mathbf{X}, \\ V &= \mathbf{b}'\mathbf{Y} \end{aligned}$$

donde $\mathbf{a}$ y $\mathbf{b}$ son vectores de coeficientes. Además, sean

$$\begin{aligned} Var(U) &= \mathbf{a}'\Sigma_{XX}\mathbf{a} \\ Var(V) &= \mathbf{b}'\Sigma_{YY}\mathbf{b} \\ Cov(U,V) &= \mathbf{a}'\Sigma_{XY}\mathbf{b}. \end{aligned}$$

Entonces el objetivo principal del ACC es buscar los vectores de coeficientes $\mathbf{a}$ y $\mathbf{b}$ tales que

$$\rho = Corr(U, V) = \frac{\mathbf{a}'\Sigma_{XY}\mathbf{b}}{\sqrt{\mathbf{a}'\Sigma_{XX}\mathbf{a}}\sqrt{\mathbf{b}'\Sigma_{YY}\mathbf{b}}} \tag{1.1}$$

sea máxima. Sujeto a

$$\mathbf{a}'\Sigma_{XX}\mathbf{a} = \mathbf{b}'\Sigma_{YY}\mathbf{b} = 1. \tag{1.2}$$

Los vectores de coeficientes $\mathbf{a}$ y $\mathbf{b}$ que cumplen esta condición son los primeros vectores canónicos, así entonces esta *primera pareja de variables canónicas*, es la pareja de combinaciones lineales U_1, V_1 con varianzas unitarias, las cuales maximizan la correlación (1.1) (primera correlación canónica). La *segunda pareja de variables canónicas*, es la pareja de combinaciones lineales U_2, V_2 con varianzas unitarias, las cuales maximizan la correlación (1.1) (segunda correlación canónica) y que no están relacionados con la primer pareja de variables canónicas. Así, la *k-ésima pareja de variables canónicas*, es la pareja de combinaciones lineales U_k, V_k con varianzas unitarias, las cuales maximizan la correlación (1.1) (k-ésima correlación canónica) y que no están relacionados con las $(k-1)$ parejas de variables canónicas anteriores [13].
El problema de las correlaciones canónicas no es más que un problema de extremos condicionados, por lo que tradicionalmente se resuelve aplicando el método de los multiplicadores de Lagrange, que conlleva a resolver un problema de valores y vectores propios aunque otra vía más es usar la descomposición singular de una matriz. El procedimiento a seguir para determinar las correlaciones y variables canónicas es descrito a continuación. Se tiene el problema de maximizar

$$f(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}'\Sigma_{XY}\mathbf{b}}{\sqrt{\mathbf{a}'\Sigma_{XX}\mathbf{a}}\sqrt{\mathbf{b}'\Sigma_{YY}\mathbf{b}}}$$

bajo las restricciones

$$\begin{aligned} g_1(\mathbf{a}, \mathbf{b}) &= \mathbf{a}'\Sigma_{XX}\mathbf{a} - 1 = 0 \\ g_2(\mathbf{a}, \mathbf{b}) &= \mathbf{b}'\Sigma_{YY}\mathbf{b} - 1 = 0, \end{aligned}$$

es decir, lo anterior plantea un problema de máximos con restricciones, entonces para encontrar su solución como se mencionó anteriormente se usa el método de multiplicadores de Lagrange. Considérese la función de Lagrange

$$F(\mathbf{a}, \mathbf{b}, \lambda, \beta) = \mathbf{a}'\Sigma_{XY}\mathbf{b} - \frac{\lambda}{2}(\mathbf{a}'\Sigma_{XX}\mathbf{a} - 1) - \frac{\beta}{2}(\mathbf{b}'\Sigma_{YY}\mathbf{b} - 1)$$

donde λ y β son multiplicadores de Lagrange. Para obtener el valor máximo se deben cumplir las condiciones de Karush-Kuhn-Tucker [22],

1. $\nabla_{\mathbf{a},\mathbf{b}} F(\mathbf{a}, \mathbf{b}, \lambda, \beta) = 0$,

2. $\nabla_{\lambda,\beta} F(\mathbf{a}, \mathbf{b}, \lambda, \beta) = 0$.

Entonces, derivando con respecto a cada variable e igualando a 0 se obtiene,

$$\frac{\partial F}{\partial \mathbf{a}} = \Sigma_{XY}\mathbf{b} - \lambda \Sigma_{XX}\mathbf{a} = 0, \tag{1.3}$$

$$\frac{\partial F}{\partial \mathbf{b}} = \Sigma_{YX}\mathbf{a} - \beta \Sigma_{YY}\mathbf{b} = 0, \tag{1.4}$$

$$\frac{\partial F}{\partial \lambda} = -\frac{1}{2}(\mathbf{a}'\Sigma_{XX}\mathbf{a} - 1) = 0, \qquad \frac{\partial F}{\partial \beta} = -\frac{1}{2}(\mathbf{b}'\Sigma_{YY}\mathbf{b} - 1) = 0. \tag{1.5}$$

Las ecuaciones (1.5) recuperan las reestricciones. Multiplicando (1.3) por $\mathbf{a}'$ y (1.4) por $\mathbf{b}'$, se tiene

$$\begin{aligned} \mathbf{a}'\Sigma_{XY}\mathbf{b} - \lambda \mathbf{a}'\Sigma_{XX}\mathbf{a} &= 0, \\ \mathbf{b}'\Sigma_{YX}\mathbf{a} - \beta \mathbf{b}'\Sigma_{YY}\mathbf{b} &= 0, \end{aligned}$$

entonces $\lambda = \beta = \mathbf{a}'\Sigma_{XY}\mathbf{b}$.
Despejando a $\mathbf{b}$ de (1.4) se obtiene

$$\lambda \Sigma_{YY}\mathbf{b} = \Sigma_{YX}\mathbf{a} \quad \Rightarrow \quad \mathbf{b} = \lambda^{-1}\Sigma_{YY}^{-1}\Sigma_{YX}\mathbf{a} \tag{1.6}$$

suponiendo que Σ_{YY} es invertible. Ahora se sustituye $\mathbf{b}$ en la ecuación (1.3), entonces se obtiene

$$\begin{aligned} & \Sigma_{XY}(\lambda^{-1}\Sigma_{YY}^{-1}\Sigma_{YX}\mathbf{a}) - \lambda \Sigma_{XX}\mathbf{a} = 0 \\ \Rightarrow \quad & \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\mathbf{a} = \lambda^2 \Sigma_{XX}\mathbf{a}. \end{aligned} \tag{1.7}$$

Como se puede observar, es un problema de valores y vectores propios generalizado, el cual requiere que Σ_{YY} sea invertible. De manera similar se trabaja para $\mathbf{b}$, despejando $\mathbf{a}$ de (1.3) y sustituyendo en (1.4), se obtiene

$$\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\mathbf{b} = \lambda^2 \Sigma_{YY}\mathbf{b}. \tag{1.8}$$

Note que una vez obtenidos los valores de $\mathbf{a}$ y λ se calcula el valor de $\mathbf{b}$ mediante la ecuación (1.6).
A continuación se presentan tres métodos, mencionados en la literatura, para determinar las correlaciones canónicas y vectores canónicos a partir de la ecuación (1.7). Para cada uno de los métodos presentados en las subsecciones anteriores se realizó un programa (ver Apéndice C, Cuadros C.1, C.2 y C.3) usando el software OCTAVE [30]. Posteriormente se presentan ejemplos en donde se hace una comparación de los resultados obtenidos mediante el programa de cada método.

1.1.1. ACC usando Descomposición en Valores Singulares

Éste método es presentado por Mardia *et al.* [23], quienes definen una matriz $\mathbf{K}$, le aplica el teorema de Descomposición en Valores Singulares (SVD con siglas en inglés) llegando a un problema de vectores y valores propios.
Bajo la suposición de que Σ_{XX} y Σ_{YY} son invertibles, sea

$$\mathbf{K} = \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}, \tag{1.9}$$

una vez obtenidas las inversas de Σ_{XX} y Σ_{YY} en las ecuaciones (1.7) y (1.8) respectivamente, se definen,

$$\mathbf{M}_1 = \Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} = \Sigma_{XX}^{-1/2}\mathbf{N}_1\Sigma_{XX}^{1/2}, \tag{1.10}$$

$$\mathbf{M}_2 = \Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} = \Sigma_{YY}^{-1/2}\mathbf{N}_2\Sigma_{YY}^{1/2}. \tag{1.11}$$

en donde $\mathbf{N}_1 = \mathbf{K}\mathbf{K}'$ y $\mathbf{N}_2 = \mathbf{K}'\mathbf{K}$. Note que $\mathbf{N}_1$ y $\mathbf{M}_1$ son matrices de orden $p \times p$ y las matrices $\mathbf{N}_2$ y $\mathbf{M}_2$ son matrices de orden $q \times q$. Del teorema A.1.1 se prueba que $\mathbf{N}_1, \mathbf{M}_1, \mathbf{N}_2$ y $\mathbf{M}_2$ tienen los mismos valores propios diferentes de cero. Como se demuestra abajo se tiene que $\mathbf{N}_1$ es definida positiva y tiene los mismos valores propios que $\mathbf{N}_2$ y éstos son positivos. Por las propiedades 6 y 7 en A.1.1 se tiene que $k = rank(\mathbf{K}) = rank(\Sigma_{XY})$, donde k denota el número de valores propios diferentes de cero. Por el teorema SVD, se tiene que:

$$\mathbf{K} = (\mathbf{s}_1, ..., \mathbf{s}_k)\mathbf{D}(\mathbf{t}_1, ..., \mathbf{t}_k)'.$$

donde $\mathbf{s}_i$ y $\mathbf{t}_i$ son vectores propios ortogonales estandarizados de $\mathbf{N}_1$ y $\mathbf{N}_2$, respectivamente, ya que

$$\mathbf{N}_1 = \mathbf{K}\mathbf{K}' = (\mathbf{s}_1, ..., \mathbf{s}_k)\mathbf{D}^2(\mathbf{s}_1, ..., \mathbf{s}_k)',$$

e igualmente para $\mathbf{N}_2$. Además, $\mathbf{K}$ tiene valores propios λ_i, contenidos en $\mathbf{D}$, donde los valores propios para $\mathbf{N}_1$ y $\mathbf{N}_2$, son λ_i^2. Se tiene el siguiente problema de valores y vectores propios,

$$\begin{aligned} \mathbf{N}_1\mathbf{s}_i &= \lambda_i^2\mathbf{s}_i, \\ \Sigma_{XX}^{-1/2}\mathbf{N}_1\Sigma_{XX}^{1/2}\Sigma_{XX}^{-1/2}\mathbf{s}_i &= \lambda_i^2\Sigma_{XX}^{-1/2}\mathbf{s}_i, \\ \mathbf{M}_1\mathbf{a}_i &= \lambda_i^2\mathbf{a}_i, \end{aligned}$$

en donde,

$$\mathbf{a}_i = \Sigma_{XX}^{-1/2}\mathbf{s}_i.$$

se deduce que $\mathbf{N}_1$ y $\mathbf{M}_1$ tienen los mismos valores propios y además los vectores propios de $\mathbf{N}_1$ y $\mathbf{M}_1$ están relacionadas por las ecuaciones anteriores. De igual forma se obtiene que $\mathbf{b}_i = \Sigma_{YY}^{-1/2}\mathbf{t}_i$. En general, se resolvieron dos problemas de valores y vectores propios de la forma $\mathbf{A}\theta = \rho\theta$ donde en uno $\mathbf{A} = \mathbf{M}_1$, $\rho = \lambda^2$ y $\theta_i = \mathbf{a}_i = \Sigma_{XX}^{-1/2}\mathbf{s}_i$ y para el otro, $\mathbf{A} = \mathbf{M}_2$, $\rho = \lambda^2$ y $\theta_i = \mathbf{b}_i = \Sigma_{XX}^{-1/2}\mathbf{t}_i$.

1.1.2. ACC usando descomposición de Cholesky

A continuación se presenta la propuesta de Hardoon *et al.* [9] que da solución a la ecuación (1.7) que representa un problema de valores y vectores propios generalizado y para ello se propone utilizar el método de descomposición de Cholesky (ver A.1.6) sobre las matrices de covarianzas para resolver un problema de valores y vectores propios simple.
Supóngase que Σ_{XX} es invertible, entonces dado que Σ_{XX} es simétrica y definida positiva, se puede descomponer, de acuerdo al Teorema de Cholesky, como

$$\Sigma_{XX} = \mathbf{R}_{XX}\mathbf{R}'_{XX}, \tag{1.12}$$

en donde $\mathbf{R}_{XX}$ es una matriz triangular inferior. Se define

$$\mathbf{w}_x = \mathbf{R}'_{XX}\mathbf{a} \quad \Rightarrow \quad \mathbf{R}_{XX}^{-1'}\mathbf{w}_x = \mathbf{a}. \tag{1.13}$$

Entonces sustituyendo (1.13) en (1.7) se obtiene

$$\Rightarrow \quad \mathbf{R}_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\mathbf{R}_{XX}^{-1'}\mathbf{w}_x = \lambda^2\mathbf{w}_x.$$

Similarmente se realiza el mismo procedimiento para determinar a $\mathbf{b}$ usando la ecuación (1.8). Por lo tanto, se resolvió un problema de valores y vectores propios simple de la forma $\mathbf{A}\theta = \rho\theta$ donde $\mathbf{A} = \mathbf{R}_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\mathbf{R}_{XX}^{-1'}$, $\rho = \lambda^2$ y $\theta = \mathbf{w}_x = \mathbf{R}'_{XX}\mathbf{a}$.

1.1.3. ACC usando factorización QR

Otra forma de obtener los valores y vectores propios es mediante el uso de un algoritmo de descomposición **QR**, presentado en Seber [38].
Como primer paso, usando el teorema de factorizaciónn QR (ver A.1.5) sobre las matrices de datos $\mathbf{X}$ y $\mathbf{Y}$ se tiene que

$$\begin{aligned} \mathbf{X} &= \mathbf{Q}_x\mathbf{T}_x \\ \mathbf{Y} &= \mathbf{Q}_y\mathbf{T}_y \end{aligned}$$

donde $\mathbf{Q}_x$ y $\mathbf{Q}_y$ son matrices ortogonales y $\mathbf{T}_x$ y $\mathbf{T}_y$ son matrices triangulares superiores.
Las matrices de varianzas y covarianzas se reescriben de la siguiente forma

$$\begin{aligned} \Sigma_{XX} &= \mathbf{T}'_x\mathbf{T}_x, \\ \Sigma_{YY} &= \mathbf{T}'_y\mathbf{T}_y, \\ \Sigma_{XY} &= \mathbf{T}'_x\mathbf{Q}'_x\mathbf{Q}_y\mathbf{T}_y. \end{aligned}$$

Se define

$$\mathbf{C} = \mathbf{T}_x^{\prime -1}\Sigma_{XY}(\mathbf{T}_y)^{-1}, \tag{1.14}$$

entonces (1.10) y (1.11) se pueden reescribir como sigue

$$\mathbf{M}_1 = \Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} = (\mathbf{T}_x)^{-1}\mathbf{C}\mathbf{C}'\mathbf{T}_x,$$

$$\mathbf{M}_2 = \Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} = (\mathbf{T}_y)^{-1}\mathbf{C}'\mathbf{C}\mathbf{T}_y.$$

Dado que $\mathbf{M}_1$ con $\mathbf{C}\mathbf{C}'$ y $\mathbf{M}_2$ con $\mathbf{C}'\mathbf{C}$ son transformaciones de similaridad, entonces tienen los mismos valores propios. De igual forma se observa que $\mathbf{M}_1$ con $\mathbf{C}\mathbf{C}'$ y $\mathbf{M}_2$ con $\mathbf{C}'\mathbf{C}$ tienen los mismos valores propios.
Ahora aplicando SVD a $\mathbf{C}$ y $\mathbf{C}'$ se tiene que

$$\bar{\mathbf{C}} \quad = \quad (\mathbf{f}_1, ..., \mathbf{f}_k)\mathbf{D}(\mathbf{g}_1, ..., \mathbf{g}_k)'$$

en donde $\mathbf{f}_i$ y $\mathbf{g}_i$ son vectores propios ortogonales de $\mathbf{C}$ y $\mathbf{C}'$ respectivamente, además,

$$\mathbf{C}\mathbf{C}' = (\mathbf{f}_1, ..., \mathbf{f}_k)\mathbf{D}^2(\mathbf{f}_1, ..., \mathbf{f}_k)',$$

e igualmente para $\mathbf{C}'\mathbf{C}$.
Como se puede observar, los valores propios de $\mathbf{C}'$ son los mismos que $\mathbf{C}$, además se tiene que los valores propios de $\bar{\mathbf{C}}\bar{\mathbf{C}}'$ y $\mathbf{C}'\mathbf{C}$ son los que se encuentran en la diagonal de la matriz $\mathbf{D}^2$. Dado que los vectores propios de $\mathbf{C}\mathbf{C}'$ son los representados por la matriz $\mathbf{F}$ que tiene a $\mathbf{f}_i$ como i-ésimo rnglón

$$\begin{aligned} \mathbf{C}\mathbf{C}'\mathbf{f}_i &= \lambda_i^2\mathbf{f}_i, \\ \mathbf{T}_x^{-1}\mathbf{C}\mathbf{C}'\mathbf{T}_x T_x^{-1}\mathbf{f}_i &= \lambda_i^2\mathbf{T}_x^{-1}\mathbf{f}_i, \\ \mathbf{M}_1\mathbf{a}_i &= \lambda_i^2\mathbf{a}_i, \end{aligned}$$

donde

$$\mathbf{a}_i = \mathbf{T}_x^{-1}\mathbf{f}_i.$$

1.1.4. Relación entre los métodos

Mediante los tres métodos descritos anteriormente se obtienen los valores de las correlaciones canónicas, sin embargo, no se presenta el mismo valor numérico en cada método, por lo tanto, es necesario determinar bajo que transformación se logra la uniformización de los resultados [24].
En el método de descomposición SVD se definió la matriz $\mathbf{M}_1$, esta ecuación puede ser reescrita en términos de la matriz A dada en el método de descomposición de Cholesky, esto es

$$\mathbf{M}_1 = \Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} = (\mathbf{R}'_{XX})^{-1}\mathbf{A}\mathbf{R}'_{XX}. \tag{1.15}$$

Que se puede ver como un problema de valores propios sustituyendo la ecuación (1.15) de la forma

$$\begin{aligned} & \mathbf{M}_1\mathbf{x} = \lambda\mathbf{x} \\ \Rightarrow \quad & (\mathbf{R}'_{XX})^{-1}\mathbf{A}\mathbf{R}'_{XX}\mathbf{x} = \lambda\mathbf{x} \\ \Rightarrow \quad & \mathbf{A}\mathbf{R}'_{XX}\mathbf{x} = \lambda\mathbf{R}'_{XX}\mathbf{x}, \end{aligned}$$

si se define

$$\mathbf{y} = \mathbf{R}'_{XX}\mathbf{x},$$

entonces a los datos obtenidos en la descomposición de Cholesky, se les aplica la transformación $\mathbf{y}$.
Se observa que para obtener las correlaciones y vectores canónicas se utilizó como datos de entrada las matrices de correlaciones para el método de Cholesky y el método de descomposición SVD, mientras que para el método de factorización QR, los datos de entrada fueron las matrices $\mathbf{X}$ y $\mathbf{Y}$. Posteriormente, para uniformizar los valores con los metodos de Cholesky y de descomposición SVD, al obtener los valores y vectores propios en el método de descomposición QR, se divide cada entrada a_{ij} sobre sus respectivas varianzas, es decir, se divide entre

$$\sqrt{var(x_i)}\sqrt{var(x_j)}. \tag{1.16}$$

1.1.5. Geometría del ACC

En el ACC clásico se quiere encontrar los vectores $\mathbf{a}$ y $\mathbf{b}$ tal que la ecuación (1.1) sea máxima bajo las restricciones (1.2). Esta ecuación puede ser reescrita de la siguiente manera

$$\rho = Corr(U_i, V_i) = \frac{\langle U_i, V_i \rangle}{\|U\| \, \|V\|}, \tag{1.17}$$

en donde $U_i = \mathbf{a}'_i\mathbf{X}$ y $V_i = \mathbf{b}'_i\mathbf{Y}$ son la i-ésima pareja de correlaciones canónicas y el número de ellas es igual al $r = min\{p, q\}$. En este caso la restricción es arbitraria en aspectos como la longitud de U_i y V_i y no afectan la correlación (1.17) mientras $\|U_i\|, \|V_i\| > 0$. Si se toma a $\Sigma_{XX} = \mathbf{X'X}$ y $\Sigma_{YY} = \mathbf{Y'Y}$. Analíticamente la maximización de (1.1) guía a los siguientes problemas de valores y vectores propios

$$(\mathbf{X'X})^{-1}\mathbf{X'Y}(\mathbf{Y'Y})^{-1}\mathbf{Y'X}\mathbf{a}_i = \lambda_i^2\mathbf{a}_i, \tag{1.18}$$

$$(\mathbf{Y'Y})^{-1}\mathbf{Y'X}(\mathbf{X'X})^{-1}\mathbf{X'Y}\mathbf{b}_i = \lambda_i^2\mathbf{b}_i, \tag{1.19}$$

describiendo a los vectores canónicos como vectores propios correspondientes a los primeros r valores propios diferentes de cero y menores que 1.
A continuación se verificará que las correlaciones canónicas representan relaciones de dependencia entre los subespacios generados por los dos conjuntos de variables [32]. Esta propiedad justifica que el ACC sea invariante ante reparametrizaciones.
Primero se describe un punto de vista de la geometría del espacio columna según Kuss *et al.* 2003 [16]. Examinando (1.17) encontramos que el coeficiente de correlación canónica $\lambda_i = Corr(U_i, V_i)$ es igual al coseno del ángulo entre las variables aleatorias U_i y V_i. Maximizar este coseno puede ser interpretado

como minimizar el ángulo entre U_i y V_i, el cual es equivalente a minimizar la distancia para variables de igual longitud

$$argmin \, \|U_i - V_i\| \tag{1.20}$$

sujeto a

$$\|U_i\| = \|V_i\| = 1. \tag{1.21}$$

De nuevo la aplicación de ortogonalidad con respecto a las parejas previamente encontradas. Sean $P_x = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ y $P_y = \mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'$ que denotan las proyecciones ortogonales sobre sus respectivos espacios columna. En vista de esas proyecciones, los problemas de valores y vectores propios en (1.18) y (1.19) dan una caracterización geométrica obvia de la solución

$$P_x P_y a_i = \lambda_i^2 \mathbf{a}_i \tag{1.22}$$
$$P_y P_x b_i = \lambda_i^2 \mathbf{b}_i \tag{1.23}$$

el espacio columna de esta pareja de variables canónicas es ilustrada en la Figura 1.1.

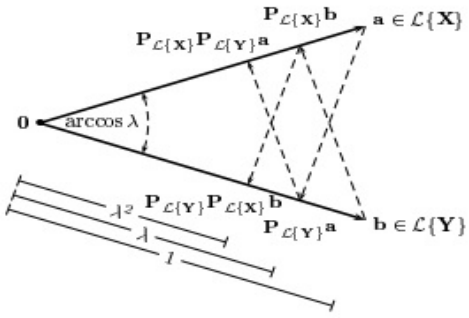


Figura 1.1: Gráfica de proyecciones de los datos[16]

1.2. Análisis de Correlación Canónica Regularizada

En la actualidad hay muchos trabajos de Estadística que utilizan la regularización en un amplio espectro de problemas. Este tema fue presentado por primera vez por Tikhonov en 1943, [42], en el contexto de la resolución de ecuaciones integrales mal planteadas, y se ha convertido en una parte importante de la Estadística. El uso de la regularización es debido a que en la mayoría de los

datos actuales, las características principales son el tamaño y complejidad [2]. El tamaño permite hacer estimaciones no paramétricas que son funciones inestables y discontinuas de la distribución subyacente de los datos, la densidad es un ejemplo típico. La complejidad de los datos, que por lo general corresponde a la alta dimensionalidad de las observaciones, hace intentar modelos muy complejos para ajustar los datos. El ajuste de los modelos con un gran número de parámetros también es inherentemente inestable. Ambas características obligan a usar regularización con el fin de obtener procedimientos más sensibles.
Por otro lado, la regularización puede usarse en el ajuste no paramétrico en diferentes formas, como lo es mediante suavizado kernel o penalización. También se usa en problemas de alta dimensión, en donde la dimensión d de los datos pueden ser del mismo orden o sustancialmente mayor que el tamaño de la muestra n [43].
Un caso específico es el ACC, que a veces no puede ser llevado a cabo cuando el número de individuos es menor al número de variables, es decir, $n \leq max(p, q)$. Esto puede llevar a situaciones en las que existen muchas covariables que están altamente correlacionadas, este tipo de comportamiento se conoce como multicolinealidad. Un enfoque para controlar los efectos de multicolinealidad es agregar un término de penalización que controla la variabilidad de los vectores propios de las matrices de covarianza muestrales dentro de los conjuntos $\mathbf{X}$ e $\mathbf{Y}$[8].
Se considera un conjunto de n individuos, cada uno de ellos con d variables. En caso de existir una fuerte multicolinealidad entre las variables a continuación habrá un subconjunto $max(p, q)$ $(< d)$ tal que los valores propios tendrán valores relativamente grandes y los restantes valores propios $d - max(p, q)$ tendrán valores comparativamente pequeñas. Este tipo de comportamiento en los valores propios puede crear inestabilidades numéricas en las matrices de covarianza muestrales. Una forma para solucionar este problema es incluir un paso de regularización en los cálculos antes realizados para el ACC. Como primer paso, los vectores canónicos pueden ser reescritos como sigue

$$\begin{aligned} \mathbf{a} &= \frac{\Sigma_{XX}^{-1}\Sigma_{XY}\mathbf{b}}{\rho_{XY}}, \\ \mathbf{b} &= \frac{\Sigma_{YY}^{-1}\Sigma_{YX}\mathbf{a}}{\rho_{XY}}, \end{aligned}$$

en donde

$$\rho_{XY} = \mathbf{a}'\Sigma_{XY}\mathbf{b}.$$

Posteriormente, las restricciones antes dadas en el ACC estarán penalizadas y se representarán de la siguiente forma

$$\begin{aligned} \Sigma_{XX}^{*}(\lambda_X) &= \mathbf{a}'\Sigma_{XX}\mathbf{a} + \lambda_X\mathbf{I}_p \\ &= \mathbf{a}'\Sigma_{XX}\mathbf{a} + \lambda_X\mathbf{a}'\mathbf{a}, \\ \Sigma_{YY}^{*}(\lambda_Y) &= \mathbf{b}'\Sigma_{YY}\mathbf{b} + \lambda_Y\mathbf{I}_q \\ &= \mathbf{b}'\Sigma_{YY}\mathbf{b} + \lambda_Y\mathbf{b}'\mathbf{b}, \end{aligned}$$

y

$$\Sigma^*_{XX}(\lambda_X) = \Sigma^*_{YY}(\lambda_Y) = 1.$$

Nuevamente se requieren encontrar combinaciones lineales tales que la correlación sea máxima sujeta a las restricciones dadas en la ecuación anterior. Usando el método de los multiplicadores de Lagrange, se tiene que

$$\begin{aligned}\phi(\mathbf{a},\mathbf{b}) &= \mathbf{a}'\Sigma_{XY}\mathbf{b} - \frac{\lambda}{2}(\mathbf{a}'\Sigma_{XX}\mathbf{a} + \lambda_X\mathbf{a}'\mathbf{a} - 1) \\ &\quad -\frac{\mu}{2}(\mathbf{b}'\Sigma_{YY}\mathbf{b} + \lambda_Y\mathbf{b}'\mathbf{b} - 1). \end{aligned} \tag{1.24}$$

Derivando con respecto a $\mathbf{a}$ y $\mathbf{b}$, respectivamente e igualando a 0 se obtiene que

$$\begin{aligned}\frac{\partial\phi}{\partial\mathbf{a}} &= \Sigma_{XY}\mathbf{b} - \lambda(\Sigma_{XX}\mathbf{a} + \lambda_X\mathbf{a}) = 0 \\ \frac{\partial\phi}{\partial\mathbf{b}} &= \Sigma'_{XY}\mathbf{a} - \mu(\Sigma_{YY}\mathbf{b} + \lambda_Y\mathbf{b}) = 0. \end{aligned} \tag{1.25}$$

Multiplicando las ecuaciones anteriores por $\mathbf{a}'$ y $\mathbf{b}'$, respectivamente, se obtiene que

$$\begin{aligned}\mathbf{a}'\Sigma_{XY}\mathbf{b} - \lambda(\mathbf{a}'\Sigma_{XX}\mathbf{a} + \lambda_X\mathbf{a}'\mathbf{a}) &= 0 \\ \mathbf{b}'\Sigma'_{XY}\mathbf{a} - \mu(\mathbf{b}'\Sigma_{YY}\mathbf{b} + \lambda_Y\mathbf{b}'\mathbf{b}) &= 0\end{aligned}$$

de lo cual se sigue que $\lambda = \mu = \mathbf{a}'\Sigma_{XY}\mathbf{b} = \rho_{XY}$.
Asumiendo que $\mathbf{b}'\Sigma_{YY}\mathbf{b} + \lambda_Y\mathbf{I}_q$ es invertible tenemos que

$$\mathbf{b} = \frac{(\mathbf{b}'\Sigma_{YY}\mathbf{b} + \lambda_Y\mathbf{I}_q)^{-1}\Sigma_{YX}\mathbf{a}}{\rho_{XY}}.$$

Sustituyéndose en la primera derivada parcial y rearreglando los términos se obtiene el siguiente problema de valores propios generalizado

$$\Sigma_{XY}(\mathbf{b}'\Sigma_{YY}\mathbf{b} + \lambda_Y\mathbf{I}_q)^{-1}\Sigma_{YX}\mathbf{a} = \rho^2_{XY}(\mathbf{a}'\Sigma_{XX}\mathbf{a} + \lambda_X\mathbf{I}_p)\mathbf{a}. \tag{1.26}$$

Se realiza el mismo procedimiento para la otra ecuación, y se obtiene que

$$\Sigma_{YX}(\mathbf{a}'\Sigma_{XX}\mathbf{a} + \lambda_X\mathbf{I}_p)^{-1}\Sigma_{XY}\mathbf{b} = \rho^2_{XY}(\mathbf{b}'\Sigma_{YY}\mathbf{b} + \lambda_Y\mathbf{I}_q)\mathbf{b}.$$

Al igual que el ACC, la ecuación (1.26) se puede ver como un problema de vectores y valores propios generalizado, para determinar a ρ_{XY}, $\mathbf{a}$ y $\mathbf{b}$. Los valores λ_X y λ_Y son obtenidos mediante el método de validación cruzada, descrito a continuación.

1.2.1. Validación cruzada

Para poder determinar los términos de penalización λ_X y λ_Y, utilizados en el ACCR se propone usar el método de validación cruzada [8].
Sea $\Lambda = (\lambda_X, \lambda_Y)$ un vector con entradas dadas por los escalares λ_X y λ_Y. Para algún valor dado de Λ, sea $\rho_\Lambda(-i)$ la primera correlación canónica del CCA, calculado de las unidades cuyas filas X_i y Y_i fueron removidas. Sean $\mathbf{a}_\Lambda^{-i}$ y $\mathbf{b}_\Lambda^{-i}$ los correspondientes vectores definidos como los primeros vectores canónicos. Se define cada uno de ellos para $i = 1, 2, ..., n$. Entonces el valor (numérico) de la validación cruzada dejando una pareja fuera para $\Lambda = (\lambda_X, \lambda_Y)$ es:

$$CV(\lambda_X, \lambda_Y) = Corr(\{X_i \mathbf{a}_\Lambda^{-i}\}_{i=1}^n, \{Y_i \mathbf{b}_\Lambda^{-i}\}_{i=1}^n).$$

Después, se elige el valor de λ_X y λ_Y que maximiza la correlación

$$\hat{\Lambda} = (\hat{\lambda}_X, \hat{\lambda}_Y) = argmax_{\lambda_X, \lambda_Y} CV(\lambda_X, \lambda_Y),$$

donde $\hat{\lambda}_X$ y $\hat{\lambda}_Y$ son elegidos con respecto a las primeras variables canónicas y se fijan para variables canónicas de orden mayor.

1.3. Análisis de Correlación Canónica Generalizada

En el ACC se tiene como objetivo determinar el grado de relación entre dos grupos de variables, sin embargo, en Estadística existen problemas en los que se requiere estudiar la relación entre varios bloques o grupos de variables, ésto para encontrar variables de un grupo, directamente relacionadas con las de otros grupos. Ejemplos para este análisis se encuentran en gran variedad de campos como la bioinformática, análisis sensorial, marketing, investigación de alimentos, entre otros. El análisis realizado para este tipo de datos es llamado Análisis de Correlación Canónica Generalizada (ACCG), y al igual que en el ACC clásico se busca determinar las variables canónicas que determinen la correlación canónica más grande.
El ACCG fue primeramente trabajado por Horst, quién introdujo dos métodos de solución del ACCG, métodos que posteriormente trabajó Kettenring en [15] junto con tres métodos más. Estas cinco técnicas para el análisis de muchos grupos son las siguientes:

- Método de la suma de las correlaciones (SUMCOR).
- Método de la varianza máxima (MAXVAR).
- Método de la suma de las correlaciones cuadradas (SSQCOR).
- Método de la varianza mínima (MINVAR).
- Método de la varianza generalizada (GENVAR).

En donde cada uno está diseñado para detectar una forma diferente de relación lineal entre los conjuntos y entonces es conveniente, especialmente en estudios exploratorios, emplear más de uno y tal vez todos estos métodos.
Una propuesta más reciente, es la presentada por Tenenhaus *et al.* en 2011, [41], y lo que hacen es combinar los métodos de análisis de datos multi-bloques y la flexibilidad de los algoritmos de mínimos cuadrados parciales (PLS), ya que en ambos métodos se trata con el mismo tipo de datos y se quieren relacionar varios bloques de variables observadas en el mismo conjunto de individuos. El poder de análisis de datos de múltiples bloques está en el hecho de que incluye una gran variedad de métodos con criterios bien identificados para ser optimizados. La gran flexibilidad del modelado ruta PLS radica en la posibilidad de tener en cuenta ciertas hipótesis sobre las conexiones entre bloques: el investigador decide qué bloques están conectados y los que no lo están. Desafortunadamente, los criterios optimizados por las diversas opciones de algoritmos de modelado de ruta PLS son a menudo poco claras. Este método es descrito a continuación y es parte del desarrollo de Tenenhaus *et al.* [41].

1.3.1. Análisis de Correlación Canónica Generalizada

Considere J vectores columna aleatorios con media cero y p_j-dimensionales $\mathbf{x}_j = (x_{j1}, ..., x_{jp_j})'$ definidos en la misma población y J vectores columna no aleatorios p_j-dimensional $\alpha_j = (x_{j1}, ..., x_{lp_j})'$. Considere una red de conexión entre los vectores aleatorios que definen una matriz de diseño, $\mathbf{C} = \{c_{jk}\}$ en donde $c_{jk} = 1$ si dos grupos están conectados y 0 en el otro caso, está matriz describe las relaciones existentes entre los grupos.
Considere dos componentes lineales $\eta_j = \sum_h \alpha_{jh} x_{jh} = \alpha_j' \mathbf{x}_j$ y $\eta_k = \sum_h \alpha_{kh} x_{kh} = \alpha_k' \mathbf{x}_k$. La correlación entre las dos variables aleatorias η_i y η_k es

$$\rho(\alpha_j' \mathbf{x}_j, \alpha_k' \mathbf{x}_k) = \frac{\alpha_j' \Sigma_{jj} \alpha_{\mathbf{j}}}{(\alpha_j' \Sigma_{jj} \alpha_{\mathbf{j}})^{1/2} (\alpha_j' \Sigma_{jj} \alpha_{\mathbf{j}})^{1/2}}, \tag{1.27}$$

en donde $\Sigma_{jj} = E(\mathbf{x}_j \mathbf{x}_j')$, $\Sigma_{kk} = E(\mathbf{x}_k \mathbf{x}_k')$ y $\Sigma_{jk} = E(\mathbf{x}_j \mathbf{x}_k')$. Todas las Σ_{jj} se suponen de rango completo.
Entonces en el Análisis de Correlación Canónica Generalizada se tiene como objetivo maximizar

$$\sum_{j,k=1, j \neq k}^{J} cjkg(\rho(\alpha_j' \mathbf{x}_j, \alpha_k' \mathbf{x}_k)), \tag{1.28}$$

sujeto a las restricciones

$$Var(\alpha_j' \mathbf{x}_j) = 1, \qquad j = 1, ..., J, \tag{1.29}$$

en donde g es la función identidad, la función valor absoluto o la función cuadrática. En terminología PLS y cuando se toma a g como la función identidad se denomina el esquema de Horst (propuesto por Kramer, 2007, citado por [41]);

cuando g es el valor absoluto se trabaja con el esquema centroide (introducido por Wold, 1985 [41]); y el esquema factorial dado cuando g es una función cuadrática (propuesto por Lohmoller, 1989 [41]). El problema (1.28) es equivalente a maximizar

$$\sum_{j,k=1,j\neq k}^{J} cjkg(\alpha_j'\mathbf{\Sigma}_{jk},\alpha_k) \tag{1.30}$$

sujeto a las restricciones

$$\alpha_j'\mathbf{\Sigma}_{jj},\alpha_j = 1, \qquad j = 1,...,J. \tag{1.31}$$

Al igual que como en el ACC y ACCR, el ACCG se ve como un problema de máximos con reestricciones. Nuevamente, mediante el método de multiplicadores de Lagrange se le da solución.

$$F(\alpha_{\mathbf{1}},...,\alpha_{\mathbf{J}},\lambda_1,...,\lambda_J) = \sum_{j,k=1,j\neq k}^{J} cjkg(\alpha_j'\mathbf{\Sigma}_{jk},\alpha_k) - \varphi\sum_{j=1}^{J}\frac{\lambda_j}{2}(\alpha_j'\mathbf{\Sigma}_{jj},\alpha_j - 1)$$

donde $\lambda_1,...,\lambda_J$ son multiplicadores de Lagrange y $\varphi = 1$ cuando g es la identidad o el valor absoluto y $\varphi = 2$ cuando g es la función cuadrática. Si tomamos derivadas con respecto a cada variable e igualando a cero, y además, tomando a g como el valor absoluto y considerando su derivada, se obtienen las siguientes ecuaciones estacionarias

$$\frac{1}{\varphi}\Sigma_{jj}^{-1}\sum_{k=1,k\neq j}^{J} cjkg(\alpha_j'\mathbf{\Sigma}_{jk}\alpha_k)\mathbf{\Sigma}_{jk}\alpha_k = \lambda_j\alpha_{\mathbf{j}}, \qquad j = 1,...,J$$

con la normalización de las restricciones

$$\alpha_j'\mathbf{\Sigma}_{jj},\alpha_j = 1, \qquad j = 1,...,J. \tag{1.33}$$

Sin embargo, estas ecuaciones no tienen solución analítica, pero pueden ser usadas para construir un algoritmo monótonamente convergente para el problema de optimización (1.30).

Considere J grupos de variables, $\mathbf{X}_1,...,\mathbf{X}_J$, en un conjunto de n individuos. Una fila de $\mathbf{X}_j$ representa una realización del vector fila aleatorio $\mathbf{x}_j^i$, una columna $\mathbf{x}_{jh}$ de $\mathbf{X}_j$ es considerado como una variable observada en n individuos, x_{jhi} es el valor de la variable de $\mathbf{x}_{jh}$ para el individuo i.

Sea $\mathbf{C} = \{c_{jk}\}$ donde $c_{jk} = 1$ si dos grupos están conectados y 0 en el otro caso, $\mathbf{C}$ es llamada matriz de diseño, está matriz describe las relaciones existentes entre los grupos.

Para encontrar una mejor estimación de la matriz de covarianza $\mathbf{\Sigma}_{jj}$ es necesario considerar la clase de combinaciones lineales

$$\left\{\hat{\mathbf{S}}_{jj} = \tau_j\mathbf{I} + (1-\tau_j)\mathbf{S}_{jj}, 0 \leq \tau_j \leq 1\right\},$$

de la matriz identidad $\mathbf{I}$ y la matriz de covarianza muestral $\mathbf{S}_{jj}$. $\hat{\mathbf{S}}_{jj}$ es denominada estimación de la contracción de $\mathbf{\Sigma}_{jj}$ y τ_j es la constante de contracción, donde

$$\hat{\tau}_j = \frac{\sum_{k\neq l=1}^{p_j} Var(s'_{j,kl}) + \sum_{k=1}^{p_j} Var(s'_{j,kk})}{\sum_{k\neq l=1}^{p_j} (s'_{j,kl})^2 + \sum_{k=1}^{p_j} (s'_{j,kl})^2},$$

aquí $s'_{j,kl}$ es una entrada de $\mathbf{S}'_{jj} = \frac{n}{n-1}\mathbf{S}_{jj}$. Para cada grupo se presenta un vector de peso exterior $\mathbf{a}_j$, una componente exterior $\mathbf{y}_j = \mathbf{X}_j\mathbf{a}_j$ y una componente interior definida como sigue

$$\mathbf{z}_j = \sum_{k=1,k\neq j}^{J} c_{jk} w(Cov(\mathbf{y}_j, \mathbf{y}_k))\mathbf{y}_k,$$

en donde la función $w(x)$ será igual a 1 (llamado esquema de Horst), x (esquema factorial) o $sign(x)$ (esquema centroide).
Además, sea g una función (identidad, cuadrada o valor absoluto para el esquema de Horst, factorial o centroide, respectivamente). Y por último, sean $\tau_1, ..., \tau_J$ constantes de contracción. Entonces se quiere maximizar:

$$\sum_{j,k=1,j\neq k}^{J} c_{jk} g(Cov(X_j\mathbf{a_j}, X_k\mathbf{a}_k)), \tag{1.34}$$

sujeto a las restricciones

$$\tau_j \left\|\mathbf{a}_j\right\|^2 + (1-\tau_j)Var(X_j\mathbf{a}_j) = 1$$

para $j = 1, ..., J$. Mediante esta maximización obtendremos las ecuaciones $\mathbf{a}_1, ..., \mathbf{a}_J$ donde

$$\mathbf{a}_j = \left[\mathbf{z}'_j\mathbf{X}_j\left[\tau_j\mathbf{I} + (1-\tau_j)\frac{1}{n}\mathbf{X}'_j\mathbf{X}_j\right]^{-1}\mathbf{X}'_j\mathbf{z}_j\right]^{\frac{-1}{2}}\left[\tau_j\mathbf{I} + (1-\tau_j)\frac{1}{n}\mathbf{X}'_j\mathbf{X}_j\right]^{-1}\mathbf{X}'_j\mathbf{z}_j \tag{1.35}$$

para $j = 1, ..., J$.

1.3.2. Análisis de Correlación Canónica Regularizada Generalizada

En el Análisis de Correlación Canónica Regularizada Generalizada se quieren encontrar las relaciones lineales entre varios bloques o grupos de variables, observadas en el mismo conjunto de individuos. Se quieren encontrar combinaciones lineales de las variables del grupo, tales que estas variables que se supone son conectadas, están altamente correlacionadas.
Considere J grupos de variables, $\mathbf{X}_1, ..., \mathbf{X}_J$, la matriz de diseño $\mathbf{C} = \{c_{jk}\}$, la función g y $\tau_1, ..., \tau_J$ constantes de contracción. Entonces se quiere maximizar:

$$\sum_{j,k=1,j\neq k}^{J} c_{jk} g(Cov(X_j\mathbf{a_j}, X_k\mathbf{a}_k)),$$

sujeto a las restricciones

$$\tau_j \|\mathbf{a}_j\|^2 + (1-\tau_j)Var(X_j\mathbf{a}_j) = 1,$$

para $j = 1, ..., J$. Mediante esta maximización obtendremos las ecuaciones $\mathbf{a}_1, ..., \mathbf{a}_J$. Las ecuaciones estacionarias (1.34) obtenidas anulando las derivadas de la función de Lagrange relacionadas con el problema de optimización son exactamente las ecuaciones estacionarias (1.35).

1.4. Aplicaciones del ACC y su extensiones a datos de manglares

En las siguientes subsecciones se presentan algunas aplicaciones de los métodos introducidos en las secciones anteriores. Principalmente se ejemplifican los métodos usando una base de datos de información de manglares. En la primera subsección se realiza un aplicación del ACC a los datos de manglar y se hace uso del paquete CCA en el software R, de la misma forma se trabaja en la segunda subsección con la diferencia de que en los dos grupos de variables el número de individuos es menor que el número de variables, se usaron los datos de cada laguna . En la tercera subsección se usó nuevamente una base de datos de manglares pero divido en tres grupos de variables y con un número de individuos menor al número de variables, razón por la cuál se aplicó el método ACCRG a estos datos y esto se hizo mediante el paquete RGCCA en el software R [33]. Esta ejemplificación ayudará al entendimiento de la aplicación de los métodos, a la interpretación de los resultados y a presentar las herramientas de R que estan disponibles para trabajar este análisis.

Datos de manglares

El sistema lagunar de Chacahua-Pastorías, en el estado de Oaxaca, fue declarado como Parque Nacional el 09 de julio de 1937 y es relevante resaltar que la información sobre la ecología, grado de conservación o impacto que presenta los ecosistemas de manglar para esta laguna, han sido muy poco estudiados [4]. Es de gran importancia realizar estudios sobre los parámetros fisicoquímicos del agua intersticial y las variables biológicas en los bosques de manglar que bordean los sistemas lagunares de Chacahua, Chacahua-Zapotalito, Chacahua-Corral, Pastoría, Salina y Miniyua.

1.4.1. Aplicación del ACC a los datos de manglares

La base de datos se dividió en dos grupos de variables (ver Cuadro), el primero correspondiente a la estructura forestal y el segundo grupo con los parámetros químicos y físicos del agua intersticial. Según la notación expresada previamente, estos dos grupos de variables constituyen las matrices $\mathbf{X}$ y $\mathbf{Y}$, el primero con 5 variables observadas y el segundo con 8 variables observadas.

Primer grupo de variables (X)	Segundo grupo de variables (Y)
Crecimiento del diámetro	Nitrato (NO3)
Hojarasca	Fosfato (PO4)
Densidad individual	Sulfato (SO4)
Área basal	Amonio (NH4)
Altura	Salinidad
	REDOX
	pH
	Temperatura

Cuadro 1.1: Variables de los datos de cabezas.

En el ámbito computacional se han desarrollado funciones para realizar aplicaciones del ACC, en el software R se encuentran la función cancor y el paquete CCA [8] que ayuda a solucionar el ACC. Los datos de manglares fueron procesados mediante estas herramientas.

En la Figura 1.2, se encuentra la representación de las correlaciones entre los dos conjuntos de variables, en donde en el primer cuadro se representan las correlaciones entre las variables del primer grupo, en el segundo están las correlaciones entre las variables del segundo grupo y en el tercero están las correlaciones entre las variables del primer grupo con el segundo grupo. Las correlaciones están representadas mediante colores, el color rojo oscuro representa correlación positiva entre dos variables y el color azúl representa correlación negativa.

En el Cuadro 1.2 se muestran las correlaciones canónicas de las parejas de combinaciones lineales posibles. Obsérvese que los tres primeros coeficientes de correlación canónica son altos y positivos, con lo que se comprueba la relación directamente proporcional entre la estructura forestal del bosque de manglar y el comportamiento de las propiedades fisicoquímicas del agua intersticial.

Puede apreciarse que en las combinaciones lineales U y V, correspondientes a la primera correlación canónica, se destaca el crecimiento del diámetro como importante en U, mientras que en V se presentan las variables PO4, NH4 y Temperatura como las más importantes y expresando una relación de oposición de NH4 con las otras dos. Estas mismas relaciones se presentan para las combinaciones lineales U y V correspondientes a la segunda correlación canónica. Las combinaciones lineales U y V correspondientes a la tercera correlación canónica destacan también la importancia de la variable Crecimiento del diámetro, aunque expresando una oposición con la Altura y tomando a NH4 y a la Temperatura como importantes pero no expresando oposición entre ellas.

Se evidencia que la producción de hojarasca, el valor del potencial REDOX, el grado de eutrofización del agua intersticial y la concentración de la salinidad son buenos estimadores del grado de conservación y vulnerabilidad actual y tendencial de los ecosistemas de mangle. Los altos valores de las correlaciones canónicas encontradas en el ACC, son resultados que concuerdan con otras investigaciones realizadas en el contexto de los bosques de manglar. En [4] se señala que la

Correlaciones Canónicas			
0.9987540	0.9807228	0.8183075	0.5872993	0.3674736
Variables Canónicas			
Primer Conjunto de Variables			
	[, 1]	[, 2]	[, 3]
CrecDia	4.176843e-01	1.100011e+00	-8.290888e-02
Hojarasca	-1.876847e-04	-5.330234e-05	-4.128376e-04
DenInd	3.298849e-06	1.807091e-06	2.401979e-05
Areabasal	2.749390e-03	-1.131958e-02	7.380853e-03
Altura	4.401872e-02	3.783937e-04	1.826498e-02
Segundo Conjunto de Variables			
	[, 1]	[, 2]	[, 3]
NO3	8.361735e-02	0.1151037816	0.2443736584
PO4	3.461938e-01	0.2337203548	0.2133681933
SO4	9.805263e-05	0.0002883568	-0.0005062320
NH4	-5.370533e-01	-0.6999402660	-0.4802956963
Salinidad	-1.407023e-02	-0.0077595335	0.0088343483
REDOX	7.078662e-05	0.0002160093	0.0005012678
pH	4.410312e-02	0.1058239001	0.0629053169
Temperatura	1.310780e-01	0.2459245303	-0.4066501369

Cuadro 1.2: Correlaciones y vectores canónicos obtenidos de los dos grupos de variables.

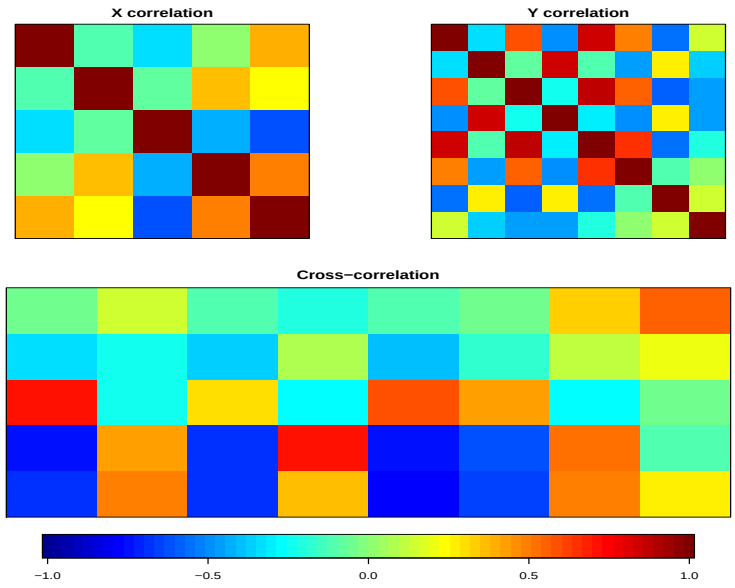


Figura 1.2: Correlaciones entre los dos grupos de variables.

disponibilidad de los nutrientes en el suelo y en el agua intersticial se ven reflejadas en la estructura forestal de los bosques de manglar. Además, se destaca que la distribución, composición y fisonomía de los bosques de manglar están determinadas por los cambios latitudinales en la temperatura y precipitación, aunque localmente depende de la fisiografía, geomorfología, sustrato, salinidad, nivel de inundación y relieve.

Posteriormente, se presenta la Figura 1.3 en la que se encuentran las correlaciones canónicas presentadas en el Cuadro 1.2.

En la Figura 1.4, se pueden observar las variables y las unidades experimentales para las dos primeras variables canónicas.

1.4.2. Aplicación del ACCR a los datos de manglares

Se aplicó el ACCR a dos grupos de variables extraídas del sistema lagunar de Chacahua-Pastorías para las lagunas Salina, Chacahua y Pastorías. Nuevamente se determinó el grado de relación entre las variables físicas de la estructura forestal de los bosques de manglar y las variables físicas y químicas del agua intersticial [4]. Estos dos grupos de variables son los conjuntos $\mathbf{X}$ y $\mathbf{Y}$, respectivamente. Se decidió utilizar el ACCR ya que el número de variables es mayor al número de individuos o mediciones por laguna.

Se realizó el análisis usando el paquete CCA en el lenguaje R [33].

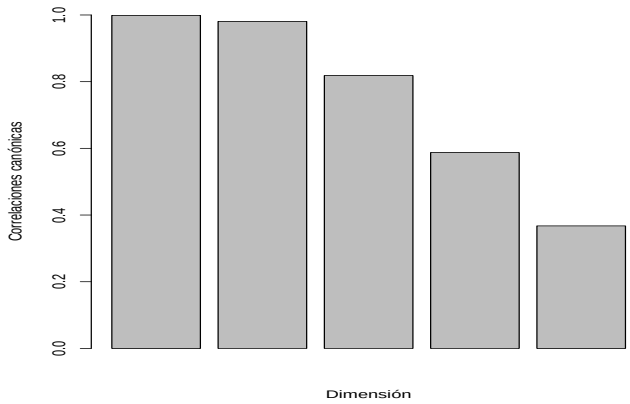


Figura 1.3: Correlaciones canónicas de los datos de manglares.

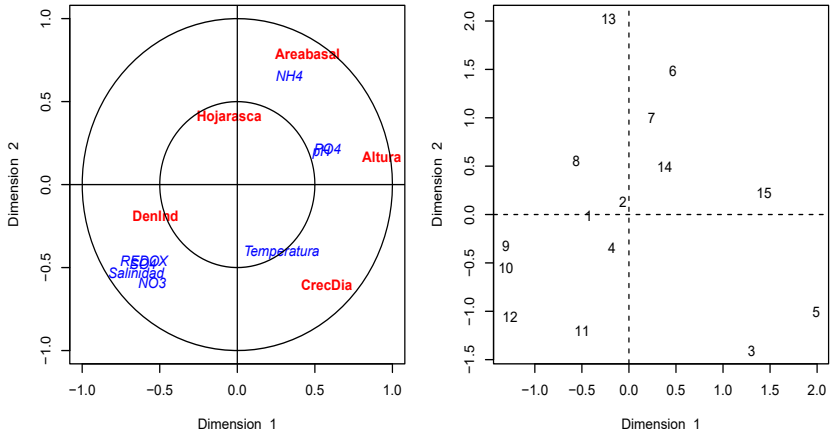


Figura 1.4: Primera y segunda variables canónicas.

Aplicación del ACCR en la laguna Salina

Se aplicó el ACCR en primera instancia a los datos de la Laguna Salina. En la Figura 1.5 se presentan las correlaciones entre los dos grupos de variables y se

observan correlaciones altas tanto negativas como positivas.

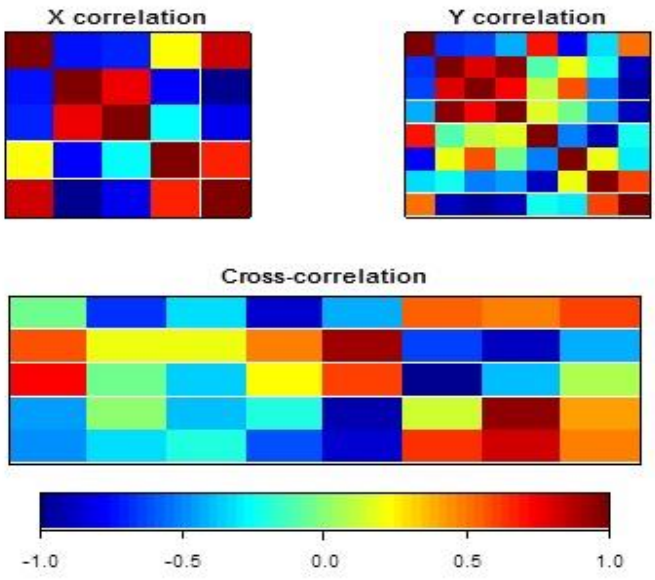


Figura 1.5: Correlaciones entre los dos grupos de variables

Como primer paso, se requiere determinar los valores λ_1 y λ_2, que son los parámetros de regularización. Usando el método de validación cruzada, se obtuvieron los siguientes valores para cada λ

$$\begin{aligned} \lambda_1 = 0{,}001, \qquad & \lambda_2 = 0{,}25075 \\ \mathbf{CV} - score \quad = \quad & 0{,}9617042, \end{aligned}$$

estos mismos valores están representados en la Figura 1.6, si se observa en la imagen un color muy bajo, este representará el valor óptimo para λ_1 y λ_2.

En el Cuadro 1.3 se presentan los valores para las correlaciones canónicas y los vectores canónicos obtenidos. Se presentan las tres primeras correlaciones canónicas, ya que fueron los valores más altos, así mismo se presentan los tres primeros vectores canónicos para cada grupo. Se observa que en la primera variable canónica, para el primer grupo, la variable representativa es el área basal mientras que en el segundo grupo es la variable REDOX. Por otro lado, en la segunda variable canónica para el primer grupo, nuevamente el área basal representa más a la nueva variable junto con la altura pero en sentido negativo; para el segundo grupo se tiene a la salinidad y REDOX. Como se puede ver, estó indica que algunas variables fisicas de la estructura forestal del bosque de manglar están relacionadas con las variables químicas del agua intersticial.

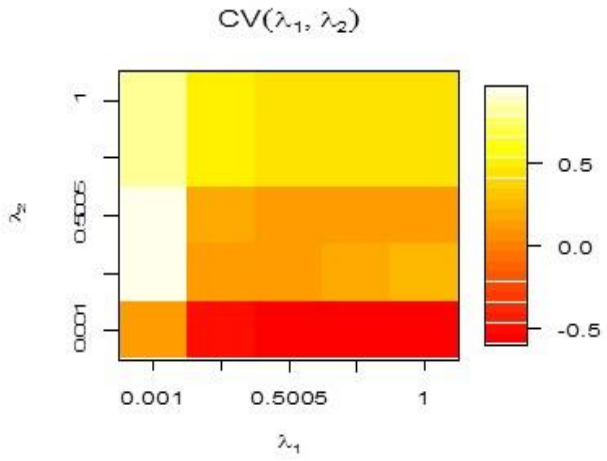


Figura 1.6: Valores λ de los términos de regularización para los datos de la laguna Salina

Aplicación del ACCR en la laguna Chacahua

Se realizó el análisis a datos obtenidos de la laguna Chacahua. Como primer paso se determinaron los valores λ_1 y λ_2. Usando el método de validación cruzada, se obtuvieron los siguientes valores para cada λ

$$\begin{aligned} \lambda_1 = 1, \qquad & \lambda_2 = 0{,}25075 \\ \mathbf{CV} - score \quad = \quad & 0{,}9994614, \end{aligned}$$

estos mismos valores están representados en la Figura 1.7, corroborándose que los valores λ_1 y λ_2 son los antes mencionados. Se pueden observar en el Cuadro 1.4, las correlaciones y vectores canónicos. Para el primer grupo, observando la primera variable canónica, se tiene que la variale representativa es la área basal y la altura pero en sentido negativo, mientras que para el segundo grupo es la salinidad con valor negativo. Para la segunda variable canónica, en el primer grupo, nuevamente el área basal es una variable que sobresale mientras que para el segundo grupo es la salinidad.

Aplicación del ACCR en la laguna Pastorías

Ahora se toma la base de datos obtenida de la laguna Pastorías y al igual que en el ejemplo anterior primero se determinan los valores λ_1 y λ_2. Usando el método de validación cruzada, se obtuvieron los siguientes valores para cada λ y estos

Correlaciones Canónicas			
0.9998863 0.9997461 0.9950637			
Variables Canónicas			
Primer conjunto de variables			
	$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$
CreDia	-2.6087e-02	-4.2066e-02	7.9140e-02
Hojarasca	6.5019e-03	1.0397e-02	-7.9777e-03
Densidad	-9.6672e-05	-6.3701e-05	5.7697e-05
área Basal	2.9963e-01	4.8316e-01	-9.0916e-01
Altura	-7.1219e-02	-1.1484e-01	2.1600e-01
Segundo conjunto de variables			
	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$
NO3	1.9030e-04	1.5004e-03	0.0183
PO4	-1.2944e-04	-1.0907e-04	-0.00299
SO4	4.3521e-04	1.1868e-03	-0.00095
NH4	-5.6928e-05	-8.6787e-06	-0.0009
Salinidad	1.4931e-03	1.5709e-02	0.18428
REDOX	2.2863e-02	-2.7965e-02	0.03813
pH	-3.8830e-05	-4.1876e-04	-0.00489
Temperatura	3.3305e-04	-1.4049e-04	0.00334

Cuadro 1.3: Correlaciones y vectores canónicos obtenidos para la laguna Salina.

están representados graficamente en la Figura 1.8.

$$\begin{aligned} \lambda_1 = 0{,}001, \qquad & \lambda_2 = 0{,}75025 \\ \mathbf{CV} - score \quad = \quad & 0{,}0214645, \end{aligned}$$

Se pueden observar en el Cuadro 1.5, las correlaciones y vectores canónicos. También se puede observar que solo las dos primeras correlaciones canónicas tienen valores altos, 0.9999985 y 0.9945238 respectivamente. En el primer grupo, observando la primera variable canónica se tiene que la hojarasca es la variable que más representa al grupo y en el segundo grupo es el sulfato con valor negativo. Para la segunda variable canónica se obtiene el mismo resultado agregando a la variable REDOX.

1.4.3. Aplicación del ACCRG a los datos de manglares

Se aplicó el ACCRG a la base de datos del ecosistema de manglar, del sistema lagunar de Chacahua - Pastorías, específicamente se trabajó con los datos de la laguna de Pastorías [25]. Se aplicó el ACCRG a tres grupos de variables, cada variable con 10 individuos: el primer grupo con las variables físicas (ramas, misceláneos y hojarasca) de la estructura forestal del bosque de manglar, el segundo grupo con las variables físico-químicas (Salinidad, Nitrato, Sulfato, Amonio, Fosfato, pH, REDOX y Temperatura) del agua intersticial y el tercer

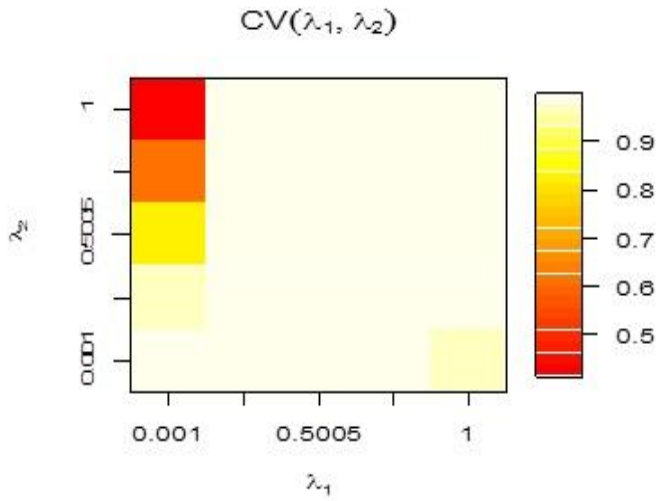


Figura 1.7: Valores λ de los términos de regularización para los datos de la laguna Chacahua

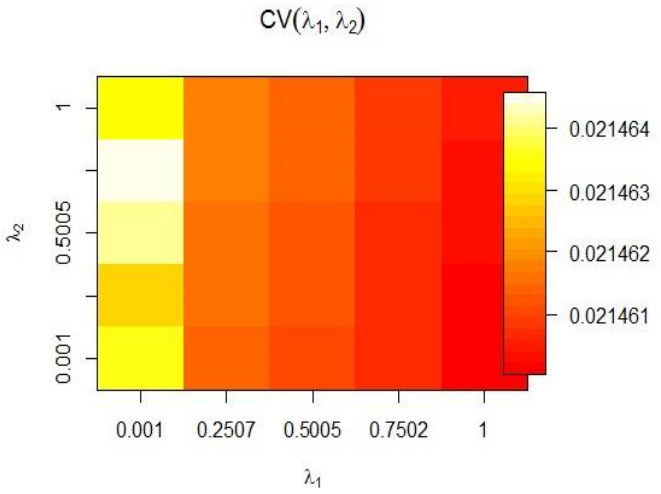


Figura 1.8: Valores λ de los términos de regularización para los datos de la laguna Pastorías

grupo con las variables de productividad primaria de los bosques de mangle (hojas, flores, ramas y estípulas). El objetivo es determinar el grado de relación

Correlaciones Canónicas			
0.9939351 0.9880264 0.9493930			
Variables Canónicas			
Primer conjunto de variables			
	$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$
CreDia	0.00032	-4.655e-03	-1.272e-02
Hojarasca	0.00086	1.2494e-03	-3.681e-04
Densidad	-0.00016	-9.184e-05	-1.342e-05
área Basal	0.07281	-1.030e-01	-4.277e-02
Altura	-0.0495	-1.483e-02	-2.242e-01
Segundo conjunto de variables			
	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$
NO3	-0.00831	-0.0232	-0.03799
PO4	0.01112	-0.00409	-0.15894
SO4	0.00385	0.00888	-0.00235
NH4	0.00467	0.00039	-0.05629
Salinidad	-0.12484	-0.20717	0.27309
REDOX	-0.02715	0.03271	0.01215
pH	0.01457	0.05431	0.15407
Temperatura	0.00724	-0.00558	-0.11609

Cuadro 1.4: Correlaciones y vectores canónicos obtenidos para la laguna Chacahua.

entre los grupos de variables.
Los datos fueron procesados con el paquete RGCCA en el lenguaje R, desarrollado por Tenenhaus [41] .
Se usó la siguiente matriz de diseño

$$\mathbf{C} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

Se obtuvieron los siguientes valores óptimos de las constantes de contracción

$$\tau_1 = 0{,}9124262 \qquad \tau_2 = 0{,}2359993$$
$$\tau_3 = 0{,}3842536.$$

Así también se obtuvieron los siguientes primeros vectores canónicos para cada grupo denotados por $\mathbf{a}_j$ para $j = 1, ..., J$, dados en (1.35). Estas nuevas variables son presentadas en el Cuadro 1.6.

Las $\mathbf{a}_i$ para $i = 1, 2, 3$, son los coeficientes estimados tales que maximizan (1.34). Se puede observar en la tabla que en $\mathbf{a}_1$ la variable PO4 con el valor $-0{,}691$ es

Correlaciones Canónicas			
0.9999985 0.9945238 2.104173e-13			
Variables Canónicas			
Primer conjunto de variables			
	$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$
CreDia	-1.9550e-07	-1.0147e-07	0.06939
Hojarasca	3.8022e-03	1.7490e-03	0.92218
Densidad	2.8134e-04	5.4124e-04	0.25119
área Basal	-1.4570e-04	-7.0767e-05	25.7918
Altura	9.8985e-06	4.1766e-06	18.2719
Segundo conjunto de variables			
	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$
NO3	1.041578e-07	0.00395	0.13048
PO4	-4.5308e-06	-0.02173	0.02753
SO4	-1.9983e-03	0.00215	0.02471
NH4	-1.1487e-06	-0.00682	-0.03627
Salinidad	-5.6110e-05	-0.07828	-0.85609
REDOX	-2.2503e-06	0.08861	-0.75438
pH	1.0477e-06	-0.00505	-0.00079
Temperatura	2.5278e-06	0.00368	-0.10572

Cuadro 1.5: Correlaciones y vectores canónicos obtenidos para la laguna Pastorías.

la más representativa del grupo así como salinidad y PH. En $\mathbf{a}_2$ se encuentran los frutos con el valor del coeficiente de 0,839 así como las hojas con 0,651. Finalmente, en $\mathbf{a}_3$ las ramas con el valor del coeficiente de 0,373 y los miscelaneos con 0,304.

En la Figura 1.9 se presentan las relaciones existentes entre las $\mathbf{Y}_i$, que son las combinaciones lineales que son obtenidas para cada grupo, una vez obtenidas las $\mathbf{a}_i$. Las combinaciones $\mathbf{Y}_2$ y $\mathbf{Y}_3$ presentan una relación lineal mas notable con respecto a las otras parejas. Esto indicaría que el grupo de las variables físico-químicas y el grupo de las variables de producción primaria presentan un grado de relación alto.

En general, en este capítulo se profundizó en el estudio de datos de manglares del sistema lagunar en el estudio se realizó un analisis usando ACC y sus extensiones. Mediante la aplicación del ACC a esta base de datos y con la ayuda del la función cancor y el paquete CCA en R se obtuvo que las variables químicas del agua intersticial están altamente correlacionadas con el grupo de variables físicas de la estructura forestal del bosque, lo que nos lleva a concluir que el desarrollo de los árboles de manglares depende de los nutrientes que pueda obtener del agua con la que se alimentan. Por otra parte, utilizando estas mismas técnicas se realizó un análisis para cada una de las lagunas de este sistema lagunar, utilizándose el ACCR debido al número de mediciones tomadas, ya que estas son

Vectores Canónicos					
$\mathbf{X}_1$	$\mathbf{a}_1$	$\mathbf{X}_2$	$\mathbf{a}_2$	$\mathbf{X}_3$	$\mathbf{a}_3$
NO3	0.233	Estípulas	0.228	Rama	0.373
PO4	-0.691	Hoja	0.651	Misceláneos	0.304
SO4	-0.242	Flor	-0.070	Hoj/dia	0.257
NH4	-0.282	Fruto	0.839	Hoj/mes	0.142
Salinidad	-0.383			Hoj/año	0.257
Redox	0.060				
PH	-0.340				
Temperatura	0.008				

Cuadro 1.6: Primer vector canónico para cada grupo de variables, dados en $\mathbf{a}_i$.

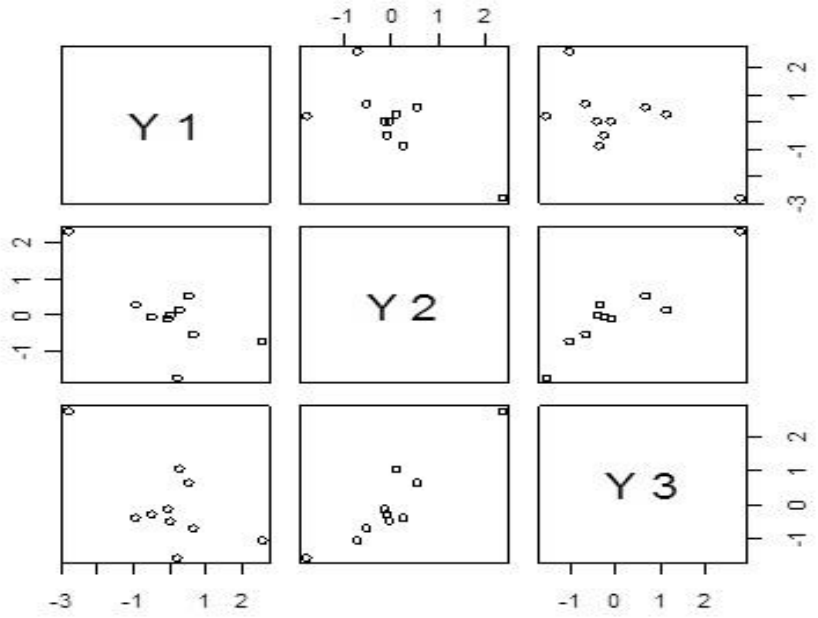


Figura 1.9: Gráfica de las $\mathbf{Y}_i$, que representan la combinación lineal para cada grupo.

menores al número de variables, y se obtuvieron resultados similares al análisis anterior. Posteriormente, se planteó el problema simultáneo de trabajar con una base de datos de manglares con mas de dos grupos de variables aplicándose el ACCRG [25], debido al número de individuos y de grupos. Lo resultante fue que sólo se observó relación entre las primeras variables canónicas del segundo grupo y el tercero, variables químicas y de producción primaria respectivamente, que lleva a concluir que estos nutrientes ayudan a la producción de hojarasca.

Capítulo 2

Análisis de Correlación Canónica con Kernel

Como se mencionó en el capítulo introductorio, no es hasta finales de los años 90's del siglo pasado que se desarrollaron procedimientos no lineales para las técnicas clásicas de la Estadística Multivariada. Una de las metodologías sugeridas ha sido la introducción del método kernel, que ha tenido gran influencia en todos los procedimientos multivariados y que, en particular, ha dado lugar a un nuevo método para el cálculo de las correlaciones canónicas. En esta tesis se tiene como objetivo principal abordar el problema del ACCK. En las siguientes secciones se presenta la teoría trabajada hasta el momento en diferentes referencias. Inicialmente se presentan las funciones kernel y algunas propiedades de ellas. Posteriormente, se define el ACCK y se presenta su método de solución basada en la propuesta de Hardoon et al. [9]. Se presenta además una forma diferente de trabajar con las funciones kernel a través de los métodos HSIC y KTA que son una alternativa para trabajar datos que no presentan correlación lineal.

2.1. Funciones kernel

Los métodos basados en kernel utilizan una aplicación implícita de los datos de entrada en un espacio de características de alta dimensión definido por una función kernel. De manera general, según Shawe-Taylor y Cristianini [39] mediante las funciones kernel se tienen los siguientes objetivos

- Los datos se encontrarán en un espacio vectorial que usualmente es llamado el espacio de características.
- En este nuevo espacio, se buscan las relaciones lineales entre las nuevas imágenes de los elementos de datos.

- Los algoritmos se aplican en una forma tal que no se necesitan las coordenadas de los puntos inmersos, sólo las parejas de productos internos.

- Las parejas de productos internos se pueden calcular de manera eficiente directamente de los elementos de datos originales utilizando una función kernel.

Lo anterior es presentado de manera gráfica en la figura 2.1.

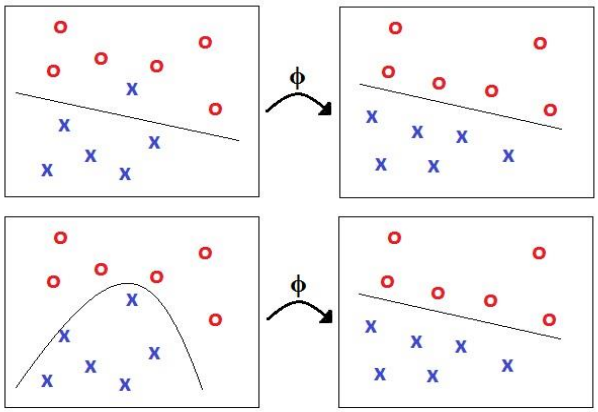


Figura 2.1: Función ϕ.

En ciertas aplicaciones se puede tener un conocimiento suficiente acerca del problema de tal manera que se puede diseñar un mapeo ϕ apropiado. Si esta asignación no es demasiado compleja y el espacio en el que se encuentran los datos no presenta dimensión tan alta, se podría simplemente aplicar de forma explícita esta asignación o mapeo a nuestros datos y con eso queda terminado. Algunas veces no se tiene conocimiento previo sobre como linealizar nuestro problema o el mapeo es difícil de calcular, ya sea en términos de complejidad computacional o en términos de requisitos de almacenamiento. Por ejemplo, considere un espacio de entradas de dos dimensiones con un espacio de características

$$\begin{aligned} \phi : \mathbb{R}^2 &\rightarrow \mathbb{R}^3 \\ (x_1, x_2) &\mapsto (x_1^2, \sqrt{2}x_1x_2, x_2^2). \end{aligned} \tag{2.1}$$

En donde se puede observar en la Figura 2.2, que para separarlo en dos clases se requiere un hiperplano lineal.
Sin embargo, lo que se puede hacer es calcular el producto interno en este espacio. La composición de la función característica con el producto interno en

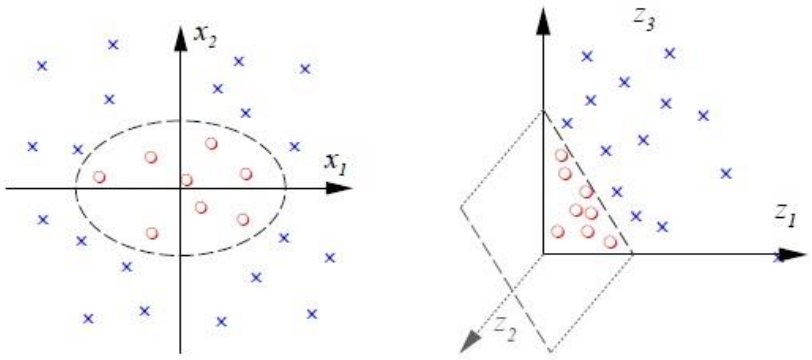


Figura 2.2: Inicialmente los datos no pueden ser separados linealmente, mediante una transformación se pueden separar usando un hiperplano lineal [29].

el espacio de características puede ser evaluada como sigue

$$\begin{aligned}\langle\phi(\mathbf{x}),\phi(\mathbf{z})\rangle &= \left\langle (x_1^2,\sqrt{2}x_1x_2,x_2^2),(z_1^2,\sqrt{2}z_1z_2,z_2^2)\right\rangle = x_1^2z_1^2+2x_1x_2z_1z_2+x_2^2z_2^2 \\ &= (x_1z_1+x_2z_2)^2 = \langle\mathbf{x},\mathbf{z}\rangle^2 .\end{aligned}$$

La solución anterior hace uso de los productos internos en el espacio de características. Los productos internos pueden, sin embargo, a veces ser calculados de manera más eficiente como una función directa de las características de entrada, sin calcular explícitamente el mapeo ϕ. En otras palabras, el paso de la representación característica-vector puede ser evitado, usando una función kernel.

Definición 2.1.1 *Un* ***kernel*** *es una función k tal que para toda $x, y \in \mathbf{X}$ satisface*

$$k(x,y) = \langle\phi(x),\phi(y)\rangle$$

donde ϕ es un mapeo de X a un espacio de características (producto interno o espacio de Hilbert) F

$$\phi : x \mapsto \phi(x) \in F.$$

Dada la definición anterior, se tienen las siguientes observaciones

- El llamado *truco del kernel* consiste en tomar el algoritmo original y formularlo de tal manera, que sólo se utilice $\phi(x)$ en los productos escalares o producto interno.
- Si se puede evaluar de manera eficiente estos productos escalares no es necesario llevar a cabo la asignación ϕ explícitamante y el problema puede resolverse aún en el espacio de características F.
- Mejor aún, no es necesario conocer la función ϕ, solo la función kernel.

A continuación se tiene la siguiente definición.

Definición 2.1.2 *Sea $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p \in \mathbf{X}$ un conjunto de observaciones, la matriz K de dimensión $p \times p$ con elementos*

$$K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$$

es llamada la ***matriz de Gram****.*

Hasta el momento sólo se tiene una manera de verificar que una función es un kernel, esto es, construyendo un espacio de características para las que la función corresponde a la primera representación de la asignación característica y luego calcular el producto interno entre las dos imágenes.
A continuación se introduce un método alternativo de demostrar que una función candidata es un kernel. Esto proporcionará una de las herramientas teóricas necesarias para crear nuevos kernel, y combinar los kernel antiguos para formar otros nuevos. Una de las observaciones más importantes es la relación con las matrices semidefinidas positivas, (ver definición en Apéndice A.1.16). Se tienen los siguientes resultados

Proposición 2.1.1 *Las matrices de Gram y kernel son semidefinidas positivas.*

Las demostraciones de los resultados presentados en este capítulo no se presentan pero pueden ser consultadas en [39]. Dados los resultados anteriores, se tiene el siguiente teorema de caracterización de funciones kernel.

Teorema 2.1.1 *Una función $k : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$ que es continua o tiene un dominio finito, puede descomponerse como $k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$ en un mapeo de características ϕ en un espacio de Hilbert F si y sólo si satisface la propiedad de ser finita semidefinida positiva.*

En [29], para k una función kernel semidefinida positiva sobre un conjunto X, se define el espacio de características $H = \{f | f : \mathbf{X} \to \mathbb{R}\}$ y el mapeo $\phi : \mathbf{X} \to \mathbb{R}^X$ como

$$\phi(x) = k(\cdot, x).$$

Se puede probar que el conjunto de todas las combinaciones lineales de la forma

$$f(\cdot) = \sum_{i=1}^{p} \alpha_i k(\cdot, x_i),$$

para p arbitrario y para $\alpha_i \in \mathbb{R}$ arbitrario forman un espacio vectorial. Especialmente para todas las funciones de la forma anterior se tiene que

$$f(x) = \langle k(\cdot, x), f \rangle_F ,$$

en donde $\langle \cdot, \cdot \rangle_F$ denota el producto interno en algún espacio de Hilbert. En particular se tiene que

$$\langle k(\cdot, \mathbf{x}), k(\cdot, \mathbf{z}) \rangle_F = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle_H = k(\mathbf{x}, \mathbf{z}).$$

Definición 2.1.3 *Sea* $\mathbf{X}$ *un conjunto no vacío, y* F *un espacio de Hilbert de funciones* $f : \mathbf{X} \to \mathbb{R}$. *Entonces* F *es llamado el* ***espacio de Hilbert reproductor de kernel*** *dotado con el producto interno* $\langle \cdot, \cdot \rangle$ *si existe un funcion* $k : \mathbf{X} \to \mathbb{R}$ *con las propiedades siguientes*

- k *tiene la propiedad de reproducción,* $\langle f, k(\cdot, \mathbf{x}) \rangle = f(\mathbf{x})$ *para toda* $f \in F$, *en particular,* $\langle k(\cdot, \mathbf{x}), k(\cdot, \mathbf{z}) \rangle = k(\mathbf{x}, \mathbf{z})$
- k *se extiende por H, es decir, H=* $\overline{span\,\{k(\cdot, \mathbf{x}) | \mathbf{x} \in X\}}$

Para las diferentes funciones kernel se tiene la siguiente proposición.

Proposición 2.1.2 *Sea* k_1 *y* k_2 *kernel sobre* $X \times X$, $X \subseteq \mathbb{R}^n$, $a \in \mathbb{R}^+$, $f(\cdot)$ *una función de valor real en* X

$$\phi : X \to \mathbb{R}^m$$

con k_3 *un kernel sobre* $\mathbb{R}^m \times \mathbb{R}^m$, *y* $\mathbf{B}$ *una matriz semidefinida positiva simétrica de orden* $n \times n$. *Entonces las siguientes funciones son kernel*

1. $k(\mathbf{x}, \mathbf{z}) = k_1(\mathbf{x}, \mathbf{z}) + k_2(\mathbf{x}, \mathbf{z})$,
2. $k(\mathbf{x}, \mathbf{z}) = ak_1(\mathbf{x}, \mathbf{z})$,
3. $k(\mathbf{x}, \mathbf{z}) = k_1(\mathbf{x}, \mathbf{z})k_2(\mathbf{x}, \mathbf{z})$,
4. $k(\mathbf{x}, \mathbf{z}) = f(\mathbf{x})f(\mathbf{z})$,
5. $k(\mathbf{x}, \mathbf{z}) = k_3(\phi(\mathbf{x}), \phi(\mathbf{z}))$,
6. $k(\mathbf{x}, \mathbf{z}) = \mathbf{x}'\mathbf{B}\mathbf{z}$.

Existen diferentes tipos de kernel [14], algunos de ellos son:

1. Lineal. $k(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$.
2. Gaussiano. $k(\mathbf{x}, \mathbf{y}) = exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right)$.
3. Polinomial. $k(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + c)^d$.
4. Tangente Hiperbólica. $k(\mathbf{x}, \mathbf{y}) = tanh(\mathbf{x} \cdot \mathbf{y} + c)$.

2.2. Análisis de Correlación Canónica con Kernel

En el Análisis de Correlación Canónica a veces no se pueden extraer descriptores útiles de los datos debido a que no presentan linealidad. El Análisis de Correlación Canónica con Kernel (ACCK) ofrece una solución alternativa por medio de la proyección de los datos en un espacio de características de alta dimensión.

Siguiendo el trabajo de Hardoon [9] se desarrolla este método como se describe a continuación.
Las matrices de varianzas y covarianzas pueden ser reescritas como las matrices de datos de $\mathbf{X}$ y $\mathbf{Y}$, es decir,

$$\mathbf{C}_{xx} = \mathbf{X}'\mathbf{X}$$
$$\mathbf{C}_{xy} = \mathbf{X}'\mathbf{Y}.$$

Las direcciones W_x y W_y pueden reescribirse como la proyección de los datos en las direcciones α y β

$$\mathbf{W}_x = \mathbf{X}'\alpha$$
$$\mathbf{W}_y = \mathbf{Y}'\beta.$$

Como en el caso del modelo lineal, el objetivo principal del ACCK, es encontrar los vectores canónicos en términos de los coeficientes α y β tal que

$$\rho = \max_{\alpha,\beta} \frac{\alpha'\mathbf{X}\mathbf{X}'\mathbf{Y}\mathbf{Y}'\beta}{\alpha'\mathbf{X}\mathbf{X}'\mathbf{X}\mathbf{X}'\alpha\beta'\mathbf{Y}\mathbf{Y}'\mathbf{Y}\mathbf{Y}'\beta}.$$

Sean $K_x = \mathbf{X}\mathbf{X}'$ y $K_y = \mathbf{Y}\mathbf{Y}'$ las matrices kernel (matriz de Gram) correspondientes a las dos representaciones. Entonces sustituyendo en la ecuación anterior se obtiene que

$$\rho = \max_{\alpha,\beta} \frac{\alpha' K_x K_y \beta}{\alpha' K_x^2 \alpha \beta' K_y^2 \beta}.$$

La ecuación anterior no se ve afectada por el cambio de escala de α y β, ya sea juntos o de forma independiente. Por lo tanto, el problema de optimización ACCK formulado en la ecuación antes descrita es equivalente a maximizar el numerador sujeto a

$$\begin{aligned} \alpha' K_x^2 \alpha &= 1 \\ \beta' K_y^2 \beta &= 1. \end{aligned}$$

2.2.1. Solución del ACCK mediante el formalismo de Lagrange

En la literatura se propone, para dar solución al ACC, usar el método de multiplicadores de Lagrange, ya que el ACC se puede ver como un problema de máximos con restricciones, que posteriormente se ve como un problema de valores y vectores propios generalizado.
Al igual que en el ACCC clásico, para encontrar los valores de α y β en el ACCK, se usa el método de multiplicadores de Lagrange, siguiendo la propuesta de Hardoon. Como primer paso se define el lagrangiano

$$L(\alpha, \beta, \lambda_\alpha, \lambda_\beta) = \alpha' K_x K_y \beta - \frac{\lambda_\alpha}{2}(\alpha' K_x^2 \alpha - 1) - \frac{\lambda_\beta}{2}(\beta' K_y^2 \beta - 1) \quad (2.2)$$

donde λ_α y λ_β son multiplicadores de Lagrange. Aplicando las condiciones de Karush-Kuhn-Tucker se obtiene que

$$\frac{\partial L}{\partial \alpha} = K_x K_y \beta - \lambda_\alpha K_x^2 \alpha = 0. \tag{2.3}$$

$$\frac{\partial L}{\partial \beta} = K_y K_x \alpha - \lambda_\beta K_y^2 \beta = 0. \tag{2.4}$$

$$\frac{\partial L}{\partial \lambda_\alpha} = -\frac{1}{2}(\alpha' K_x^2 \alpha - 1) = 0. \tag{2.5}$$

$$\frac{\partial L}{\partial \lambda_\beta} = -\frac{1}{2}(\beta' K_y^2 \beta - 1) = 0. \tag{2.6}$$

Las ecuaciones (2.5) y (2.6) recuperan las reestricciones. Multiplicando (2.3) por α' y (2.4) por β', se tiene que

$$\begin{aligned} \alpha' K_x K_y \beta - \lambda_\alpha \alpha' K_x^2 \alpha &= 0 \\ \beta' K_y K_x \alpha - \lambda_\beta \beta' K_y^2 \beta &= 0 \end{aligned}$$

entonces $\lambda_\alpha = \lambda_\beta = \lambda$. Considere el caso en donde las matrices kernel K_x y K_y son invertibles. Despejando a β de (2.4) se obtiene que

$$\begin{aligned} & \lambda K_y K_y \beta = K_y K_x \alpha \\ \Rightarrow \quad & \beta = \lambda^{-1} K_y^{-1} K_y^{-1} K_y K_x \alpha \\ \Rightarrow \quad & \beta = \lambda^{-1} K_y^{-1} K_x \alpha. \end{aligned}$$

Ahora sustituimos β en la ecuación (2.3), entonces se obtiene que

$$\begin{aligned} & K_x K_y (\lambda^{-1} K_y^{-1} K_x \alpha) - \lambda_\alpha K_x^2 \alpha = 0 \\ \Rightarrow \quad & \lambda^{-1} K_x K_y K_y^{-1} K_x \alpha = \lambda_\alpha K_x^2 \alpha \\ \Rightarrow \quad & K_x K_x \alpha = \lambda_\alpha^2 K_x^2 \alpha \\ \Rightarrow \quad & \mathbf{I}\alpha = \lambda_\alpha^2 \alpha. \end{aligned} \tag{2.7}$$

Este problema se puede ver como un problema de valores y vectores propios $\mathbf{Ex} = \rho\mathbf{x}$. De la ecuación (2.7) se puede deducir que $\lambda = 1$ para cada vector de α. De aquí, se puede elegir las proyecciones α como vectores unitarios j_i, para $i = 1..., m$, mientras que β son las columnas de $\frac{1}{\lambda} K_y^{-1} K_x$. Por lo tanto, cuando K_x o K_y son invertibles, se puede formar una correlación perfecta.
Por otro lado, un problema computacional que puede surgir al usar este método es el uso de grandes conjuntos de datos, ya que esto puede conducir a problemas de rango incompleto, es decir, que las matrices K_x o K_y sean no invertibles . Para superar este problema, se aplica ortogonalización parcial de Gram-Schmidt para reducir la dimensionalidad de las matrices del kernel.

Hasta ahora se ha considerado que las matrices del kernel son invertibles, aunque no necesariamente lo son. Se puede ver que el uso de grandes conjuntos de datos, puede conducir a problemas numéricos. A continuación se utiliza la ortogonalización parcial de Gram-Schmidt (PGSO) para aproximar las matrices kernel de tal manera que se pueda volver a representar a la correlación con dimensionalidad reducida.

Método de ortogonalización parcial de Gram-Schmidt

Sea K una matriz de orden $m \times n$ con $m \geq n$, sean $\mathbf{a}_i$ para $i = 1, ..., m$ los vectores columna de K. El objetivo en este método es encontrar vectores ortonormales, es decir, dada una suscesión de vectores linealmente independientes, mediante el método crea la base ortogonalizando cada vector a todos los vectores $\mathbf{a}_i$, donde $\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_m$ son tales vectores ortonormales. Para determinar estos vectores $\mathbf{q}_i$ se realiza el siguiente método

$$\begin{aligned}
\mathbf{q}_1 &= \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|_2} \text{ donde,} \mathbf{v}_1 = \mathbf{a}_1, \\
\mathbf{q}_2 &= \frac{\mathbf{v}_2}{\|\mathbf{v}_2\|_2} \text{ donde,} \mathbf{v}_2 = \mathbf{a}_2 - (q_1' \mathbf{a}_2)\mathbf{q}_1, \\
&\vdots \\
\mathbf{q}_i &= \frac{\mathbf{v}_i}{\|\mathbf{v}_i\|_2} \text{ donde,} \mathbf{v}_i = \mathbf{a}_i - \sum_{k=1}^{i-1} (q_k' \mathbf{a}_i)\mathbf{q}_k.
\end{aligned}$$

En donde los vectores $\mathbf{q}$ son ahora ortogonales. En particular, este método de ortogonalización es parcial ya que se utiliza un parámetro de precisión para detener el algoritmo cuando este criterio no se cumpla. El algoritmo de ortogonalización parcial de Gram-Schmidt en encuentra en el Apéndice C, Cuadro C.4 y cabe mencionar que mediante este método se obtendrá una matriz triangular inferior R. La matriz original K se puede escribir como $K = RR'$. Ahora, para verificar lo anteriormente descrito, mediante el programa MATLAB se obtuvo aleatoriamente una matriz de 30×70. En la Figura 2.3 se puede observar la matriz obtenida mediante el método PGSO, sin embargo, no se puede observar si es que esta matriz es triangular ya que los datos se encuentran dispersos. Mediante una permutación de filas, en la Figura 2.4 se puede observar que ahora se obtiene una matriz triangular. Se utilizaron diferentes valores de precisión para la misma matriz, sin embargo, se obtuvo la misma matriz triangular.

Análisis de correlacion canónica con Kernel usando PGSO

Como primer paso se descomponen las matrices kernel K_x y K_y mediante el método PGSO anteriormente descrito, obteniéndose así que

$$\begin{aligned}
K_x &= R_x R_x', \\
K_y &= R_y R_y'
\end{aligned}$$

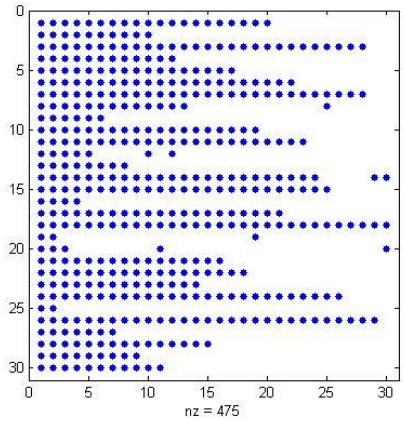


Figura 2.3: Matriz obtenida mediante el método PGSO.

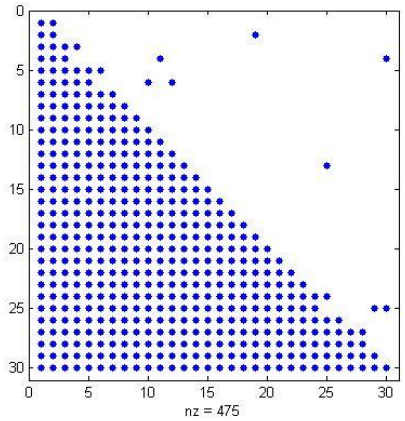


Figura 2.4: Matriz con las filas ordenadas.

endonde R_x y R_y son matrices triangulares inferiores.
Sustituyendo las ecuaciones anteriores en (2.3) y (2.4) se obtiene que

$$\frac{\partial L}{\partial \alpha} = R_x R_x' R_y R_y' \beta - \lambda R_x R_x' R_x R_x' \alpha = 0, \tag{2.8}$$

$$\frac{\partial L}{\partial \beta} = R_y R_y' R_x R_x' \alpha - \lambda R_y R_y' R_y R_y' \beta = 0, \tag{2.9}$$

multiplicando (2.8) por R'_x y (2.9) por R'_y

$$R'_x R_x R'_x R_y R'_y \beta - \lambda R'_x R_x R'_x R_x R'_x \alpha = 0, \tag{2.10}$$

$$R'_y R_y R'_y R_x R'_x \alpha - \lambda R'_y R_y R'_y R_y R'_y \beta = 0. \tag{2.11}$$

Se redefine como $\mathbf{Z}$ a la nueva matriz de correlaciones con la dimensión reducida, entonces las submatrices de correlaciones son:

$$\begin{aligned} Z_{xx} &= R'_x R_x, \\ Z_{yy} &= R'_y R_y, \\ Z_{xy} &= R'_x R_y, \\ Z_{yx} &= R'_y R_x. \end{aligned}$$

Defínase a $\tilde{\alpha}$ y $\tilde{\beta}$ como:

$$\tilde{\alpha} = R'_x \alpha, \tag{2.12}$$

$$\tilde{\beta} = R'_y \beta,$$

sustituyendo todas las ecuaciones anteriores en (2.10) y (2.11) se obtienen las ecuaciones

$$Z_{xx} Z_{xy} \tilde{\beta} - \lambda Z_{xx}^2 \tilde{\alpha} = 0, \tag{2.13}$$

$$Z_{yy} Z_{yx} \tilde{\alpha} - \lambda Z_{yy}^2 \tilde{\beta} = 0. \tag{2.14}$$

Supóngase que Z_{xx} y Z_{yy} son invertibles. Multiplicando (2.13) por Z_{xx}^{-1} y (2.14) por Z_{yy}^{-1} se obtiene que

$$Z_{xy} \tilde{\beta} - \lambda Z_{xx} \tilde{\alpha} = 0, \tag{2.15}$$

$$Z_{yx} \tilde{\alpha} - \lambda Z_{yy} \tilde{\beta} = 0. \tag{2.16}$$

Despejando a $\tilde{\beta}$ de (2.16) se tiene que

$$\tilde{\beta} = \frac{Z_{yy}^{-1} Z_{yx} \tilde{\alpha}}{\lambda}$$

y sustituyendo en (2.15) se obtiene

$$Z_{xy} Z_{yy}^{-1} Z_{yx} \tilde{\alpha} = \lambda^2 Z_{xx} \tilde{\alpha}. \tag{2.17}$$

Se tiene ahora un problema de valores y vectores propios generalizado, lo cual aún sigue impidiendo obtener la solución. Sea SS' igual a la descomposición

incompleta de Cholesky de Z_{xx} tal que $Z_{xx} = SS'$, donde S es una matriz triangular inferior, y sea

$$\hat{\alpha} = S'\tilde{\alpha}. \tag{2.18}$$

Sustituyendo en la ecuación (2.19) se tiene que

$$S^{-1}Z_{xy}Z_{yy}^{-1}Z_{yx}S^{-1\prime}\hat{\alpha} = \lambda^2\hat{\alpha}. \tag{2.19}$$

ésta ecuación de valores singulares resuelve el problema del análisis de correlación canónica con kernel.
Cabe mencionar que originalmente se quieren los valores de α y β. Para ello sabemos que de acuerdo a la ecuación (2.12)

$$\tilde{\alpha} \quad = \quad R_x'\alpha$$

y también de (2.18)

$$\hat{\alpha} = S'\tilde{\alpha}.$$

Despejando a α se obtiene que

$$(R_x')^{-1}(S')^{-1}\hat{\alpha} \quad = \quad \alpha,$$

en donde $\hat{\alpha}$ es un valor que se determinó mediante el problema de valores y vectores propios, el valor de S también es conocido, la matriz R_x es conocida también, sin embargo, es de rango incompleto y esto dificulta obtener la inversa de $(R_x')^{-1}$ y en consecuencia el valor de α, por lo que se sugiere tomar a

$$(R_x')^{-1} = R_x(R_x'R_x)^{-1}.$$

El algoritmo realizado para el método ACCK se presenta en el Apéndice C, Cuadro C.5. Además, se realizaron algoritmos para obtener las matrices kernel de los datos, usando los kernel lineal, polinomial, gaussiano y tangente hiperbólica.para obtener las correlaciones canónicas mediante el algoritmo ACCK.

2.2.2. Análisis de Correlación Canónica usando KTA

En el artículo de Chang et al. [5] se proponen dos criterios para encontrar asociaciones no lineales como solución al ACCK, estos son: el Criterio de Independencia de Hilbert-Schmidt (HSIC) y el Criterio de la Alineación Objetivo del Kernel Centrado (KTA). Éstos métodos motivan a superar las limitaciones del ACCK, como lo son, la falta de interpretabilidad debido a la transformación de los vectores de datos en un espacio de Hilbert abstracto y la incapacidad de eliminar rasgos irrelevantes desde las variables originales.
El primer método asociado con F_x y F_y, es el criterio de independencia de Hilbert-Scmidt (HSIC) que es es la norma Hilbert-Schmidt cuadrada del operador de covarianza cruzada entre el espacio de probabilidad X e Y. El HSIC se

puede utilizar como una medida de independencia cuando se asocia con algún kernel. En general, al igual que el ACCK, en el HSIC se tiene como objetivo calcular los vectores u y v tales que maximizan a

$$\rho_h = \frac{1}{(n-1)^2} tr(K^u \tilde{K}^v), \tag{2.20}$$

sujeto a $\|u\| = \|v\| = 1$, donde K^u es la matriz de Gram para los datos proyectados $u'x_i$ y $\tilde{K}^v$ la matriz de Gram centrada con ij-ésima entrada

$$\tilde{K}^v_{ij} = K^v_{ij} - \frac{1}{n}\sum_{i=1}^{n} K^v_{ij} - \frac{1}{n}\sum_{j=1}^{n} K^v_{ij} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K^v_{ij}. \tag{2.21}$$

Por otro lado, se define el método de la alineación objetivo del kernel centrado (KTA), que no es más que la versión normalizada del HSIC, y al igual que éste método, el KTA también es una medida de independencia. El método KTA al igual que el ACCK tiene como objetivo calcular los vectores u y v tal que maximizen

$$\rho_a = \frac{tr(K^u \tilde{K}^v)}{\sqrt{tr(K^u \tilde{K}^u) tr(K^v \tilde{K}^v)}}, \tag{2.22}$$

sujeto a $\|u\| = \|v\| = 1$.
Para resolver por los dos métodos anteriores en [5] se propone utilizar el método del gradiente descendente. A continuación se presenta el procedimiento usando el kernel Gaussiano. Se definen las funciones kernel como sigue:

$$\begin{aligned} k(x_i, x_j) &= exp(-\sigma_x \|x_i - x_j\|^2), \\ &\text{y} \\ k(y_i, y_j) &= exp(-\sigma_y \|y_i - y_j\|^2), \end{aligned} \tag{2.23}$$

y las funciones kernel para los datos proyectados son:

$$\begin{aligned} k^u(x_i, x_j) &= Exp(-\sigma_x \|u'(x_i - x_j)\|^2), \\ k^v(y_i, y_j) &= Exp(-\sigma_y \|v'(y_i - y_j)\|^2). \end{aligned} \tag{2.24}$$

Sin tomar en cuenta el término constante, los gradientes de (2.20) con respecto a u' y v' son

$$\begin{aligned} \frac{\partial \rho_h}{\partial u'} &= -2\sigma_x u' \sum_{i=1}^{n}\sum_{j=1}^{n} K^u_{ij} \tilde{K}^v_{ij} (x_i - x_j)(x_i - x_j)', \\ \frac{\partial \rho_h}{\partial v'} &= -2\sigma_y v' \sum_{i=1}^{n}\sum_{j=1}^{n} \tilde{K}^u_{ij} K^v_{ij} (y_i - y_j)(y_i - y_j)'. \end{aligned} \tag{2.25}$$

Por otro lado, para (2.22), se considerarán los gradientes $log(\rho_a)$. Los gradientes $log(\rho_a)$ con respecto a u' y v' están dados a continuación

$$\frac{\partial log(\rho_a)}{\partial u'} = u' \sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}^{u}(x_i - x_j)(x_i - x_j)',$$
$$\frac{\partial log(\rho_a)}{\partial v'} = v' \sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}^{v}(y_i - y_j)(y_i - y_j)'. \tag{2.26}$$

en donde

$$W_{ij}^{u} = -2\sigma_x K_{ij}^{u}\left(\frac{\tilde{K}_{ij}^{v}}{tr(K^u\tilde{K}^v)} - \frac{\tilde{K}_{ij}^{u}}{tr(K^u\tilde{K}^u)}\right)$$
$$W_{ij}^{v} = -2\sigma_y K_{ij}^{v}\left(\frac{\tilde{K}_{ij}^{u}}{tr(K^u\tilde{K}^v)} - \frac{\tilde{K}_{ij}^{v}}{tr(K^v\tilde{K}^v)}\right). \tag{2.27}$$

Posteriormente, se estimarán los valores de u y v mediante el algoritmo de gradiente descendente, con un gradiente modificado para asegurar la longitud unitaria de las restricciones.
El algoritmo implementado mediante este método se encuentra en el software R y el paquete lleva el nombre de hsicCCA.

2.3. Aplicación del ACCK y del ACC no lineal mediante KTA

A continuación se presentan aplicaciones a diferentes bases de datos. En la primera sección se realiza una comparación de las correlaciones canónicas obtenidas al usar diferentes valores de las constantes en los kernel usando los datos de manglar. Posteriormente se presentan aplicaciones a dos bases de datos que mediante el ACC clásico no se obtuvieron correlaciones altas. Una base de datos tiene información de características físicas de niños así el número de enfermedades y medicamentos que les fue recetado. La segunda base de datos es referente a información de variables psicológicas y academícas de alumnos de determinada escuela.

2.3.1. Aplicación del ACCK a los datos de manglar

Se aplicó el ACCK a dos grupos de variables extraídas del sistema de lagunar de Chacahua-Pastorías [26].
Usando los kernel de la primera sección se presenta el siguiente análisis

Kernel polinomial

Utilizando la función kernel polinomial se aplicó el análisis de correlación canónica con kernel al ejemplo. Se tomaron como constantes c los valores: 0, 10,

100, 1000, 10000, 1000000, -10, -100, -1000 y -1000000; y para los grados del polinomio los valores: 1, 2 y 3. Como una observación, cuando se toma $c = 0$ y $d = 1$, se trata del caso del kernel lineal. Los resultados son presentados en los Cuadros 2.1, 2.2 y 2.3.

d	1	1	1	1	1	1	1	1	1
c	0	10	100	1000	1000000	-10	-100	-1000	-100000
	0.2713	0.2565	0.3675	0.3675	0.3675	0.7248	0.6782	0.5393	0.4417
	0.7374	0.3956	0.5873	0.5873	0.5873	0.8107	0.7745	0.7842	0.9079
	0.9418	0.7478	0.8183	0.8183	0.8183	0.9646	0.9739	0.9283	
	0.9629	0.9494	0.9807	0.9807	0.9807	0.9985	0.9962		
	0.9987	0.9964	0.9988	0.9988	0.9988				
		1	1	1	1				

Cuadro 2.1: Valores propios.

d	2	2	2	2	2	2	2
c	0	10	1000	-10	-100	-1000	-100000
	1	1	1	0.4672	0.3766	0.5102	0.0032
	1	1	1	0.9265	0.8229	0.7162	0.0032
	1	1	1	1	0.981	0.8501	0.8276
	1	1	1	1	0.9943	0.9651	
	1	1	1	1	1		
	1	1	1	1	1		
	1	1	1	1			
	1	1	1				
	1	1	1				
	1	1	1				
	1	1	1				
	1	1	1				
	1	1	1				
	1	1	1				
	1	1	1				

Cuadro 2.2: Valores propios.

Al realizar este análisis se obtuvieron diferentes resultados.

- El número de valores propios es igual al rango de K_x (Ran(K_x)=Ran(R_x)).
- Cuando se tiene un valor par o impar para d, se logran obtener los valores propios.
- Para valores de c mayores que 10, los valores propios quedarán fijos, y serán los mismos.
- Para valores de c negativos, el número de valores propios diferentes disminuye, además, los valores de la correlación decrecen.

d	3	3	3	3	3
c	0	10	-10	-100	-1000000
	1	1	0.3108	0.4295	0.0032
	1	1	0.5569	0.6545	0.9369
	1	1	0.8137	0.9398	
	1	1	1	0.97	
	1	1	1	1	
	1	1	1	1	
	1	1	1	1	
	1	1	1		
	1	1	1		
	1	1			
	1	1			
	1	1			
	1	1			
	1	1			
	1	1			

Cuadro 2.3: Valores propios.

Kernel gaussiano

Se realizó el ACCK utilizando el ejemplo anterior, en este caso se tomaron como matrices kernel, las obtenidas mediante el kernel gaussiano, sin embargo, desde el planteamiento de esta función se puede observar que para obtener K_x y K_y algunos elementos se pueden hacer cero. Como resultado del ACCK se obtuvieron valores de 0 y 1 para los vectores propios α.

Kernel tangente hiperbólica

Para c con diferentes valores positivos y negativos, siempre se obtiene un solo valor propio con su valor igual a 1.
Dado que como se observa anteriormente, en el análisis del caso de maglares, los valores de las correlaciones canónicas son altos aplicando el ACC, así que no es necesario aplicar el ACCK. A continuación se presenta el caso en que mediante el ACC se obtienen valores de la correlación bajos y debido a eso se aplica ACCK

2.3.2. Aplicación a datos de interacciones medicamentosas

Se han realizado investigaciones sobre las reacciones adversas hacia ciertos medicamentos y su relación con la calidad de vida de los pacientes. En particular, en [18], [21], [27] se realizó un estudio para analizar las posibles causas de las reacciones medicamentosas en pacientes pediátricos con datos de un consultorio de la Ciudad de Puebla, explorando las relaciones existentes entre la edad, el peso y la talla de niños de 2 a 6 años con ciertas enfermedades que han padecido con los tipos y número de medicamentos que han ingerido.

Se cuenta con una base de datos tomados en un consultorio de la Ciudad de Puebla. La información obtenida es de 292 pacientes pediátricos y se requiere estudiar las relaciones entre los grupos de variables presentados en el Cuadro 2.4.

Primer grupo de variables (X)	Segundo grupo de variables (Y)
Edad	Número de enfermedades
Peso	Número de medicamentos
Talla	Número de fármacos

Cuadro 2.4: Variables del estudio de interacciones medicamentosas

Para determinar el grado de relación entre los dos grupos de variables, se realizó el ACC usando el paquete CCA del software R [8].
Como primera parte, se obtuvieron las correlaciones entre cada variable, éstas correlaciones son presentadas gráficamente en la Figura 2.5. Se observa que las correlaciones entre las variables del primer grupo son altas y positivas (están presentadas en el primer cuadro superior), por otro lado, se observa que la correlación entre las variables del segundo grupo es alta y positiva solo entre la variable número de medicamentos y número de fármacos. Y las correlaciones entre las variables del grupo $\mathbf{X}$ con el grupo $\mathbf{Y}$ son bajas.
Los resultados para el ACC clásico están tabulados Cuadro 2.5. El valor de la primer correlación canónica obtenida es 0.178, lo que indica que la relación entre los dos grupos de variables existe pero es baja, ésto es, podría no ser lineal. De igual forma, la segunda y tercer correlaciónes canónicas.

Correlaciones canónicas			
0.178119 0.114667 0.020926			
Variables canónicas			
Primer Conjunto de Variables			
	$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$
Edad	0.775292	1.471942	-0.459640
Peso	0.000157	-0.000302	-0.000270
Talla	-0.130479	0.019326	0.081394
Segundo Conjunto de Variables			
	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$
Enfermedades	-0.2419006	-1.529568	-1.197908
Medicamentos	0.3143961	-1.290641	1.413167
Fármacos	0.5213939	1.268991	-1.306821

Cuadro 2.5: Correlaciones y vectores canónicos obtenidos para los datos de medicamentos.

En la Figura 2.6 se encuentran las gráficas de las variables canónicas (primera y segunda) proyectadas en los datos. Se observa que los datos se agrupan en lineas horizontales, pero no se observa alguna relación lineal entre las proyecciones.

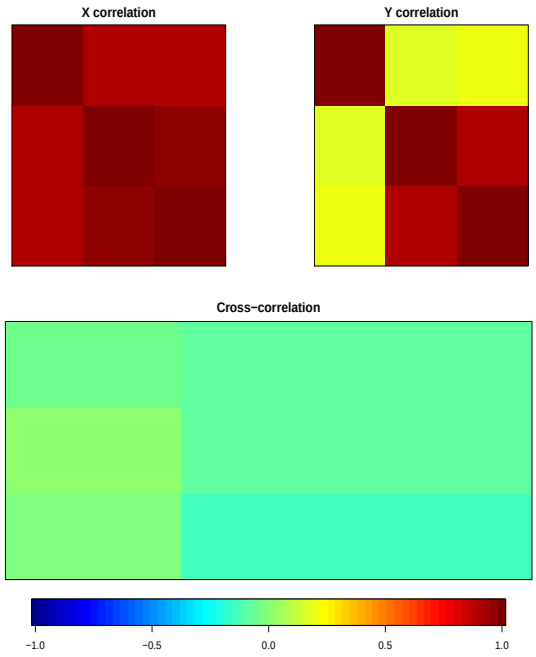


Figura 2.5: Correlaciones entre las variables

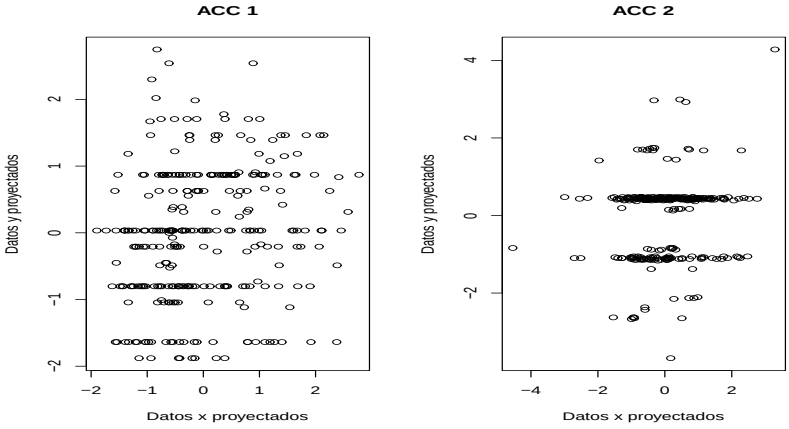


Figura 2.6: Proyecciones de las variables canónicas

Como se observó anteriormente, las relaciones entre los grupos de variables es muy pequeña y podría no ser lineal. Se propone entonces realizar el ACCK a

los datos usando el kernel gaussiano, polinomial y tangente hiperbólica. Este análisis se llevó a cabo usando el paquete kernlab en R [14]. Las correlaciones canónicas son presentados en la Cuadro 2.6. Para el kernel gaussiano se usó $\sigma = 0{,}1$, para el kernel tangente hiperbólica se usó $c = 0$ y para el kernel polinomial se utilizaron diferentes valores de c y d, pero no se logró obtener algún resultado.

Método Kernel	ρ_1	ρ_2	ρ_3
Gaussiano	0.966398	0.931275	0.916570
Tangente Hiperbólica	0.999657	0.530523	

Cuadro 2.6: Comparación de las correlaciones canónicas obtenidas mediante los diferentes métodos kernel.

Como se puede ver en el Cuadro 2.6, la primer correlación canónica obtenida usando el kernel gaussiano y tangente hiperbólica es alta, 0.966 y 0.999, respectivamente. Esto indica que al transformar los datos se obtuvo una correlacion alta.

Observando las proyecciones de las variables canónicas en la Figura 2.7 de la primera y segunda correlación canónica se tiene que los datos representados por el ACCK con kernel gaussiano parecen formar una linea horizontal, lo que indicaría una relación lineal. Los datos representados por el ACCK con kernel tangente hiperbólica para los primeros vectores canónicos se acumulan todos en un círculo en un determinado lugar mientras que las proyecciones de los datos de los segundos vectores canónicos continúan dentro de un círculo pero están dispersos.

Como se mencionó anteriormente, el ACCK carece de interpretabilidad debido a la transformación de los vectores hacia un espacio de Hilbert. Usando el paquete en R hsicCCA [5] se realizó el ACCK usando los métodos HSIC y KTA, los resultados obtenidos son presentados en el Cuadro 2.7

Método	ρ		**u**			**v**		
HsicCCA	ρ_1	0.00276	0.8536	-0.5150	-0.0776	0.2200	0.6615	0.7168
	ρ_2	0.00165	-0.0815	0.0149	-0.9965	-0.9516	0.3070	0.0088
KtaCCA	ρ_1	0.13038	0.9285	-0.0231	0.3704	-0.5109	-0.6399	-0.5739
	ρ_2	0.10559	-0.3704	0.0063	0.9288	0.8521	-0.2889	-0.4363

Cuadro 2.7: Resultados del ACCK usando los métodos HSIC y KTA.

Se puede observar que para ρ_1 en los vectores $\mathbf{u_1}$ la variable representativa es la edad en ambos métodos mientras que para el vector $\mathbf{v_1}$ las variables representativas son el número de medicamentos y el número de fármacos para el método HSIC y para el método KTA son las tres variables. Para ρ_2 la variable representativa es la talla para $\mathbf{u_2}$ en ambos métodos y para $\mathbf{v_2}$ la variable representativa es el número de enfermedades. Las proyecciones son presentadas en la Figura

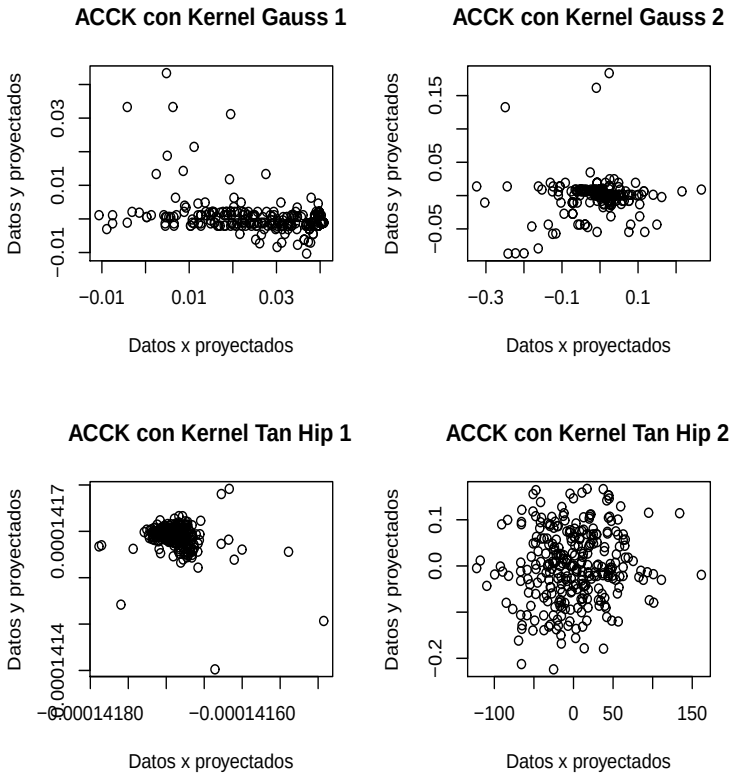


Figura 2.7: Proyecciones de las variables canónicas usando kernel

2.8. Se observa que los datos proyectados de cada pareja de vectores canónicos se comportan igual, éstos forman semicírculos pero dentro de éstos los datos se distribuyen en diferentes líneas.

En la ciencia médica se ha constatado que las interacciones entre medicamentos pueden provocar reacciones adversas de complejidad en dependencia de las características de los pacientes. En este estudio en donde se utilizan características físicas de pacientes pediátricos, al llevar a cabo el ACC con enfoque clásico se obtienen valores bajos en las correlaciones canónicas. El resultado inferencial llevado a cabo en [20] permitió arribar a la conclusión, de que a pesar de que los valores de las correlaciones son bajos, existe interrelación entre las características físicas y las reacciones medicamentosas. La utilización del ACCK abre una nueva alternativa en estos estudios, ya que como se pudo apreciar se logró una evaluación más objetiva de la mencionada interrelación.

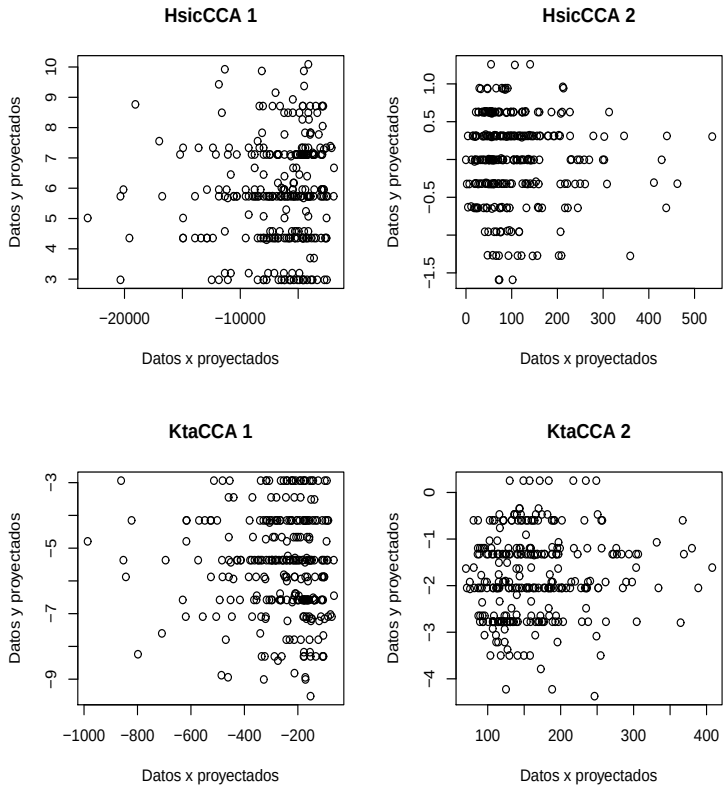


Figura 2.8: Proyecciones de las variables canónicas usando HSIC y KTA

2.3.3. Aplicación a datos de un estudio de educación

Se presenta un ejemplo aplicado a datos almacenados en el archivo mmreg.dat [1] , en el cual se aplicaron las metodologías descritas en el ejemplo anterior. Un investigador recopiló datos sobre tres variables psicológicas, cuatro variables académicas (puntajes de las pruebas estandarizadas) y de género, para 600 estudiantes de primer año de universidad. Se está interesado en la forma en que el conjunto de variables psicológicas se relaciona a las variables académicas, ver Cuadro 2.8.

Los resultados del ACC para el ejemplo son presentados en el Cuadro 2.9. Al comparar las variables psicológicas con variables académicas, mediante el ACC clásico se obtienen valores bajos en las correlaciones canónicas.

Realizando el ACCK a los datos del ejemplo, usando el paquete kernlab en R, se obtienen las correlaciones canónicas dadas en el Cuadro 2.10

[1] http://www.ats.ucla.edu/stat/r/dae/canonical.htm

Var. psicológicas (X)	Var. académicas (Y)
Centro de control	Lectura
Autoconcepto	Escritura
Motivación	Matemáticas
	Ciencias

Cuadro 2.8: Variables de estudio.

Correlaciones canónicas			
0.4464 0.1533 0.0225			
Variables canónicas			
Primer Conjunto de Variables			
	$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$
C. de control	0.0510	-0.0312	-0.0202
Autoconcepto	-0.0096	-0.0344	0.0492
Motivación	0.0510	0.1077	0.0446
Segundo Conjunto de Variables			
	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$
Lectura	0.0017	6.50e-5	0.0036
Escritura	0.0022	3.69e-3	-0.0039
Matemáticas	0.0007	1.20e-4	0.0035
Ciencias	-0.0001	-5.075e-3	-0.0036

Cuadro 2.9: Correlaciones y vectores canónicos de los datos.

Método Kernel	ρ_1	ρ_2	ρ_3
Gaussiano	0.9712	0.9449	0.9360

Cuadro 2.10: Correlaciones canónicas obtenidas mediante el kernel Gaussiano.

Mediante el uso del paquete hsicCCA en el software R se presentan los resultados obtenidos para el ejemplo dado, ver el Cuadro 2.11.

Crit.	**u**			**v**			
HSIC	0.8	0.3	0.2	-0.5	-0.5	-0.4	0.3
	3.2	-3.0	-6.3	0.6	-0.5	0.2	0.3
KTA	-0.9	0.1	-0.2	-0.7	-0.2	0.1	-0.6
	0.1	0.9	0.0	0.4	-0.2	-0.7	-0.4

Cuadro 2.11: Resultados del ACCK usando los métodos HSIC y KTA.

En resumen, primero se realizó un programa para obtener las correlaciones canónicas usando las diferentes funciones kernel, se realizó una comparación entre los resultados para diferentes valores del grado d y del coeficiente c usando el kernel polinomial obteniéndose que a medida que el grado del kernel aumenta,

las correlaciones canónicas son cercanas a 1; por otro lado, mediante el kernel gaussiano, para diferentes valores de σ se obtuvieron correlaciones canónicas igual a 1 y finalmente, al usar el kernel tangente hiperbólica, con diferentes valores de c (positivos y negativos), siempre se obtuvo una sola correlación canónica con valor igual a 1; este programa se aplicó en el problema de manglares mencionado anteriormente aunque no era necesario ya que mediante el ACC se obtuvieron valores altos para las correlaciones canónicas, sin embargo, este ejemplo facilitó observar el comportamiento de los valores de las correlaciones mediante el uso de las diferentes funciones kernel.

Se aplicó el ACCK a dos bases de datos, una de información de medicamentos y otra de educación, ya que al aplicarles el ACC se obtuvieron correlaciones canónicas bajas. Para la base de datos de medicamentos se obtuvo que al transformar los datos mediante una función kernel las correlaciones fueron más altas, especialmente usando el kernel gaussiano. Se concluye que el número y cantidad de medicamentos se relaciona con la calidad de vida del paciente aunque esta relación inicialmente no es lineal. Posteriormente mediante el método KTA se realizó el ACC no lineal, para interpretar los resultados de las variables canónicas y para los datos de medicamentos se obtuvo la misma interpretación que en el ACC, es decir, para las primeras variables canónicas la edad resulta ser el factor que más influye para determinar el número de medicamentos y fármacos. Para la base de datos de educación se obtuvo que al transformar los datos mediante el kernel gaussiano las correlaciones canónicas son altas ($\rho_1 = 0{,}9712, \rho 2 = 0{,}9449, \rho_3 = 0{,}9360$) y obteniéndose además que el centro de control está altamente correlacionado con las variables de lectura y escritura.

Capítulo 3

ACC y ACCK usando Algoritmos Genéticos

Las técnicas que dan solución al ACC y al ACCK presentadas en los capítulos anteriores son usadas bajo el supuesto de que las matrices de varianzas y covarianzas son invertibles, a continuación se presenta una alternativa para dar solución a estos problemas, mediante el uso de la técnica de Algoritmos Genéticos, ya que se trabaja el problema directamente de la definición de optimización de correlación canónica sin necesidad de hacer alguna suposición sobre los datos.

3.1. Algoritmos Genéticos

Los Algoritmos Genéticos fueron desarrollados por Holland entre los años 60's y 70's, es un método que imita la teoría de la evolución biológica de Darwin y es usado para resolver problemas de optimización global.
Cuando se tiene un problema de optimización que tiene una función objetivo a maximizar o minimizar bajo ciertas restricciones y con un espacio de soluciones factibles las cuales cumplen las restricciones se puede modelar mediante un modelo computacional, un modelo frecuentemente usado es el de Algorítmos Genéticos (AG) constituido por una población, una función de aptitud y cromosomas. La función de aptitud está relacionada con la función objetivo. El espacio de soluciones factibles representado por los cromosomas. Se conoce como fenotipo a cada solución factible, y genotipo a su representación computacional codificada en una cadena de caracteres también conocida como cromosoma. Cada cromosoma es constituido por un conjunto de genes. Los cromosomas evolucionan a través de iteraciones, llamadas generaciones. Al paso de cada generación, los cromosomas son evaluados mediante una función de aptitud, que indicará el nivel de adaptación del individuo. En cada generación a los cromosomas se les aplica el operador de cruza y mutación generando una nueva población la cual converge al óptimo global, ver Figura 3.1.
Según Coello [6] la aplicación más común de los AG ha sido la solución de

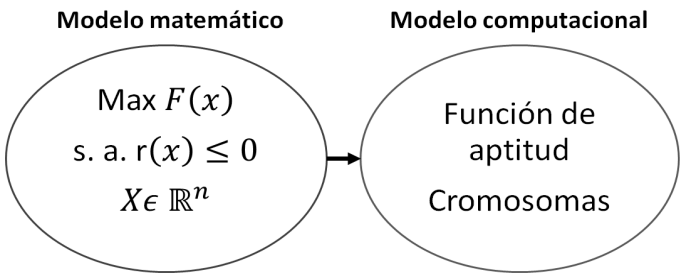


Figura 3.1: Modelación del problema.

problemas de optimización, en donde han mostrado ser muy eficientes y confiables, sin embargo, se recomienda en general tomar en cuenta en el problema las siguientes características antes de usarlos:

- Su espacio de búsqueda (posibles soluciones) debe estar delimitado dentro de un cierto rango.
- Debe poderse definir una función de aptitud que nos indique que tan buena o mala es una cierto cromosoma (solución).
- Las soluciones deben codificarse de una forma que resulte relativamente fácil de implementar en la computadora.

Algunas ventajas de los AG sobre los algoritmos de optimización tradicionales son [44], [6]:

- La habilidad de tratar con problemas complejos. Los AG tratan con varios tipos de optimización, con variable continua o discreta, con función objetivo y restricciones lineal o no lineal.
- Operan de forma simultánea con varias soluciones, en vez de trabajar de forma secuencial como las técnicas tradicionales, es decir, se puede paralelizar.
- Cuando se usan para problemas de optimización (maximizar una función objetivo) se obtienen óptimos globales.
- Los operadores de cruza y mutación se aplican con un porcentaje de probabilidad.

Algunas desventajas son:

- El modelado del cromosoma que representa el espacio de soluciones factibles.

- La definición de la función de aptitud que permite medir la mejora de la solución.
- La definición de las probabilidades para cruza y mutación.
- Pueden tardar mucho en converger, o no converger en absoluto dependiendo del modelado.

Los AG se lleva a cabo mediante el contenido genético de una población que contiene potencialmente la solución, o una mejor solución, a un problema dado de adaptación. Los cromosomas mejor adaptados representarán la mejor solución al problema de optimización. Para encontrarla se debe cambiar la información genética usando mutación y cruza, y seleccionar la mejor población [6].
De manera general en el mecanismo evolutivo se tiene lo siguiente:
Cromosoma. Los individuos (posibles soluciones del problema), pueden representarse como un conjunto de parámetros, denominados genes, los cuales agrupados forman una cadena o serie de valores, referida como cromosoma. Un ejemplo de cromosoma es el siguiente

$$\begin{bmatrix} 6 & 1 & 9 & 4 & 7 & 0 & 0 & 7 & 5 & 7 & 0 & 3 & 5 & 4 & 7 \end{bmatrix}$$

Fenotipo. Es el conjunto de números reales representando la solución factible del problema de optimización.
Genotipo. Es la representación de la solución factible codificada en un cromosoma.
La transformación fenotipo genotipo y viceversa es mostrada en la Figura 3.2.

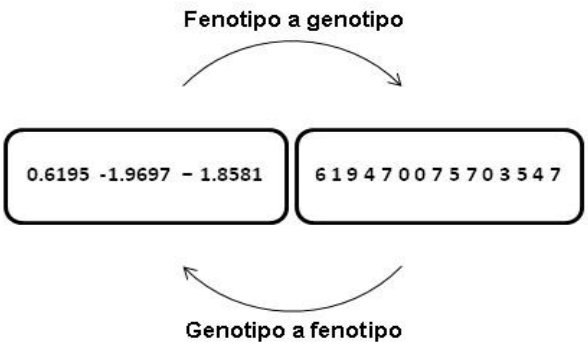


Figura 3.2: Transformación fenotipo-genotipo.

Sobre esta población elegida, algunos individuos son seleccionados aleatoriamente para reproducirse. Durante la reproducción, los nuevos individuos de la población resultan de modificaciones e intercambio genético de los padres. Hay esencialmente dos operadores genéticos.

Mutación. Consiste de simplemente cambiar un gen en el cromosoma por algún otro elegido aleatoriamente (usualmente con probabilidad pequeña), como en el siguiente ejemplo

$$\begin{bmatrix} 6 & 1 & 9 & 4 & 7 & 0 & 0 & 7 & 5 & 7 & 0 & 3 & 5 & 4 & 7 \\ & & & \downarrow & & & & & & & & & & & \\ 6 & 1 & 9 & 8 & 7 & 0 & 0 & 7 & 5 & 7 & 0 & 3 & 5 & 4 & 7 \end{bmatrix}.$$

Cruza. Es una recombinación de la información durante la reproducción de los individuos seleccionados. Se eligen dos padres y se cortan sus cadenas de cromosomas en una posición elegida al azar, para producir dos subcadenas iniciales y dos subcadenas finales. Después se intercambian las subcadenas iniciales, produciéndose dos nuevos cromosomas completos. Ambos descendientes heredan genes de cada uno de los padres. Esto es presentado a continuación:
Antes de la cruza

$$\left[\begin{array}{ccccc|ccccccccc} 6 & 1 & 9 & 4 & 7 & 0 & 0 & 7 & 5 & 7 & 0 & 3 & 5 & 4 & 7 \\ 7 & 2 & 5 & 8 & 3 & 1 & 0 & 6 & 2 & 3 & 9 & 7 & 2 & 1 & 7 \end{array}\right].$$

Después de la cruza

$$\left[\begin{array}{ccccc|ccccccccc} 7 & 2 & 5 & 8 & 3 & 0 & 0 & 7 & 5 & 7 & 0 & 3 & 5 & 4 & 7 \\ 6 & 1 & 9 & 4 & 7 & 1 & 0 & 6 & 2 & 3 & 9 & 7 & 2 & 1 & 7 \end{array}\right].$$

Función de adaptación, aptitud o fitness. La función de adaptación debe ser diseñada para cada problema de manera específica. Ésta función asigna un número real a un cromosoma particular, el cual refleja su nivel de adaptación.La función debe estar relacionada con la función objetivo a maximizar.
Los AG se inician con una población generada aleatoriamente y de un tamaño bien definido. Esta población en cada generación será modificada a través de la cruza y mutación usando la función de adaptación se determina la nueva población para la siguiente generación. El criterio de par del algoritmo está dado por el número de generaciones definidas al inicio del algorítmo.

A continuación, en el Cuadro 3.1 se presenta el pseudocódigo utilizado para el mecanismo de AG.

Dadas las condiciones que se tienen para el ACC y ACCK, como lo es la invertibilidad de las matrices de varianzas y covarianzas (o matrices de correlaciones), se realizó un programa que involucra AG como método de solución. Como primera instancia se realizó un algoritmo partiendo del problema de valores y vectores propios generalizado ecuación (1.7) para obtener las correlaciones canónicas y los vectores canónicos. Posteriormente se realizó un algoritmo que da una solución directa de la definición de correlación canónica ecuación (1.1). Finalmente se realizó uno algoritmo mas para dar solución al ACCK ecuación (2.22). A continuación se describe cada uno de ellos.

Pseudocódigo
1) Definir los parámetros (función de adaptación, tamaño de la población, razón de cruza, razón de mutación, número de generaciones).
2) Generar una población inicial aleatoriamente de posibles soluciones.
3) Evaluar la función de adaptación.
4) Obtener los genotipos de la población.
5) Realizar mutación y cruza de acuerdo a la razón de cruza y mutación.
6) Determinar la nueva población.
7) Transformar el genotipo al fenotipo de la nueva población.
8) Se regresa al paso 3 hasta llegar al número de generaciones.

Cuadro 3.1: Pseudocódigo del algoritmo genético.

3.2. Programa elaborado para resolver el ACC mediante AG en el formalismo de Lagrange

En el ACC se requieren encontrar los vectores ρ, $\mathbf{a}$ y $\mathbf{b}$. Siguiendo la teoría presentada en el primer capítulo, se llegó a un problema de valores y vectores propios generalizado, dado en la ecuación (1.7). Para trabajar con Algoritmos Genéticos se debe tener una función objetivo que no es más que la función a maximizar. Para éste problema de ACC, se tiene la siguiente función objetivo

$$F(\mathbf{a}, \lambda) = \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\mathbf{a} - \lambda^2\Sigma_{XX}\mathbf{a} = 0. \tag{3.1}$$

Dado que se quieren encontrar los valores en los que $F(\mathbf{a}, \lambda)$ sea igual a 0 una manera más práctica de tratar el problema es cambiar la forma de verlo. Lo que se quiere encontrar es una raíz para $F(\mathbf{a}, \lambda)$ lo cual es equivalente a obtener el mínimo $F'(\mathbf{a}, \lambda)F(\mathbf{a}, \lambda)$ (ver Figura 3.3), es decir,

$$minF'(\mathbf{a}, \lambda)F(\mathbf{a}, \lambda) \tag{3.2}$$

para la que se restringe la condición $0 < \lambda^2 \leq 1$, ya que λ nos determinará el valor de la correlación.Como se vió anteriormente, no es necesario incluir a $\mathbf{b}$, ya que se encuentra en términos de $\mathbf{a}$, ecuación **??**, es decir, el espacio de soluciones factibles está dado por λ y $\mathbf{a}$.

Como se puede observar, es necesario determinar las matrices de varianzas y covarianzas entre los grupos de variables ya que son necesarios para la función. En particular, en el problema se requiere encontrar valores $\rho = \lambda^2$ y $\mathbf{a}$ tal que (3.2) sea máxima, usualmente se modela el problema de minimización con la función de aptitud $1/F'F$. Las soluciones factibles ρ y $\mathbf{a}$ están representadas por un cromosoma de dimensión $N + 1$. Cada entrada del vector se define como una cadena de 5 dígitos entre 0 y 9 de acuerdo a la propuesta de Ison et al [12]. Posteriormente se elige una población inicial aleatoria y se le realiza el procedimiento descrito en la subsección anterior. Cabe aclarar que para determinar los mejores resultados o los mejores valores se utilizaron dos criterios, el primero es

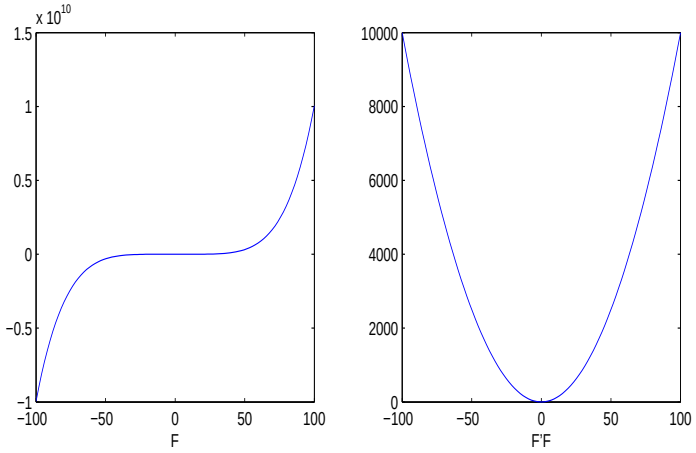


Figura 3.3: Función objetivo ACC

tomar los valores con menor error de $F'F$ y el segundo es verificar las reestricciones dadas en la ecuación (1.2), es decir, se tomarán los valores que tengan un error pequeño.

3.3. Solución directa del ACC usando Algoritmos Genéticos

A continuación se presenta una solución directa del ACC usando el método AG. En el ACC se requieren encontrar los vectores ρ, $\mathbf{a}$ y $\mathbf{b}$. Se propone encontrar la solución directamente de la ecuación (1.1), que se denotará de ahora en adelante como

$$F(\mathbf{a},\mathbf{b}) = \frac{\mathbf{a}'\Sigma_{XY}\mathbf{b}}{\sqrt{\mathbf{a}'\Sigma_{XX}\mathbf{a}}\sqrt{\mathbf{b}'\Sigma_{YY}\mathbf{b}}}, \tag{3.3}$$

y mediante la cual se determinarán $\mathbf{a}$ y $\mathbf{b}$ tal que al sustituirse en la ecuación se obtenga un máximo global, que representará la primera correlación canónica ρ, todo esto sujeto a las restricciones (1.2) y además se debe cumplir la siguiente condición $0 < F(\mathbf{a},\mathbf{b}) \leq 1$.
En el problema se requiere encontrar valores $\mathbf{a}$ y $\mathbf{b}$ tal que (3.3) sea máxima. Para el programa se tomó como función objetivo $-F$, para que el problema sea visto como un problema de mínimos y se toma $1/F$ como función de adaptación o fitness. Para este programa, si N_1 es el número de variables de X y N_2 es el número de variables de Y, entonces el cromosoma tendrá $N_1 + N_2$ genes. Para determinar el valor óptimo que maximiza (3.3), se realizó de manera análoga el procedimiento de AG descrito en la sección anterior. Para determinar los mejores resultados o los mejores valores se verificaron las reestricciones dadas en la ecuación (1.2), es decir, se tomaron los valores con un error pequeño.

3.4. Análisis de correlación canónica no lineal usando Algoritmos Genéticos

En su artículo, Chang desarrolla la solución para el HSIC y KTA mediante el uso del método de gradiente descendente. A continuación se propone determinar la solución específicamente del KTA mediante AG, ya que mediante éste método se obtiene una solución directa de la definición de KTA.
Se quiere dar solución al ACCK, usando AG para el método KTA. Mediante éste método se requieren encontrar los vectores ρ, $\mathbf{u}$ y $\mathbf{v}$. Para el problema de ACCK (o ACC no lineal) con el método KTA, se propone encontrar la solución directamente de la ecuación (2.22), que se denota como

$$F(\mathbf{u}, \mathbf{v}) = \frac{tr(K^u \tilde{K}^v)}{\sqrt{tr(K^u \tilde{K}^u) tr(K^v \tilde{K}^v)}} \tag{3.4}$$

y mediante la cual se determinarán $\mathbf{u}$ y $\mathbf{v}$ tal que al sustituirse en la ecuación se obtenga un máximo global, que será la primera correlación canónica ρ_a, todo esto sujeto a las restricciones $\|u\| = \|v\| = 1$ y además se debe cumplir la siguiente condición $0 < F(\mathbf{u}, \mathbf{v}) \leq 1$.
Como primer paso, se aplicará el kernel gaussiano a las matrices de datos $\mathbf{X}$ y $\mathbf{Y}$ para obtener las matrices kernel. La mátriz kernel ésta definida en (2.23) y la matriz kernel centrada en (2.24). Para poder calcular las matrices kernel usando el kernel gaussiano se debe introducir el valor de sigma para cada grupo de variables, en éste caso, al igual que Chang, se utilizó el llamado ´´median trick"para poder determinar éste parámetro, que no es más que la mediana de la distancia Euclideana entre las parejas (x_i, x_j). El código para determinar a σ_x y σ_y se encuentra en el Cuadro 3.2. El código para la obtención de las matrices kernel se encuentra en el Cuadro 3.3.
En el problema se requiere encontrar valores $\mathbf{u}$ y $\mathbf{v}$ tal que (3.4) sea máxima. Para el programa se tomó como función objetivo $-F$, para que el problema sea visto como un problema de mínimos y se toma $1/F$ como función de adaptación o fitness.
Para este programa, si N_1 es el número de variables de X y N_2 es el número de variables de Y, entonces el cromosoma tendrá $N_1 + N_2$ genes. Para determinar el valor óptimo que maximiza (3.4), se realizó de manera análoga el procedimiento de AG descrito en la sección anterior. Para determinar los mejores resultados o los mejores valores se verificaron las reestricciones $\|u\| = \|v\| = 1$, es decir, se tomaron los valores con un error pequeño.

3.5. Características del AG para el ACC y ACCK

Se realizó un algoritmo en el software OCTAVE [30] para encontrar las correlaciones canónicas y los vectores canónicos usando AG siguiendo el pseudocódigo dado en la sección 3.1. El programa fué corrido en una computadora con procesador Intel(R)Core(TM)i525000 @ 3.30 GHz con memoria 8.00 GB RAM. Se

Datos de entrada
X % *Matriz de datos del primer grupo de variables*
Y % *Matriz de datos del segundo grupo de variables*
Método
$[N1, m1] = size(X)$;
$dimensionx = N1 * (N1 - 1)/2$;
$distanciax = zeros(1, dimensionx)$;
$banx = 0$;
$fork = 1 : N1 - 1$
$forj = k + 1 : N1$
$banx = banx + 1$;
$distanciax(banx) = norm(X(k,:) - X(j,:))$;
end
end
$sigmax = median(distanciax)$;
% *Se realiza lo mismo para Y*

Cuadro 3.2: Algorítmo para calcular σ_x, σ_y.

realizó un algoritmo para cada caso de, los antes mencionados. El código del programa utilizado para el resolver el problema de ACCK KTA para éste trabajo son dados en los cuadros 3.4 y 3.5, para los otros problemas el algoritmo se trabajó de una forma similar.
Para obtener los mejores resultados se llevó a cabo un procedimiento de calibración sobre los parámetros: tamaño de la población, número de generación, razón de mutación y razón de cruza. El procedimiento consiste en fijar tres parámetros mientras el cuarto es corrido en un rango de valores de posibles mejores soluciones. Por ejemplo, primero se fijan los parámetros
Número de generaciones = 300
Razón de mutación = 0.2
Razón de cruza = 0.8
y el programa es ejecutado iterando el tamaño de la población en el rango entre 10 y 100 (con lapsos de 10 en 10). Supóngase que la mejor aproximación a la solución fué obtenida en
Tamaño de la población = 20,
posteriormente se realizará lo mismo para los otros parámetros corriendo el programa con: número de generación entre 100 y 1500 (con lapsos de 100); razón de mutación entre 0.01 y 0.3 (con lapsos de 0.01); razón de cruza entre 0.7 y 0.9 (con lapsos de 0.01). Obteniéndose los mejores valores para cada parámetro.
El programa que se realizó tiene como criterio de paro el número de generaciones.

Datos de entrada
X % *Matriz de datos del primer grupo de variables*
Y % *Matriz de datos del segundo grupo de variables*
Método
function [rho,Kx,Kyc]=funcion$_K TA(X, Y, u, v, sigmax, sigmay, N1)$
% Kernel Gaussiano
$fori = 1 : N1$
$forj = i : N1$
$AUX1 = X(i,:) - X(j,:);$
$prod1 = u * AUX1';$
$Kx(i,j) = exp(-sigmax * prod1^2);$
$Kx(j,i) = Kx(i,j);$
% Se realiza lo mismo para Ky
end
end
% Kernel Gaussiano centrado
$auxtotal1 = sum(sum(Kx));$
$auxi1 = sum(Kx);$
$auxj1 = sum(Kx');$
$fori = 1 : N1$
$forj = 1 : N1$
$Kxc(i,j) = Kx(i,j) - auxi1(j)/N1 - auxj1(i)/N1 + auxtotal1/(N1^2);$
end
end
% Se realiza lo mismo para Kyc
$rho1 = sqrt(trace(Kx * Kxc) * trace(Ky * Kyc));$
$rho = trace(Kx * Kyc)/rho1;$

Cuadro 3.3: Algorítmo para calcular matrices kernel.

Datos de entrada
X % *Matriz kernel para el primer grupo de variables* Y % *Matriz kernel para el segundo grupo de variables* Largo % *Tamaño del cromosoma* Ngen % *Número de generaciones* Pob % *Tamaño de la población* Mut % *Razón de mutación* Cross % *Razón de cruza*
Método
$[N1, m1] = size(X)$; $[N2, m2] = size(Y)$; $N = m1 + m2$ $sigmax = sigma(X)$ $sigmay = sigma(Y)$ $minN = 1$; $maxN = N$; $for\ j = minN : maxN$ $min(j) = -1$; $max(j) = 1$; *end*
Generación de población
$funTotal = rand(1, pob)/rand$; $pobla = rand(pob, N)$; $for\ k = 1 : pob$ $pobla(k, 1 : m1) = pobla(k, 1 : m1)/norm(pobla(k, 1 : m1))$; $pobla(k, m1 + 1 : N) = pobla(k, m1 + 1 : N)/norm(pobla(k, m1 + 1 : N))$; *end*
Transformación fenotipo-genotipo[12]
$for\ generacion = 1 : Ngen$ $rank = 1./funTotal$; *pobn=fentogen(pobla,min,max,largo);* *pobn=pareja2(pobn,rank,mut,cross);* *poblafin=gentofen(pobn,min,max,largo);* *pobla=poblafin;* *[pob,m]=size(pobla);*

Cuadro 3.4: Algorítmo para KTA con AG.

Evalua cada solución de la población
$for\ kpob = 1 : pob$
$pobla(kpob)$;
$u = pobla(kpob, minN : m1)$;
$v = pobla(kpob, m1 + 1 : maxN)$;
$funrho = funcion_K TA(X, Y, u, v, sigmax, sigmay, N1)$;
$funTotal(kpob) = -funrho + 1$;
end
$funTotal$;
$[Pob, Indice] = sort(funTotal)$;
end
$error1 = norm(u) - 1$;
$error2 = norm(v) - 1$;
$rho = funrho$;
Datos de salida
rho % correlación canónica
u % vector canónico para el primer grupo
v % vector canónico para el segundo grupo
$error1$
$error2$

Cuadro 3.5: Algorítmo para KTA con AG (continuación).

Capítulo 4

Aplicaciones y estudios comparativos

4.1. Comparación de tres métodos de optimización del ACC en casos reales

A continuación se presentan diferentes comparaciones realizadas tanto para el ACC como el ACCK. Estas comparaciones se realizaron entre diferentes bases de datos encontradas en la literatura, de simulación y datos reales, y fueron utilizadas conforme a la adecuación a cada método ACC. Principalmente se realizan comparaciones entre los resultados de los métodos usados clásicamente con los resultados obtenidos mediante Algorítmos Genéticos.

4.1.1. Caso de manglares

Se obtuvieron los valores de las correlaciones canónicas y los vectores canónicas usando los algoritmos de los tres métodos mencionados en el primer capítulo (SVD, Cholesky y QR). En el Cuadro 4.1, se presentan las correlaciones canónicas y vectores canónicos obtenidos mediante el método de descomposición QR.
Se puede observar que los tres primeros valores de las correlaciones canónicas son altos, lo que indica que se tomarán sólo los tres primeros vectores canónicos para cada grupo de variables. Los resultados ya fueron analizados en el primer capítulo. Lo que se realizó posteriormente fue la determinación de las correlaciones y vectores canónicos en cada uno de los métodos, que se entiende son valores similares debido a que se les aplicó una transformación que relaciona los valores entre ellos presentada en la sección 1.1.4.
Para determinar que algoritmo trabaja con mejor precisión, una vez obtenidos los valores de **a** y **b**, se verificó que se cumplieran las restricciones, es decir, se sustituyeron los valores de **a** y **b** en las restricciones (1.2). Para cada método se obtuvieron los errores y estos están dados en el Cuadro 4.2.

Correlaciones canónicas					
0.9988 0.9807 0.8183 0.5873 0.3675					
Variables canónicas					
Primer Conjunto de Variables					
	$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$	$\mathbf{a}_4$	$\mathbf{a}_5$
CrecDia	0.2378	-0.6262	0.0472	-0.8801	0.3291
Hojarasca	-0.2941	0.0835	0.6470	-0.5307	-0.6332
DenInd	0.1494	-0.0818	-1.0879	-0.4668	-0.5272
Areabasal	0.1768	0.7279	-0.4746	-0.6364	0.6340
Altura	0.9084	-0.0078	-0.3769	0.5611	-0.8464
Segundo Conjunto de Variables					
	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$	$\mathbf{b}_4$	$\mathbf{b}_5$
NO3	0.2944	-0.4053	-0.8605	0.4322	-0.8074
PO4	1.2336	-0.8329	-0.7603	1.2386	-0.0823
SO4	0.3318	-0.9758	1.7132	0.1631	0.7257
NH4	-0.6614	0.8621	0.5915	-1.0156	0.2195
Salinidad	-1.1640	0.6419	-0.7309	-1.7107	1.0996
REDOX	0.0289	-0.0882	-0.2046	0.6777	-0.4449
pH	0.0717	-0.1720	-0.1023	-0.9633	0.7517
Temperatura	0.2971	-0.5575	0.9218	-0.7954	0.0585

Cuadro 4.1: Correlaciones y vectores canónicos obtenidos por el método de descomposición QR para el sistema lagunar Chacahua-Pastorias.

Se puede observar que, los resultados obtenidos mediante el método de descomposición QR son mas exactos en comparación con el método de Cholesky y el método de descomposición SVD, que presentan errores similares.

4.1.2. Caso datos de educación

Se aplicó el ACC a los grupos de variables dadas en el Cuadro 2.8 del Capítulo 2 de los datos de educación, utilizando los tres diferentes algoritmos presentados anteriormente. Usando las matrices de correlaciones obtenidas para cada grupo se obtuvieron los resultados para el método que utiliza descomposición de Cholesky y el método SVD y para el método de factorización QR se utilizaron directamente las matrices de datos de los dos grupos. En la el Cuadro 4.3 se presentan los resultados obtenidos para cada método. Observe que los valores son diferentes. Para demostrar la equivalencia entre el método SVD y el método de Cholesky, se aplica la transformación (1.15) a este último. Posteriormente, para demostrar la equivalencia entre el método SVD y QR, los resultados de este último se dividen entre la expresión (1.16).
Posteriormente, una vez obtenidos los valores de $\mathbf{a}$ y $\mathbf{b}$, se verificó que se cumplieran las restricciones, es decir, se sustituyeron los valores de $\mathbf{a}$ y $\mathbf{b}$ en las restricciones (1.2) y para cada método se obtuvieron los errores, estos se pre-

Métodos			
	Cholesky	SVD	QR
$\mathbf{a}_1$	1.671692118065948e-09	4.069932391104203e-09	-4.440892098500626e-016
$\mathbf{a}_2$	4.237717954325149e-09	1.581991870835964e-09	-6.661338147750939e-016
$\mathbf{a}_3$	-2.809401322689809e-08	-5.373831268862261e-09	-1.554312234475219e-015
$\mathbf{a}_4$	-3.659740732686601e-09	-1.859771003864807e-09	2.220446049250313e-016
$\mathbf{a}_5$	1.590959963770899e-08	1.581671904560267e-09	-8.881784197001252e-016
$\mathbf{b}_1$	-3.453718733226197e-08	1.501404645054549e-08	0
$\mathbf{b}_2$	-1.564867124770331e-08	-2.740188687067047e-08	-4.440892098500626e-016
$\mathbf{b}_3$	-6.933913199169695e-08	-2.240607910319881e-08	1.332267629550188e-015
$\mathbf{b}_4$	1.827544116039803e-08	-2.711097590868405e-08	-4.440892098500626e-016
$\mathbf{b}_5$	-1.042203912060558e-08	3.840211149075401e-08	-1.110223024625157e-016

Cuadro 4.2: Errores obtenidos en cada método al verificar las restricciones.

Correlaciones Canónicas									
0.4641	0.1675		0.1040						
Variables Canónicas									
	Cholesky			SVD			QR		
	$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$	$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$	$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$
C. de control	0.9045	-0.3879	-0.1775	-0.8404	-0.4166	-0.4435	0.0512	-0.0254	-0.0270
Autoconcepto	-0.1360	-0.6565	0.7419	0.2479	-0.8379	0.5833	-0.0144	-0.0485	0.0338
Motivación	0.4043	0.6469	0.6465	-0.4327	0.6948	0.6865	0.0516	0.0828	0.0817
	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$
Lectura	-0.8408	-0.3582	0.1352	-0.4508	-0.0496	0.2160	0.0018	-0.0002	0.0009
Escritura	-0.4466	0.3688	0.2160	-0.3490	0.4092	0.8881	0.0015	0.0017	0.0037
Matemáticas	-0.1096	-0.2215	-0.0013	-0.2205	0.0398	0.0885	0.0010	0.0002	0.0004
Ciencias	0.0310	-0.6657	-0.5448	-0.0488	-0.8266	-1.0661	0.0002	-0.0035	-0.0045
Sexo	-0.2840	0.4934	-0.7989	-0.3150	0.5406	-0.8944	0.0258	0.0443	-0.0733

Cuadro 4.3: Resultados obtenidos usando cada uno de los métodos, sin realizar alguna transformación.

sentan en el Cuadro 4.4. En donde se puede observar que el método con menor error es el método de descomposición QR.
Se observa que al aplicar el ACC a un ejemplo en particular, se obtuvo que el algoritmo con menor error al calcular las correlaciones canónicas y los vectores canónicos es el que utiliza factorización QR ya que reduce el número de operaciones.

4.2. Estudios comparativos del Algoritmo Genético en ACC

Se utilizaron diversos ejemplos para determinar las correlaciones y vectores canónicos usando AG y se compararon con los resultados para los tres métodos aplicados anteriormente (SVD, Cholesky y QR).

Errores obtenidos usando **a** en las restricciones			
	Cholesky	SVD	QR
$\mathbf{a}_1$	0	4.44089209850063e-16	-6.661338147750939e-016
$\mathbf{a}_2$	-4.44089209850063e-16	-1.11022302462516e-16	4.440892098500626e-016
$\mathbf{a}_3$	-1.11022302462516e-16	-6.66133814775094e-16	0
Errores obtenidos usando **b** en las restricciones			
	Cholesky	SVD	QR
$\mathbf{b}_1$	1.15394054311935e-08	1.54838777355337e-08	4.440892098500626e-016
$\mathbf{b}_2$	2.34737154158893e-08	3.43547208458972e-08	4.440892098500626e-016
$\mathbf{b}_3$	7.96218141418947e-08	-3.10447266782532e-08	1.554312234475219e-015

Cuadro 4.4: Errores obtenidos en cada método al verificar las restricciones.

4.2.1. Comparaciones del ACC con AG mediante solución de valores y vectores propios

Caso datos de cabezas

A continuación se presenta un ejemplo publicado por Mardia et al. [23]. Se consideran los datos de $n = 25$ familias para las variables en donde se tomaron medidas de la cabeza del primer y segundo hijo (Cuadro 4.5).

Primer grupo de variables (X)	Segundo grupo de variables (Y)
Longitud de la cabeza del primer hijo	Ancho de la cabeza del primer hijo
Longitud de la cabeza del segundo hijo	Ancho de la cabeza del segundo hijo

Cuadro 4.5: Variables de los datos de cabezas

Como se puede ver, cada grupo tiene dos variables, entonces el cromosoma tendrá tres genes representando a λ y al vector propio propio **a** con dos entradas. Por ejemplo, el fenotipo

$$\left[\begin{array}{ccc} \lambda & \mathrm{a}_1 & \mathrm{a}_2 \end{array}\right] = \left[\begin{array}{ccc} 0.4368 & 1.4397 & 1.0553, \end{array}\right]$$

es transformado al siguiente genotipo

$$\left[\begin{array}{ccc} \lambda & \mathrm{a}_1 & \mathrm{a}_2 \end{array}\right] = \left[\begin{array}{ccccccccccccccc} 4 & 3 & 6 & 7 & 8 & 5 & 0 & 7 & 1 & 9 & 5 & 0 & 5 & 2 & 7 \end{array}\right],$$

haciendo una codificación de cinco dígitos por gen.
Se calibró el algoritmo, obteniéndose lo siguientes parámetros
Número de generaciones = 1400
Tamaño de la población = 20
Razón de mutación = 0.06
Razón de cruza = 0.65
Una vez calibrado el programa, éste se corrió para obtener la correlación canónica y los vectores canónicos. Se aplicaron los tres métodos del ACC a éste

ejemplo para calcular las correlaciones y vectores canónicos, además, se aplicó el método de algoritmos genéticos; para llevar a cabo este análisis se usaron los programas en OCTAVE de cada método. Los resultados son presentados en el Cuadro 4.6 y los errores en las restricciones en el Cuadro 4.7 en donde ρ_1 y ρ_2 representan la primera y segunda correlación canónica, $\mathbf{a}_1$ con $\mathbf{b}_1$ los primeros vectores canónicos y $\mathbf{a}_2$ con $\mathbf{b}_2$ los segundos vectores canónicos.

Método	Cor. canónica		Vectores canónicos			
ACC	ρ		**a**		**b**	
Chol	ρ_1	0.78855	0.55215	0.52155	0.50450	0.53824
	ρ_2	0.05382	-1.36650	1.37846	-1.76827	1.75829
SVD	ρ_1	0.78855	-0.55215	-0.52155	-0.50450	-0.53824
	ρ_2	0.05382	-1.36650	1.37846	-1.76827	1.75829
QR	ρ_1	0.78850	-0.55218	-0.52153	-0.50444	-0.53828
	ρ_2	0.05373	-1.36637	1.37836	-1.76856	1.75856
AG	ρ_1	**0.78866**	-0.55800	-0.52799	-0.51016	-0.54436
	ρ_2	**0.05310**	1.33800	-1.35000	1.75302	-1.74716

Cuadro 4.6: Correlaciones y vectores canónicos de los datos de cabezas.

ACC	ρ	Error 1	Error 2
Chol	ρ_1	4.440892098500626e-016	-3.330669073875470e-016
	ρ_2	0	-7-771561172376096e-016
SVD	ρ_1	-2.220446049250313e-016	-3.330669073875470e-016
	ρ_2	0	-2.220446049250313e-016
QR	ρ_1	-2.220446049250313e-016	-1.110223024625157e-015
	ρ_2	2.220446049250313e-016	-1.221245327087672e-015
AG	ρ_1	7.332638494119999e-008	2.300958079999749e-002
	ρ_2	4.710887503482643e-008	-4.107195999997837e-002

Cuadro 4.7: Errores en las restricciones de los datos de cabezas.

Como se puede observar, los resultados entre los métodos propuestos en la literatura con los resultados obtenidos mediante el programa de AG son similares, y además concuerdan con los publicados por Mardia et al. [23].

Caso datos de educación

Para la base de datos de educación, nuevamente se calibró el algoritmo para éstos datos y se obtuvieron los parámetros siguientes
Número de generaciones=200
Tamaño de la población=60
Razón de mutación=0.17
Razón de cruza=0.75
A continuación se presentan las tres correlaciones canónicas y sus vectores canónicos obtenidos usando los métodos clásicos para el ACC y los mejores resultados

obtenidos usando algoritmos genéticos (ver Cuadro 4.8 y 4.9). Se observa que los resultados obtenidos mediante cada método son similares con errores muy pequeños.

Mét	Cor. canónica		Vectores canónicos						
ACC	ρ		**a**			**b**			
Chol	ρ_1	0.4462	0.8388	-0.1664	0.4265	0.4457	0.5337	0.1832	-0.0358
	ρ_2	0.1525	-0.5116	-0.5955	0.9035	-0.0214	-0.8751	-0.0323	1.2101
	ρ_3	0.0228	-0.3332	0.8493	0.3761	-0.8921	0.9401	-0.8263	0.8525
SVD	ρ_1	0.4462	-0.8388	0.1664	-0.4265	-0.4457	-0.5337	-0.1832	0.0358
	ρ_2	0.1525	-0.5116	-0.5955	0.9035	0.0214	0.8751	0.0323	-1.2101
	ρ_3	0.0228	-0.3332	0.8493	0.3761	0.8921	-0.9401	0.8263	-0.8525
QR	ρ_1	0.4464	0.8379	-0.1670	0.4281	0.4450	-0.5358	0.1826	-0.0368
	ρ_2	0.1533	-0.5134	-0.5941	0.9034	0.0160	0.8794	0.0278	-1.2055
	ρ_3	0.0225	-0.3328	0.8502	0.3747	0.8924	-0.9349	0.8268	-0.8589
AG	ρ_1	**0.4472**	-0.8380	0.1680	-0.4260	-0.4441	-0.5322	-0.1826	0.0359
	ρ_2	**0.1525**	0.5200	0.5980	-0.9180	-0.0233	-0.8866	-0.0338	1.2269
	ρ_3	**0.0228**	-0.3280	0.8480	0.3680	0.8869	-0.9755	0.8214	-0.7969

Cuadro 4.8: Correlaciones y vectores canónicos de los datos de educación.

ACC	ρ	Error 1	Error 2
Chol	ρ_1	0	5.185136409124880e-007
	ρ_2	-1.110223024625157e-016	-2.531819347506570e-007
	ρ_3	-2.220446049250313e-016	2.844562994241962e-005
SVD	ρ_1	6.661338147750939e-016	-1.761943296374469e-006
	ρ_2	6.661338147750939e-016	2.121832691326375e-006
	ρ_3	2.220446049250313e-016	-1.523801746117215e-006
QR	ρ_1	0	8.881784197001252e-016
	ρ_2	4.440892098500626e-016	-1.110223024625157e-016
	ρ_3	4.440892098500626e-016	8.881784197001252e-016
AG	ρ_1	2.633681119999998e-02	2.653008930530021e-02
	ρ_2	-1.217026560000023e-02	-5.771939257599890e-03
	ρ_3	-2.573973600000112e-03	-6.918124880499921e-03

Cuadro 4.9: Errores en las restricciones de los datos de educación.

En los dos ejemplos presentados se realizan las comparaciones con cada método para verificar el buen funcionamiento del programa de algoritmos genéticos. A continuación se presenta un problema real al que se requiere encontrar el grado de relación entre los dos grupos de variables.

Caso datos de Carbono

Dado que no existe información sobre secuestro de carbono en sistemas forestales en el Estado de Puebla, México, y en particular, en la zona de Teziutlán,

se hizo necesario iniciar estudios de investigación pertinentes. El problema que se planteó investigar es que, como consecuencia de las variaciones climáticas ocurridas en los suelos que soportan vegetación forestal, en la región de Teziutlán, se considera que también han ocurrido variaciones en el contenido del carbono orgánico en esos suelos. Por ello, se aplicaron determinadas técnicas para realizar la medición correspondiente del contenido de carbono orgánico, en esos suelos, en muestras seleccionadas durante los años 1987 y 2009 [3], [20]. Se midieron variables que caracterizan la textura de los suelos (grupo X en la Cuadro 4.10) y otras variables que caracterizan la fertilidad de los suelos (grupo Y en la Cuadro 4.10) con el objetivo de evaluar el grado de asociación entre esos grupos de variables.

Primer grupo de variables (X)	Segundo grupo de variables (Y)
Densidad aparente	Porcentaje de materia orgánica
Porcentaje de arena	Porcentaje de carbono orgánico
Porcentaje de limo	Carbono orgánico en suelos
Porcentaje de arcilla	Porcentaje de nitrógeno total
	Relación carbono nitrógeno

Cuadro 4.10: Variables de los datos de carbono.

Como se desea tener una medida del grado de asociación entre estos dos grupos de variables, se decide realizar el ACC usando los programas de los métodos de Cholesky, SVD y descomposición QR, sin embargo, no se pudo realizar este análisis, ya que la matriz de varianzas y covarianzas de los datos del grupo X no es definida positiva y por lo tanto, no se puede aplicar el teorema de Cholesky a la matriz [28]. Por otro lado, la matriz de varianzas y covarianzas del grupo Y no es invertible, como se puede ver en la versión corta de la matriz de varianzas y covarianzas en (4.1), mismo efecto que se puede observar en la matriz de correlaciones (4.2) del grupo Y (denotada por R_{yy}). Este hecho hace que la matriz de varianzas y covarianzas no cumpla con la condición de invertibilidad que es un supuesto para cada uno de los métodos.

$$\Sigma_{yy} = \begin{bmatrix} 2{,}0292e+001 & 1{,}1770e+001 & 1{,}6611e+002 & 6{,}7708e-001 & 1{,}8419e+000 \\ 1{,}1770e+001 & 6{,}8273e+000 & 9{,}6352e+001 & 3{,}9274e-001 & 1{,}0684e+000 \\ 1{,}6611e+002 & 9{,}6352e+001 & 8{,}5190e+003 & 1{,}8940e+000 & -1{,}5539e+001 \\ 6{,}7708e-001 & 3{,}9274e-001 & 1{,}8940e+000 & 4{,}1738e-001 & 3{,}1020e-001 \\ 1{,}8419e+000 & 1{,}0684e+000 & -1{,}5539e+001 & 3{,}1020e-001 & 9{,}3903e+000 \end{bmatrix}, \tag{4.1}$$

$$R_{yy} = \begin{bmatrix} 1{,}0000 & 1{,}0000 & 0{,}3995 & 0{,}2327 & 0{,}1334 \\ 1{,}0000 & 1{,}0000 & 0{,}3995 & 0{,}2327 & 0{,}1334 \\ 0{,}3995 & 0{,}3995 & 1{,}0000 & 0{,}0318 & -0{,}0549 \\ 0{,}2327 & 0{,}2327 & 0{,}0318 & 1{,}0000 & 0{,}1567 \\ 0{,}1334 & 0{,}1334 & -0{,}0549 & 0{,}1567 & 1{,}0000 \end{bmatrix}. \tag{4.2}$$

Dado este problema, se propone entonces resolver el ACC usando algoritmos

genéticos directamente de la ecuación (3.3), ya que debido a la definición de ésta ecuación no es necesario encontrar la inversa de las matrices de varianza y covarianzas de los datos, este procedimiento es presentado en la siguiente sección.

4.2.2. Comparaciónes del ACC con AG mediante solución directa

Caso datos de cabezas

Se aplicó este algoritmo al ejemplo de datos de cabezas de la Sección 3. Como primer paso se llevó a cabo un procedimiento de calibración en donde se obtuvieron los parámetros siguientes
Número de generaciones=1100
Tamaño de la población=80
Razón de mutación=0.044
Razón de cruza=0.69.
Posteriormente, mediante el programa se obtuvieron los vectores canónicos **a** y **b** tales que la correlación canónica ρ es máxima, estos valores están dados en el Cuadro 4.11. Observe que los resultados obtenidos mediante este método son similares a los presentados en el Cuadro 4.6, solo que en este método sólo se logró obtener la correlación máxima que representa a la primera correlación canónica.

Cor	Vectores Canónicos				Error	Error
ρ	**a**		**b**		1	2
0.7885	0.5539	0.5202	0.5219	0.5434	0.0007	0.0437

Cuadro 4.11: Correlación canónica y vectores canónicos de los datos de cabezas.

Note que con éste método **a** y **b** son determinados con los AG directamente de la definición de ACC para obtener a ρ, a diferencia del método de AG anterior, en donde **b** era determinado a partir de **a**.

Caso datos de educación

Se aplicó este algoritmo al ejemplo de datos de educación. Se llevó a cabo un procedimiento de calibración en donde se obtuvieron los parámetros siguientes
Número de generaciones=100
Tamaño de la población=90
Razón de mutación=0.18
Razón de cruza=0.66.
Mediante el programa se obtuvieron los primeros vectores canónicos **a** y **b** con primera correlación canónica ρ, estos valores están dados en el Cuadro 4.12. Al comparar estos valores con los del Cuadro 4.8, se observa que son similares.

Cor.	Vectores Canónicos							Error	Error
ρ	**a**			**b**				1	2
0.4439	-0.7687	0.2115	-0.5052	-0.4300	-0.5588	-0.2033	0.0124	-0.0361	0.0847

Cuadro 4.12: Correlación canónica y vectores canónicos de los datos de educación.

Nuevamente, estos dos ejemplos ayudan a verificar el buen funcionamiento de este segundo programa de algoritmos genéticos.

Caso datos de Carbono

Para determinar la correlación canónica máxima y sus vectores canónicos en el problema de los datos de carbono se utilizó el programa de AG que encuentra estos valores directamente de la definición [28]. Como primer paso se realizó la calibración de los parámetros del programa, obteniéndose
Número de generaciones=900
Tamaño de la población=80
Razón de mutación=0.055
Razón de cruza=0.72.
Posteriormente, mediante el programa se determinaron los valores de los vectores canónicos de tal manera que al sustituirlos en (3.4) se obtiene el valor máximo, que representa a la primera correlación canónica. El mejor valor obtenido de la correlación y los vectores canónicos están presentados en el Cuadro 4.13 y sus errores en 4.14; se presentan además, los errores en las reestricciones que fueron los más pequeños que se obtuvieron.

Cor.	Vectores Canónicos								
ρ	**a**				**b**				
0.5741	-0.925	-1.506	-0.994	-0.640	1.124	-0.122	-0.193	-0.323	0.304

Cuadro 4.13: Correlación canónica y vectores canónicos de los datos de carbono.

Error 1	Error 2
-0.058	-0.005

Cuadro 4.14: Errores en las restricciones de los datos de carbono.

Mediante este programa se logró obtener el valor de la correlación canónica, el cuál es 0,5741, que nos indica que los grupos de variables están correlacionados pero el grado de relación no es fuerte. Se observa además, en el primer vector canónico, en el primer grupo, las variables que más aportan información son densidad aparente, porcentaje de arena y porcentaje de limo con signo negativo, mientras que para el segundo grupo solamente el porcentaje de materia

orgánica con signo positivo. El resultado refleja cómo en estos suelos con elevada producción de materia orgánica se presenta con densidades aparentes bajas.

4.3. Comparaciones del ACC no lineal usando KTA y AG

Caso datos de cabezas

A continuación se presenta la aplicación del método a los datos de cabezas.
Como ya se sabe, cada grupo tiene dos variables, entonces el cromosoma tendrá cinco genes representando a ρ_a y a los vectores propios **u** y **v** con dos entradas.
Se calibró el algoritmo, obteniéndose los siguientes parámetros
Número de generaciones = 200
Tamaño de la población = 70
Razón de mutación = 0.26
Razón de cruza = 0.76
Una vez calibrado el programa, éste se corrió para obtener la correlación canónica y los vectores canónicos usando la función KTA con solución mediante el Gradiente Descendente (GD) y AG, estos resultados son presentados en el Cuadro 4.15.

Método		Vectores canónicos				Error 1	Error 2
ACCK-KTA	ρ_a	**u**		**v**			
GD	0.9330	0.0914	0.9958	0.9847	-0.1738	2.523e-09	2.803e-10
AG	0.9332	0.0907	0.9967	0.9913	-0.1776	7.984e-4	0.0071

Cuadro 4.15: Correlaciones y vectores canónicos de los datos de cabezas.

Se puede observar que los resultados para ambos métodos son similares. Si se observan los valores de ρ_a, el del método AG es levemente mas grande que el del GD, sin embargo, los errores en el método GD son más pequeños que los del AG.
A continuación se presentan en la Figura 4.1 y 4.2 los datos proyectados para el KTA GD y KTA AG, respectivamente.
Además, se presentan los valores de la correlación de Spearman de las parejas de proyecciones usando cada uno de los métodos, estos valores se encuentran en el Cuadro 4.16

Método	Spearman
ACCK-KTA	ρ
GD	0.6308
AG	0.6309

Cuadro 4.16: Correlaciones de Spearman de los datos proyectados.

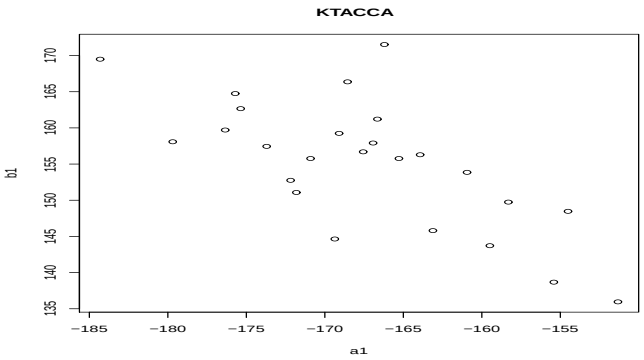


Figura 4.1: Proyección de datos KTA GD del ejemplo de cabezas.

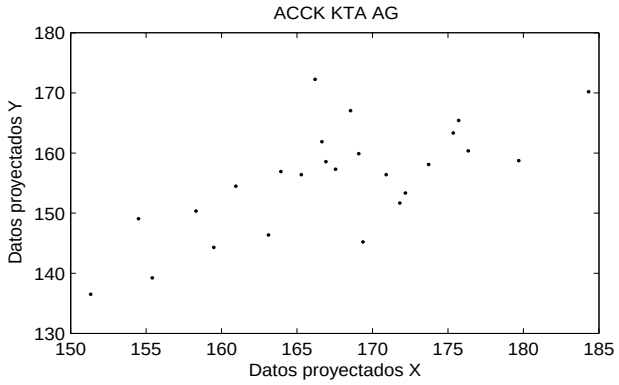


Figura 4.2: Proyección de datos KTA AG del ejemplo de cabezas.

Caso datos simulados

En [5] se presenta una simulación, de fuentes de señales generadas artificialmente, junto con algunas variables ruido, se presenta además pruebas del rendimiento de KCCA y diferentes variantes del Análisis de Correlación Canónica (ACC) para recuperar las señales de fuente en presencia de ruido y variables

irrelevantes. Las variables generadas son:

$$\begin{aligned} z &\sim uniform(-\pi, \pi) \\ \epsilon_{\mathbf{x}_1}, \epsilon_{\mathbf{y}_1} &\sim N(0, \sigma = 0{,}1), \\ \epsilon_{\mathbf{y}_2}, \epsilon_{\mathbf{y}_3} &\sim N(0, \sigma = 0{,}5), \\ \mathbf{x}_1 &= sin(z) + \epsilon_{\mathbf{x}_1}, \\ \mathbf{x}_2 &\sim N(0, 1), \\ \mathbf{x}_3 &\sim N(0, 1), \\ \mathbf{x}_4 &\sim N(0, 1), \\ \mathbf{x}_5 &\sim N(0, 1), \\ \mathbf{y}_1 &= cos(z) + \epsilon_{\mathbf{y}_1}, \\ \mathbf{y}_2 &= cos(\mathbf{x}_2 + \mathbf{x}_3) + \epsilon_{\mathbf{y}_2}, \\ \mathbf{y}_2 &= \mathbf{x}_4 + \epsilon_{\mathbf{y}_3}, \\ \mathbf{y}_4 &\sim N(0, 1). \end{aligned}$$

Se generaron $n = 100$ muestras para cada una de las variables. El primer grupo de variables es $X = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5)'$ y el segundo $Y = (\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4)'$.
Se calibró el algoritmo, obteniéndose mejores resultados con los siguientes parámetros
Número de generaciones = 200
Tamaño de la población = 70
Razón de mutación = 0.27
Razón de cruza = 0.66
Se llevó a cabo el ACCK usando la función KTA para dar solución mediante el método de gradiente descendiente y el método de AG. Los resultados son presentados en el Cuadro 4.20.

Método KTA		Vectores canónicos						Error
GD	0.4155	**u**	0.0035	-0.0508	-0.2127	0.9710	0.0955	4.508e-10
		v	0.0759	0.1399	-0.9855	-0.0575		-6.106e-10
AG	0.4197	**u**	0.0207	0.1068	0.0552	-0.6930	0.0247	0.2959
		v	0.0210	0.0758	-0.9504	0.0533		0.0448

Cuadro 4.17: Correlaciones y vectores canónicos de los datos simulados.

Nuevamente, los resultados en ambos métodos son similares. Si se observan los valores de ρ_a, del método AG es levemente mas grande que el de GD y en cuanto a los errores, en el método AG no son tan pequeños como los del GD.
A continuación se presentan en la Figura 4.3 y la Figura 4.4 los datos proyectados para el KTA GD y KTA AG, respectivamente.
Además, se presentan los valores de la correlación de Spearman de las parejas de proyecciones usando de cada uno de los métodos, estos valores se encuentran en el Cuadro 4.18

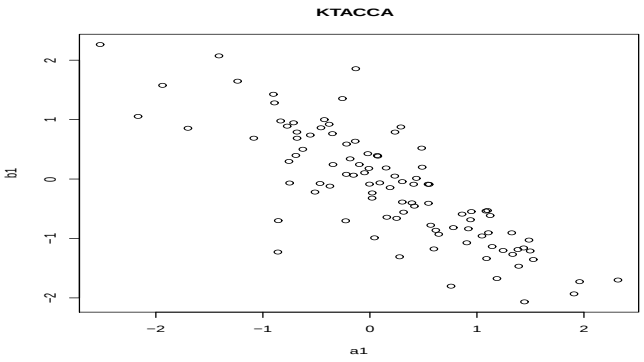


Figura 4.3: Proyección de datos KTA GD del ejemplo de datos simulados.

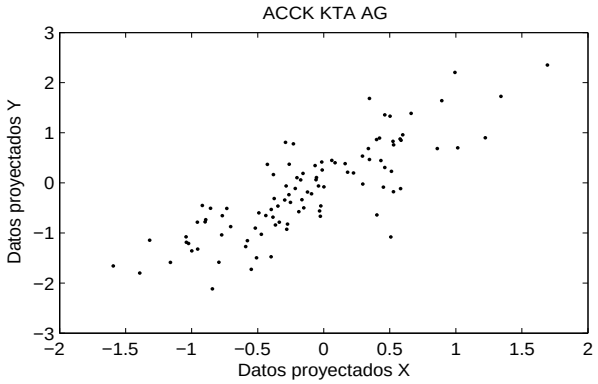


Figura 4.4: Proyección de datos KTA AG del ejemplo de datos simulados.

Método	Spearman
ACCK-KTA	ρ
GD	0.8060
AG	0.8225

Cuadro 4.18: Correlaciones de Spearman de los datos proyectados.

Caso datos medicamentos

Se cuenta con una base de datos tomados en un consultorio de la Ciudad de Puebla. La información obtenida es de 292 pacientes pediátricos y se requiere estudiar las relaciones entre los grupos de variables presentados en el Cuadro 4.19.

Primer grupo de variables (X)	Segundo grupo de variables (Y)
Edad	Número de enfermedades
Peso	Número de medicamentos
Talla	Número de fármacos

Cuadro 4.19: Variables del estudio de interacciones medicamentosas

Se calibró el algoritmo, obteniéndose mejores resultados con los siguientes parámetros
Número de generaciones = 50
Tamaño de la población = 10
Razón de mutación = 0.089
Razón de cruza = 0.685
Se llevó a cabo el ACCK usando la función KTA para dar solución mediante el método de gradiente descendiente y el método de algoritmos genéticos. Los resultados son presentados en el siguiente Cuadro 4.20.

Método KTA		Vectores canónicos				Error
GD	0.1303	**u**	0.9285	-0.0231	0.3704	-1.245e-09
		v	-0.5109	-0.6399	-0.5739	-3.081e-08
AG	0.1301	**u**	0.8640	0.0432	0.5014	1.0614e-4
		v	0.4683	0.5803	0.6662	4.7248e-5

Cuadro 4.20: Correlaciones y vectores canónicos de los datos de medicamentos.

Nuevamente, los resultados en ambos métodos son similares, si se observan los errores, nuevamente en el método de algoritmos genéticos los errores en las reestricciones no son muy pequeños comparados con los del método de gradiente descendente.
A continuación se presentan en la Figura 4.5 y la Figura 4.6 los datos proyectados para el KTA GD y KTA AG, respectivamente.
Además, se presentan los valores de la correlación de Spearman de las parejas de proyecciones usando de cada uno de los métodos, éstos valores se encuentran en el Cuadro 4.21

Método	Spearman
ACCK-KTA	ρ
GD	0.1248
AG	0.1305

Cuadro 4.21: Correlaciones de Spearman de los datos proyectados.

En conclusión, se realizó una comparación entre los resultados de los tres métodos de optimización para resolver el ACC usando diferentes ejemplos, obteniéndose que mediante el método que usa factorización QR se tiene más precisión en

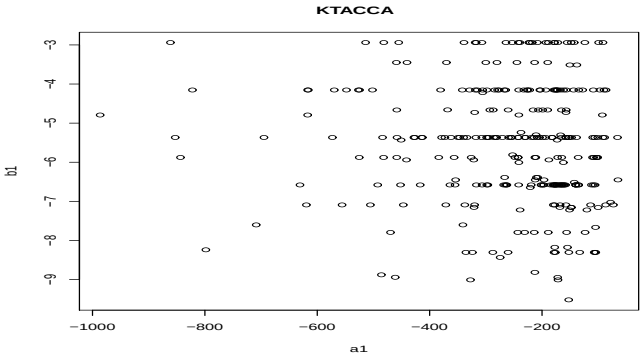


Figura 4.5: Proyección de datos KTA GD del ejemplo de medicamentos.

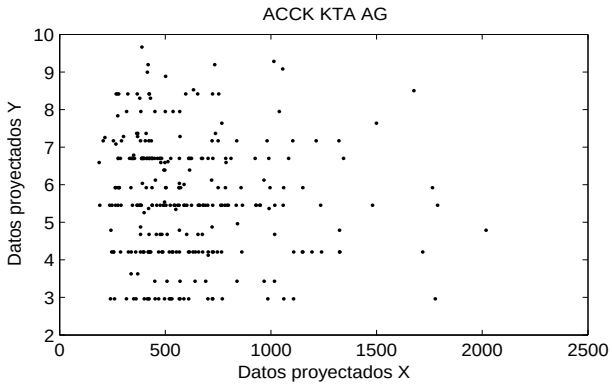


Figura 4.6: Proyección de datos KTA AG del ejemplo de medicamentos.

los resultados, es decir, el error en las restricciones es pequeño en comparación con los otros métodos de optimización.

Además, se hizo un estudio comparativo del ACC con AG, primero visto el problema como un problema de valores y vectores propios simple, nuevamente se realizaron comparaciones entre los tres métodos del ACC sin usar AG y el que se obtuvo al usar algoritmos genéticos, en cada uno de los ejemplos se obtuvo que el ACC AG obtiene valores de las correlaciones canónicas ligeramente más grandes en comparación con los otros métodos.

Para el caso del estudio de secuestro de Carbono debido a que una de sus matrices de varianzas y covarianzas no es invertible, para ello se realizó un programa que obtuviera las correlaciones canónicas de manera directa de la definición para que no se requiriera el cálculo de las matrices inversas de las matrices de varianzas y covarianzas, de igual forma se usó el método de algoritmos genéti-

cos. Mediante este programa se obtuvieron solamente las primeras correlaciones canónicas para cada uno de los ejemplos y principalmente se obtuvo el valor de la correlación canónica del ejemplo de datos de Carbono.

Capítulo 5

Conclusiones

En este trabajo de tesis se mostró que el Análisis de Correlación Canónica (ACC), que surgió para determinar la relación lineal entre dos grupos de variables ha sido extendido, en base al comportamiento de los datos, a un número mayor de grupos de variables, así como cuando se tiene un número menor de individuos que el de variables. Por otra parte, se desarrolló el caso cuando los datos ya no presentan relaciones lineales, mediante el uso del Análisis de Correlación con Kernel (ACCK), que transforma los datos y los envía a un nuevo espacio de alta dimensión para poder manejarlos desde otra perspectiva y poder obtener mejores resultados de las correlaciones canónicas. Se incluyó el estudio de los métodos que mejoraron al ACCK, que son el Criterio de Independencia de Hilbert-Schmidt (HSIC) y el Criterio de Alineación Objetivo del Kernel Centrado (KTA), ya que mediante estos métodos se obtiene la interpretación de las variables canónicas. Una nueva forma de determinar la solución del ACC mediante el método de Algorítmos Genéticos resultó interesante, en el sentido de que la solución se encuentra directamente de la definición de correlación canónica, sin realizar diferentes cálculos o usar diferentes métodos para obtener tanto la correlación canónica como los vectores canónicos.
A continuación se exponen los resultados obtenidos en este trabajo:

- Bajo el supuesto de linealidad
  - Se profundizó en el estudio de datos de manglares del sistema lagunar en el estudio se realizó un analisis usando ACC y sus extensiones.
  - Se realizó un programa para cada uno de los métodos de solución al ACC, se realizaron comparaciones entre ellos así como la relación entre cada uno de ellos.
- Bajo el supuesto de no linealidad.
  - Se presentaron los métodos kernel que dan lugar al llamado ACCK, así como otros métodos no lineales que dan solución al ACC no lineal como los son el Criterio de Independencia de Hilbert Schmidt (HSIC) y el criterio de Alineación Objetivo de Kernel Centrado (KTA)

- Se estudiaron la relación entre la calidad de vida de niños con interacciones medicamentosas en el momento que estaban enfermos. Principalmente se usó el ACCK y el ACC no lineal para obtener que si hay una fuerte relación entre las variables de cada grupo.
- Se introdujo el método de algoritmos genéticos que se presentó como un método para resolver el ACC mediante valores y vectores propios y para resolver el ACC de manera directa usando la definición de esta. Además, se usó este método para obtener otro método alternativo para resolver el ACC no lineal usando los algoritmos genéticos de forma directa sobre la definición de KTA. Se desarrollaron los programas correspondientes
- Se realizó un programa usando el método de AG para resolver el ACC no lineal, basándose en la definición del valor ρ del método KTA pero obteniendo este valor directamente de la definición. Se realizaron comparaciones entre los resultados obtenidos mediante la función HsicCCA en R y los resultados obtenidos mediante algoritmos genéticos, concluyéndose que mediante ambos métodos se logra obtener los mismos valores para el valor ρ, además, se obtuvieron las correlaciones de las proyecciones de los datos mediante la correlación de Spearman y de igual forma se obtuvieron valores similares.

El ACC continúa ofreciendo un campo de investigación promisorio para los estadísticos. La investigación sobre medidas de asociación no lineales y el empleo de pruebas de hipótesis de independencia, usando métodos kernel, es un tema de gran actualidad, dado que el ACCK no requiere la suposición de normalidad de las observaciones y por lo tanto, permite una gran aplicabilidad. Otra vía abierta de investigación es la aplicación de redes neuronales mediante el Análisis de Correlación Canónica Profundo (DCCA).

Apéndice A

Algunas herramientas matemáticas

A continuación se presentan algunos teoremas y definiciones que son de gran utilidad en la parte teórica tanto en la presentación del ACC como del ACCK. En primera instancia se mencionan algunas definiciones y teoremas algebraicos y posteriormente se introduce material para definir a un espacio de Hilbert. La información presentada es basada principalmente en Kreyszig [17] y Mardia [23].

A.1. Propiedades algebraicas

Definición A.1.1 *Un **espacio vectorial (o espacio lineal)** sobre un campo* $\mathbf{K}$ *es un conjunto no vacío* $\mathbf{X}$ *de vectores con una operación binaria de adición* $+ : \mathbf{X} \times \mathbf{X} \to \mathbf{X}$ *y una multiplicación escalar* $\cdot : \mathbf{K} \times \mathbf{X} \to \mathbf{X}$*, que satisfacen:*

1. *Si* $x, y \in \mathbf{X}$ *entonces* $x + y \in \mathbf{X}$.
2. *Para todo* $x, y, z \in \mathbf{X}$, $(x + y) + z = x + (y + z)$.
3. *Existe un vector* $0 \in \mathbf{X}$ *tal que para todo* $x \in \mathbf{X}$, $x + 0 = 0 + x = x$.
4. *Si* $x \in \mathbf{X}$, *existe un vector* $-x \in \mathbf{X}$ *tal que* $x + (-x) = 0$.
5. *Si* $x, y \in \mathbf{X}$, *entonces* $x + y = y + x$.
6. *Si* $x \in \mathbf{X}$ *y* α *es un escalar, entonces* $\alpha x \in \mathbf{X}$.
7. *Si* $x, y \in \mathbf{X}$ *y* α *es un escalar, entonces* $\alpha(x + y) = \alpha x + \alpha y$.
8. *Si* $x \in \mathbf{X}$ *y* α *y* β *son escalares, entonces* $(\alpha + \beta)x = \alpha x + \beta x$.
9. *Si* $x \in \mathbf{X}$ *y* α *y* β *son escalares, entonces* $\alpha(\beta x) = (\alpha\beta)x$.
10. *Para cada vector* $x \in \mathbf{X}$, $1x = x$.

Definición A.1.2 *Un subconjunto* $\mathbf{Y}$ *del espacio vectorial* $\mathbf{X}$*, es un* ***subespacio*** *de* $\mathbf{X}$ *si es cerrado bajo las operaciones lineales. Esto es,* $x + y \in \mathbf{Y}$ *para toda* $x, y \in \mathbf{Y}$ *y* $\alpha x \in \mathbf{Y}$ *para cada* $\alpha \in \mathbf{K}$ *y* $x \in \mathbf{Y}$*.*

Un subespacio especial de $\mathbf{X}$ es el subespacio impropio $\mathbf{Y} = \mathbf{X}$. Cualquier otro subespacio de $\mathbf{X}(\neq \{0\})$ es llamado subespacio propio.

Definición A.1.3 *Una* ***combinación lineal*** *de vectores* $x_1, x_2, ..., x_m$ *de un espacio vectorial* $\mathbf{X}$ *es una expresión de la forma*

$$\alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_m x_m$$

donde los coeficientes $\alpha_1, \alpha_2, ..., \alpha_m$ *son algunos escalares.*

Definición A.1.4 *Para algún subconjunto no vacío* $M \subset \mathbf{X}$*, el conjunto de todas las combinaciones lineales de vectores de* M *es llamado el generado de* M*.*

Definición A.1.5 *Independencia y dependencia de un conjunto* M *de vectores* $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_r$ $(r \geq 1)$ *en un espacio vectorial* $\mathbf{X}$ *son definidas mediante la ecuación*

$$\alpha_1 \mathbf{x}_1 + \alpha_2 \boldsymbol{x}_2 + ... + \alpha_r \mathbf{x}_r = 0 \tag{A.1}$$

donde $\alpha_1, \alpha_2, ..., \alpha_r \in \mathbf{K}$*. La ecuación (A.1) se cumple cuando* $\alpha_1 = \alpha_2 = ... = \alpha_r = 0$*. Si ésta es la única* r*-tupla de escalares para la cual (A.1) se cumple, el conjunto* M *se dice ser* ***linealmente independiente****.* M *se dice ser* ***linealmente dependiente*** *si* M *no es linealmente independiente, ésto es, si (A.1) se cumple para alguna* r*-tupla de escalares, no todas cero.*

Definición A.1.6 *Un espacio vectorial se dice ser* ***finito dimensional*** *si existe un número entero positivo* n *tal que* $\mathbf{X}$ *contiene un conjunto linealmente independiente de* n *vectores, mientras algún conjunto de* $n + 1$ *o más vectores de* $\mathbf{X}$ *es linealmente dependiente,* n *es llamada la* ***dimensión*** *de* $\mathbf{X}$*,* $n = dim\mathbf{X}$*.*

Definición A.1.7 *Un conjunto* B *de vectores se dice ser una* ***base*** *para un espacio vectorial* $\mathbf{X}$ *si* B *genera* $\mathbf{X}$ *y los elementos de* B *forman un conjunto linealmente independiente.*

Definición A.1.8 *El* ***rango*** *de una matriz* $\mathbf{A}$ *de orden* $n \times p$ *es definido como el número máximo de filas (columnas) linealmente independientes en* $\mathbf{A}$*. Se denota el rango de* $\mathbf{A}$ *como* $rank(\mathbf{A})$*.*

Las siguientes propiedades se cumplen

Propiedades A.1.1

1. $0 \leq rank(\mathbf{A}) \leq min(n, p)$.
2. $rank(\mathbf{A}) = rank(\mathbf{A}')$.
3. $rank(\mathbf{A} + \mathbf{B}) \leq rank(\mathbf{A}) + rank(\mathbf{B})$.

4. $rank(\mathbf{AB}) \leq min\{rank(\mathbf{A}), rank(\mathbf{B})\}$.
5. $rank(\mathbf{A'A}) = rank(\mathbf{AA'}) = rank(\mathbf{A})$.
6. *Si* $\mathbf{B}$ *de orden* $n \times n$ *y* $\mathbf{C}$ *de orden* $p \times p$ *son no singulares entonces* $rank(\mathbf{BAC}) = rank(\mathbf{A})$.
7. *Si* $n = p$ *entonces* $rank(\mathbf{A}) = p$ *si y sólo si* $\mathbf{A}$ *es no singular.*

Definición A.1.9 *Sea* $\mathbf{A}$ *una matriz cuadrada de orden* $p \times p$*, sea* $\mathbf{I}$ *la matriz identidad de orden* $p \times p$ *entonces los escalares* $\lambda_1, \lambda_2, ..., \lambda_p$ *que satisfacen la ecuación polinomial* $|\mathbf{A} - \lambda\mathbf{I}|$ *son llamados los* ***valores propios (eigenvalores o raices características)*** *de la matriz* $\mathbf{A}$*. La ecuación* $|\mathbf{A} - \lambda\mathbf{I}|$ *(que es función de lambda) es llamada la* ***ecuación característica****.*

Definición A.1.10 *Sea* λ *un valor propio y sea* $\mathbf{x}$ *un vector no cero tal que*

$$\mathbf{Ax} = \lambda\mathbf{x}$$

entonces $\mathbf{x}$ *se dice ser un* ***vector propio (eigenvector o vector característico)*** *de la matriz* $\mathbf{A}$ *asociado al valor propio* λ*.*

Definición A.1.11 *Un vector propio* $\mathbf{x}$ *con entradas reales es llamado* ***estandarizado*** *si*

$$\mathbf{x'x} = 1.$$

Teorema A.1.1 *Para* $\mathbf{A}$ *de orden* $n \times p$ *y* $\mathbf{B}$ *de orden* $p \times n$*, los valores propios no cero de* $\mathbf{AB}$ *y* $\mathbf{BA}$ *son los mismos y tienen la misma multiplicidad. Si* $\mathbf{x}$ *es un vector propio no trivial de* $\mathbf{AB}$ *para un valor propio* $\lambda \neq 0$*, entonces* $\mathbf{y} = \mathbf{Bx}$ *es un vector propio no trivial de* $\mathbf{AB}$*.*

Definición A.1.12 *Una matriz* $\mathbf{A}$ *es* ***simétrica*** *si* $\mathbf{A} = \mathbf{A'}$*.*

Si $\mathbf{A}$ es una matriz simétrica, se puede dar información detallada acerca de sus vectores y valores propios.

Teorema A.1.2 *Todos los valores propios de una matriz simétrica* $\mathbf{A}$ *de orden* $p \times p$ *son reales*

Teorema A.1.3 *De descomposición espectral. Sea* $\mathbf{A}$ *una matriz simétrica* $p \times p$*. Entonces* $\mathbf{A}$ *puede ser expresada en términos de sus* p *parejas de valores y vectores propios* $(\lambda_i, \mathbf{e}_i)$ *como*

$$\mathbf{A} = \Gamma\mathbf{D}\Gamma' = \sum_{i=1}^{k} \lambda_i \mathbf{x}_i \mathbf{x}_i'$$

donde $\mathbf{D}$ *es una matriz diagonal con los valores propios de* $\mathbf{A}$ *y* Γ *es una matriz ortogonal cuyas columnas son vectores propios estandarizados.*

El rango de $\mathbf{A}$ es igual al número de valores propios diferentes de cero.

Definición A.1.13 *Sean* $\mathbf{a}$ *y* $\mathbf{b}$ *vectores del mismo tamaño,* $\mathbf{a}$ *y* $\mathbf{b}$ *se dicen* ***ortogonales*** *si* $\mathbf{a}'\mathbf{b} = 0$. *Si* $\mathbf{a}'\mathbf{a} = 1$, *se dice que* $\mathbf{a}$ *es* ***normalizado***.

Definición A.1.14 *Una matriz* $\mathbf{A}$ *es* ***ortogonal*** *si sus columnas son normalizadas y mutuamente ortogonales.*

Propiedades A.1.2 *Si* $\mathbf{A}$ *es una matriz ortogonal cuadrada entonces se cumple que* $\mathbf{AA}' = \mathbf{A}'\mathbf{A} = \mathbf{I}$ *o* $\mathbf{A}' = \mathbf{A}^1$.

Las ideas que guían a la descomposición espectral puede ser extendida para dar una descomposición para una matriz rectangular en lugar de una matriz cuadrada. Si $\mathbf{A}$ es una matriz rectangular, entonces los vectores en la expansión de $\mathbf{A}$ son vectores propios de las matrices cuadradas $\mathbf{AA}'$ y $\mathbf{A}'\mathbf{A}$.

Teorema A.1.4 *Descomposición en Valores Singulares (SVD). Sea* $\mathbf{A}$ *una matriz de números reales* $q \times p$. *Entonces existe una matriz ortogonal* $\mathbf{U}$ *de orden* $q \times q$ *y una matriz ortogonal* $\mathbf{V}$ *de orden* $p \times p$ *tal que*

$$\mathbf{A} = \mathbf{UDV}'$$

donde la matriz $\mathbf{D}$ *de orden* $q \times p$ *es una matriz diagonal con entradad* λ_i, *con* $i = 1, 2, ..., min(p, q)$.

Teorema A.1.5 *Factorización (descomposición) QR. Sea* $\mathbf{A}$ *una de orden* $q \times p$ *y de rango* p, *entonces existe una matriz* $\mathbf{Q}$ *de orden* $q \times q$ *ortogonal y una matriz* $\mathbf{R}$ *de orden* $p \times p$ *triangular superior tal que* $\mathbf{A} = \mathbf{QR}$.

Definición A.1.15 *Una* ***forma cuadrática*** $Q(\mathbf{x})$ *en* p *variables* $x_1, x_2, ..., x_p$ *es* $Q(\mathbf{x}) = \mathbf{x}'\mathbf{Ax}$, *donde* $\mathbf{x}' = [x_1, x_2, ..., x_p]$ *y* $\mathbf{A}$ *es una matriz simétrica* $p \times p$.

Definición A.1.16

$Q(\mathbf{x})$ *es llamada una* ***forma cuadrática definida positiva*** *si* $Q(\mathbf{x}) > 0$ *para todo* $\mathbf{x} \neq 0$

$Q(\mathbf{x})$ *es llamada una* ***forma cuadrática semidefinida positiva*** *si* $Q(\mathbf{x}) \geq 0$ *para todo* $\mathbf{x} \neq 0$

Una matriz simétrica es llamada definida positiva (semidefinida positiva) si $Q(\mathbf{x})$ *es definida positiva (semidefinida positiva) y escribiremos* $\mathbf{A} > 0$ *o* $\mathbf{A} \geq 0$ *para* $\mathbf{A}$ *definida positiva o semidefinida positiva, respectivamente.*

Teorema A.1.6 *Descomposición de Cholesky. Si* $\mathbf{A}$ *de orden* $p \times p$ *es una matriz definida positiva y simétrica, entonces existe una única matriz triangular superior con elementos positivos en la diagonal* ($u_{ii} > 0$), *tal que* $\mathbf{A} = \mathbf{U}'\mathbf{U}$.

Además existe una única matriz trangular inferior $\mathbf{L}$ con elementos positivos en la diagonal tal que $\mathbf{A} = \mathbf{LL}'$.

A.2. Espacio de Hilbert

Definición A.2.1 *Un* ***espacio métrico*** *es una pareja* $(\mathbf{X}, d)$*, donde* $\mathbf{X}$ *es un conjunto y* d *es una* ***métrica*** *(o una función distancia en* $\mathbf{X}$*), esto es, una función definida en* $\mathbf{X} \times \mathbf{X}$ *tal que para todo* $x, y, z \in \mathbf{X}$ *se tiene*

- d *es de valor real, finita y no negativa.*
- $d(x, y) = 0$ *sí y sólo si* $x = y$*.*
- $d(x, y) = d(y, x)$*.*
- $d(x, y) \leq d(x, z) + d(z, y)$*.*

Definición A.2.2 *Un* ***subespacio*** $(\mathbf{Y}, \tilde{d})$ *de* $(\mathbf{X}, d)$ *es obtenido si se toma un subconjunto* $\mathbf{Y} \subset \mathbf{X}$ *y una restricción* $\tilde{d}$ *a* $\mathbf{Y} \times \mathbf{Y}$*; entonces la métrica en* $\mathbf{Y}$ *es la restricción*

$$\tilde{d} = d|_{\mathbf{Y} \times \mathbf{Y}}$$

$\tilde{d}$ *es llamada la métrica* ***inducida*** *en* $\mathbf{Y}$ *por* d*.*

Definición A.2.3 *Una sucesión* $\{x_n\}$ *en un espacio métrico* $\mathbf{X}$ *se dice que* ***converge o es convergente*** *si existe un* $x \in \mathbf{X}$ *tal que*

$$lim_{n \to \infty} d(x_n, x) = 0$$

x *es llamado el* ***límite*** *de* $\{x_n\}$*.*

Definición A.2.4 *Una sucesión* $\{x_n\}$ *en un espacio métrico* $\mathbf{X}$ *se dice ser de* ***Cauchy*** *si para todo* $\epsilon > 0$ *existe* $N = N(\epsilon)$ *tal que*

$$d(x_m, x_n) < \epsilon$$

para todo $m, n > N$*. El espacio* $\mathbf{X}$ *se dice ser* ***completo*** *si toda sucesión de Cauchy en* $\mathbf{X}$ *converge.*

Definición A.2.5 *Una* ***norma*** *en un espacio vectorial* $\mathbf{X}$ *es una función de valor real en* $\mathbf{X}$ *cuyo valor* $x \in \mathbf{X}$ *es denotado por* $\|x\|$*.*

La norma tiene las siguientes propiedades.

Propiedades A.2.1

- $\|x\| \geq \mathbf{0}$
- $\|x\| = \mathbf{0} \Longleftrightarrow x = 0$
- $\|\alpha x\| = |\alpha| \, \|x\|$
- *Desigualdad del triangulo.* $\|x + y\| \leq \|x\| + \|y\|$

Definición A.2.6 *Un* ***espacio normado*** $\mathbf{X}$ *es un espacio vectorial con una norma definida en el. Un* ***espacio de Banach*** *es un espacio normado completo.*

Definición A.2.7 *Una sucesión* $\{x_n\}$ *en un espacio normado* $\mathbf{X}$ *es llamada* ***sucesión de Cauchy*** *si para todo* $\epsilon > 0$ *existe un entero* N *tal que* $\|x_m - x_n\| < \epsilon$ *para todo* $m, n \geq N$*. Un espacio normado* $\mathbf{X}$ *es* ***completo*** *si toda sucesión de Cauchy es convergente.*

Definición A.2.8 *Un* ***producto interno*** *en* $\mathbf{X}$ *es un mapeo de* $\mathbf{X} \times \mathbf{X}$ *en el campo escalar* $\mathbf{K}$ *de* $\mathbf{X}$*; esto es, para cada par de vectores* x *y* y *existe un escalar asociado el cual es escrito como* $\langle x, y \rangle$ *es llamado el producto interno de* x *y* y*.*

El producto interno tiene las siguientes propiedades.

Propiedades A.2.2 *Para todos los vectores* $x, y, z \in X$ *y* $\alpha \in \mathbf{K}$

- $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$
- $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$
- $\langle x, y \rangle = \langle \bar{y, x} \rangle$
- $\langle x, x \rangle \geq \mathbf{0}$
  $\langle x, x \rangle = \mathbf{0} \Longleftrightarrow x = \mathbf{0}$

Definición A.2.9 *Un* ***espacio de producto interno*** *es un espacio vectorial* $\mathbf{X}$ *con un producto interno definido en* $\mathbf{X}$*. Un* ***espacio de Hilbert*** *es un espacio de producto interno completo.*

Apéndice B

Uso de R para el ACC

En los primeros dos capítulos de esta tesis se presentan ejemplos en donde se hace uso de algunos paquetes del software R, a continuación se presenta un esboce de las funciones utilizadas para el ACC, ACCR, ACCRG y ACCK.

B.1. Paquete CCA para el ACC clásico y regularizado

Para poder trabajar con los datos en R sin tener que importar una hoja de EXCEL se pueden realizar el copiado de la siguiente forma. En la hoja de excel se seleccionan los datos a ser utilizados (Figura B.1) y se selecciona la opción copiar (CTRL+C).
Sin realizar algún movimiento más, en la consola de R se teclea el siguiente comando

$$manglares1 < -read.table(file ='' clipboard'', head = T)$$

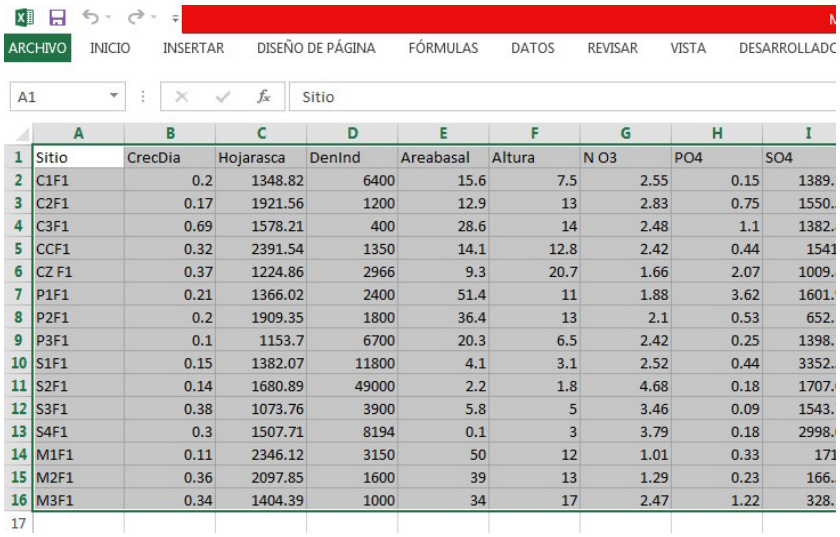
mediante el cual se está ingresando toda la información de los datos, posteriormente, si se escribe

$$manglares1$$

que es el nombre asignado a los datos, en la pantalla se despliega la información copiada como en la Figura B.2
Para trabajar el ACC clásico se hizo uso de la función cancor y el paquete CCA (que debe ser instalado y cargado para poder usarse) en R.
Como un análisis previo a las variables de los datos, para obtener los valores de las correlaciones entre ellas, se teclea

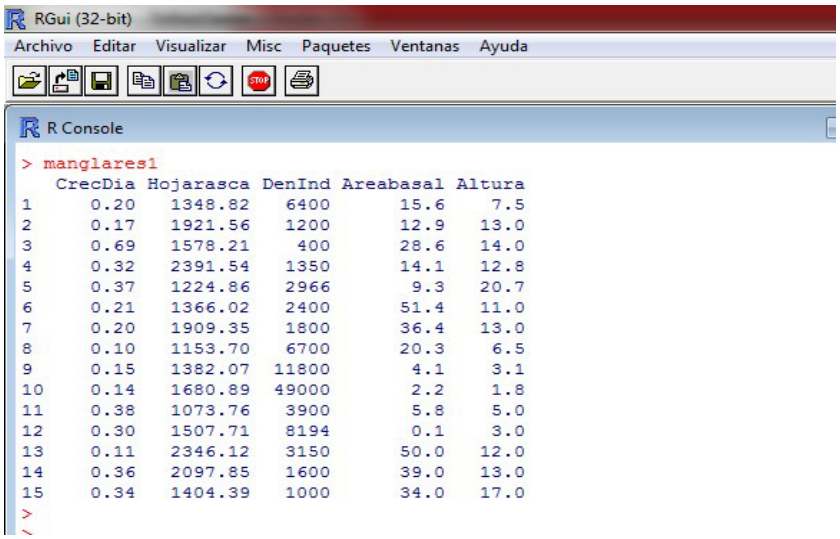
$$cor(mang1, mang2)$$

y se desplegarán los valores de las correlaciones entre las variables en forma de matriz. Estos valores pueden ser graficados para mejor entendimiento, mediante



	A	B	C	D	E	F	G	H	I
1	Sitio	CrecDia	Hojarasca	DenInd	Areabasal	Altura	N O3	PO4	SO4
2	C1F1	0.2	1348.82	6400	15.6	7.5	2.55	0.15	1389.
3	C2F1	0.17	1921.56	1200	12.9	13	2.83	0.75	1550.
4	C3F1	0.69	1578.21	400	28.6	14	2.48	1.1	1382.
5	CCF1	0.32	2391.54	1350	14.1	12.8	2.42	0.44	1541
6	CZ F1	0.37	1224.86	2966	9.3	20.7	1.66	2.07	1009.
7	P1F1	0.21	1366.02	2400	51.4	11	1.88	3.62	1601.
8	P2F1	0.2	1909.35	1800	36.4	13	2.1	0.53	652.
9	P3F1	0.1	1153.7	6700	20.3	6.5	2.42	0.25	1398.
10	S1F1	0.15	1382.07	11800	4.1	3.1	2.52	0.44	3352.
11	S2F1	0.14	1680.89	49000	2.2	1.8	4.68	0.18	1707.
12	S3F1	0.38	1073.76	3900	5.8	5	3.46	0.09	1543.
13	S4F1	0.3	1507.71	8194	0.1	3	3.79	0.18	2998.
14	M1F1	0.11	2346.12	3150	50	12	1.01	0.33	171
15	M2F1	0.36	2097.85	1600	39	13	1.29	0.23	166.
16	M3F1	0.34	1404.39	1000	34	17	2.47	1.22	328.
17									

Figura B.1: Datos de manglares



```
> manglares1
   CrecDia Hojarasca DenInd Areabasal Altura
1     0.20   1348.82   6400       15.6    7.5
2     0.17   1921.56   1200       12.9   13.0
3     0.69   1578.21    400       28.6   14.0
4     0.32   2391.54   1350       14.1   12.8
5     0.37   1224.86   2966        9.3   20.7
6     0.21   1366.02   2400       51.4   11.0
7     0.20   1909.35   1800       36.4   13.0
8     0.10   1153.70   6700       20.3    6.5
9     0.15   1382.07  11800        4.1    3.1
10    0.14   1680.89  49000        2.2    1.8
11    0.38   1073.76   3900        5.8    5.0
12    0.30   1507.71   8194        0.1    3.0
13    0.11   2346.12   3150       50.0   12.0
14    0.36   2097.85   1600       39.0   13.0
15    0.34   1404.39   1000       34.0   17.0
>
```

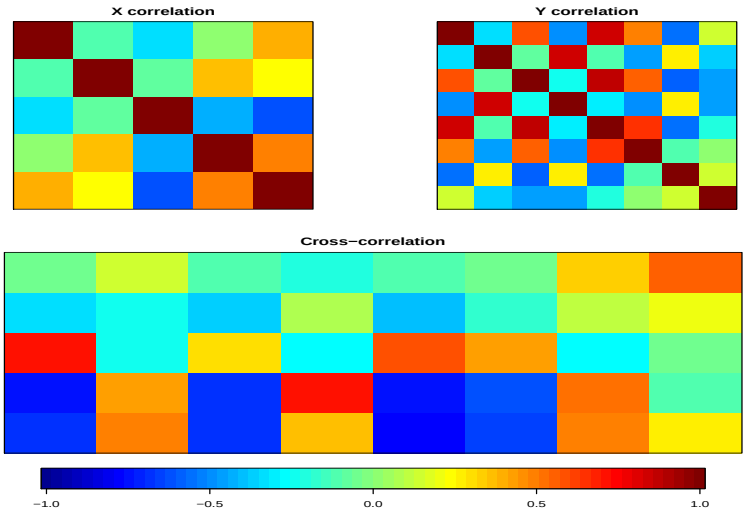
Figura B.2: Datos de manglares en R

los comandos

$$mcorrmang < -matcor(mang1, mang2)$$

$$img.matcor(mcorrmang, type = 2)$$

y se abrirá una pantalla nueva con la imagen siguiente



Para obtener las correlaciones canónicas se puede usar la función cancor

$$cancor(mang1, mang2)$$

o bien la función cc

$$corm < -cc(mang1, mang2)$$

.
Cabe señalar que con ambas funciones se obtienen tanto las correlaciones canónicas como los vectores canónicos, sin embargo, mediante la función cc se obtiene más información adicional como lo es los scores de las variables canónicas, se puede obtener el gráfico de barras de las correlaciones así como un gráfico en donde se presentan las dos primeras variables canónicas, mediante

$$barplot(corm\$cor, xlab = \text{“Dimensión”}, ylab = \text{“Correlaciones canónicas”}, ylim = c(0, 1))$$

$$plt.cc(corm, var.label = TRUE)$$

.

B.2. Paquete CCA para el ACC regularizado

Para realizar el ACCR se usó el paquete CCA presentado en la sección anterior. El uso es prácticamente similar, solo que antes de obtener las correlaciones y vectores canónicos es necesario obtener los valores de regularización, estos se obtienen al teclear

$$estim.regul(mang1, mang2),$$

supónga que estos valores son los siguientes $\lambda_x = 1$ y $\lambda_y = 0{,}25075$, entonces para obtener las correlaciones canónicas y variables canónicas se escribe

$$rcc(mang1, mang2, 1, 0{,}25075)$$

.

B.3. Paquete RGCCA para el ACC regularizado generalizado

Para realizar el ACCRG se usó el paquete RGCCA en R [41]. En primera instancia se introdujeron los datos, sin embargo, para poder trabajar con ellos adecuadamente es necesario convertirlos a una matriz, este proceso se lleva a cabo escribiendo

$$mg1 < -as.matrix(manglares1)$$

en donde manglares1 es el nombre original de los datos del primer grupo de variables y mg1 es el nombre de la matriz de datos.
Dado que en este trabajo se usaron tres grupos de variables, como siguiente paso se almacenan todas en una variable

$$Grupos = list(mg1, mg2, mg3)$$

y posteriormente se realiza el análisis especificando en la matriz de dise o C que se quiere determinar el grado de relación entre los grupos $mg1$ con $mg2$, $mg1$ con $mg3$ y viceversa.

$$rgcca(Grupos, C, tau = \text{“}optimal\text{”}, scheme = \text{“}centroid\text{”}, scale = TRUE)$$

Se observa que se eligió trabajar con el esquema centroide. Mediante esta orden se obtienen los valores de los coeficientes de cada una de las variables y finalmente se despliega una gráfica en la cual es más fácil apreciar que grupos de variables tienen una relación lineal más notable.

B.4. Paquete kernlab para el ACCK

Para llevar a cabo el ACCK se utilizó el paquete kernlab en R. Primero los datos deben escribirse en forma de matrices (usando el comando as.matrix). Posteriormente para realizar el ACCK con el kernel lineal se usa el comando siguiente

$$rmedker1 < -kcca(rmed1, rmed2, kernel = \text{"}vanilladot\text{"}, kpar = list(), ncomps = 6)$$

en donde $rmed$ representa los datos de los pacientes pediátricos. Ahora, mediante las siguientes instrucciones se obtienen las correlaciones y los coeficientes de las variables canónicas proyectadas

$$slot(rmedker1, \text{"}kcor\text{"})$$

$$a1 < -slot(rmedker1, "xcoef")$$
$$b1 < -slot(rmedker1, "ycoef")$$

y si se quieren graficar estas proyecciones se usa la instrucción

$$plot(a1[,1], b1[,1], main = \text{“Kernel lineal”}, xlab = "a1", ylab = "b1")$$

para cada una de las variables canónicas.
Para realizar el ACCK gaussiano se realiza

$$rmedker2 \quad < - \quad kcca(rmed1, rmed2, kernel = "rbfdot", kpar = list(sigma = 0{,}1), ncomps = 6)$$

en donde puede observar que es necesario introducir el valor de σ.
Para el ACCK polinomial se usa

$$rmedker3 \quad < - \quad kcca(rmed1, rmed2, kernel = "polydot", kpar = list(degree = 1, scale = 0{,}5, offset = 1), ncomps = 6)$$

y para el ACCK tangente hiperbólica

$$rmker3 \quad < - \quad kcca(rmed1, rmed2, kernel = "polydot", kpar = list(degree = 1, scale = 1, offset = 1), ncomps = 6)$$

en donde también es necesario introducir los valores para el grado y las constantes que se presentan en la definición.
Para obtener los valores de las correlaciones canónicas se realiza el mismo procedimiento que en el ACCK lineal así como para obtener las gráficas de las proyecciones.

B.5. Paquete CCA para HSIC y KTA

Para realizar el ACC no lineal se usa el paquete HsicCCA en R. Se continúa trabajando con los datos en forma de matrices, al igual que en la sección anterior. Posteriormente, si se requiere trabajar con el método HSIC, se usa la función hsicCCA como sigue:

$$rmedhsic1 \quad < - \quad hsicCCA(rmed1, rmed2, 2, sigmax = NULL, sigmay = NULL, numrepeat = 5, numiter = 100, reltolstop = 1e-04)$$

en donde rmed1 y rmed2 son los dos grupos de variables, el número 2 indica la cantidad de variables canónicas a obtener y además los valores de σ_x y σ_y son calculados mediante el programa usando el truco de la mediana. Un procedimiento similar es el que se realiza para el método KTA, usando la función ktaCCA de la siguiente manera

$$rmedkta1 \quad < - \quad ktaCCA(rmed1, rmed2, 2, sigmax = NULL, sigmay = NULL, numrepeat = 5, numiter = 100, reltolstop = 1e-04).$$

Apéndice C

Algoritmos usados para el ACC y ACCK

Para determinar las correlaciones canónicas y los vectores canónicos (valores y vectores propios) en el ACC clásico se propusieron tres métodos para obtener estos valores: mediante descomposición SVD, descomposición de Cholesky y descomposición QR. Para cada uno de estos métodos se realizó un programa en MATLAB y los algoritmos usados para cada método son presentados en los siguientes Cuadro C.1, Cuadro C.2 y Cuadro C.3, respectivamente.

Datos de entrada
Σxx Σyy Σxy $\Sigma yx = \Sigma xy'$
$K = \Sigma xx(-1/2) * \Sigma xy * \Sigma yy(-1/2)$ $[x, D, y] = svd(K)$ $lambda = diag(D)$ $a = \Sigma xx(-1/2) * x$ $b = \Sigma yy(-1/2) * y$
Datos de salida
λ valores propios a, b vectores propios

Cuadro C.1: Algoritmo usando descomposición SVD [23]

Datos de entrada
Σxx
Σyy
Σxy
$\Sigma yx = \Sigma xy'$
$L_1 = chol(\Sigma_{xx})$
$L_2 = chol(\Sigma_{yy)}$
$R_1 = L_1'$
$R_2 = L_2'$
$A = inv(R_1) * \Sigma xy * inv(\Sigma yy) * \Sigma yx * inv(R_1)'$
$B = inv(R_2) * \Sigma yx * inv(\Sigma xx) * \Sigma xy * inv(R_2)'$
$[V_1, \rho 1] = eig(A)$
$[V_2, \rho 2] = eig(B)$
$lambda = sqrt(\rho 1)$
$a = inv(L_1) * V_1$
$b = inv(L_2) * V_2$
Datos de salida
λ valores propios
a, b vectores propios

Cuadro C.2: Algoritmo usando descomposición de Cholesky [9]

En el Cuadro C.4 se encuentra el algorítmo para el método de ortogonalización parcial de Gram-Schmidt usado para descomponer una matriz en una multiplicación de matrices triangulares y en el Cuadro C.5 se presenta el algoritmo utilizado para el método ACCK.

Datos de entrada
X Y
% Generar datos centrados
$[nr, ncx] = size(X)$ $fork = 1 : ncx$ $media(k) = sum(X(:, k))/nr$ $X(:, k) = X(:, k) - media(k)$ end $[qx, Tx] = qr(X);$ $[mr, ncy] = size(Y)$ $fork = 1 : ncy$ $media(k) = sum(Y(:, k))/mr$ $Y(:, k) = Y(:, k) - media(k)$ end $[qy, Ty] = qr(Y);$
% Determinación de los rangos de Tx y TY
$dx = rank(Tx)$ $dy = rank(Ty)$ $diagy = eye(nr, dy)$ $qrqy = qy * diagy$
% Determinación de vectores y valores propios $qrqty = qrqy' * qx$ $qrqty = qrqty(:, 1 : dx)$ $[v, d, u] = svd(qrqty)$ $xcoef = Tx(1 : dx, 1 : dx)\ u$ $ycoef = Ty(1 : dy, 1 : dy)\ v$
Datos de salida
λ valores propios a, b vectores propios

Cuadro C.3: Algoritmo usando descomposición QR[38]

Datos de entrada
K % Matriz kernel eta % Parámetro de precisión
% Método
$[m, n] = size(K);$ $j = 1;$ $norm2 = zeros(m, 1); sizeK = zeros(m, 1); index = zeros(m, 1); feat = zeros(m, m);$ $fori = 1 : m$ $norm2(i) = K(i, i);$ end $suma = 0;$ $fork = 1 : m$ $suma = suma + norm2(k);$ end while $suma > eta\&j \ = m + 1$ $[mas, ij] = max(norm2);$ $index(j) = ij;$ $sizeK(j) = sqrt(norm2(ij));$ $fori = 1 : m$ $suma2 = 0;$ $fort = 1 : j - 1$ $suma2 = suma2 + feat(i, t) * feat(ij, t);$ end $feat(i, j) = (K(i, ij) - suma2)/sizeK(j);$ $norm2(i) = norm2(i) - feat(i, j) * feat(i, j);$ end $j = j + 1;$ $suma = 0;$ $fork = 1 : m$ $suma = suma + norm2(k);$ end end $feat = feat(1 : m, 1 : j - 1);$
Datos de salida
$feat$ matriz ortogonal

Cuadro C.4: Algoritmo de ortogonalización parcial de Gram-Schmidt[9]

Datos de entrada
X % Matriz kernel para el primer grupo de variables Y % Matriz kernel para el segundo grupo de variables eta % Parámetro de precisión
% Método
$Kx = X * X';$ $Ky = Y * Y';$ $[Rx, Rxsize, Rxindex] = PGSO(Kx, eta)$ $[Ry, Rysize, Ryindex] = PGSO(Ky, eta);$ $Zxx = Rx' * Rx$ $Zyy = Ry' * Ry$ $Zxy = Rx' * Ry$ $Zyx = Ry' * Rx$ $S = chol(Zxx)'$ $invS = inv(S)$ $y = invS * Zxy$ $iZyy = inv(Zyy)$ $A = y * iZyy * y'$ $A = 0{,}5 * (A' + A) + eye(size(A, 1)) * 10e - 6$ $[alpha1, D] = eig(A)$ $lambda = diag(sqrt(D))$ $alphatesta = invS' * alpha1$ $invRx = Rx * inv(Rx' * Rx)$ $alpha = invRx * alphatesta$ $aux = inv(Zyy) * Zyx * alphatesta$ $Nlambda = size(lambda)$ $fork = 1 : Nlambda$ $beta1(:, k) = aux(:, k)/lambda(k)$ end $invRy = Ry * inv(Ry' * Ry)$ $beta = invRy * beta1$
Datos de salida
$lambda$ % correlaciones canónicas $alpha$ % vector canónico para el primer grupo $beta$ % vector canónico para el segundo grupo

Cuadro C.5: Algoritmo para el ACCK[9]

Bibliografía

[1] Berlinet, A. & Thomas-Agnan C. *Reproducing Kernel Hilbert Spaces In Probability and Statistics*. Springer Science+Business Media. Primera edición. New York. ISBN 978-1-4613-4792-7. 2004.

[2] Bickel, P. & Li, B. *Regularization in statistics*. Test, 15(2), 271 344. 2006.

[3] Castillo, M., Linares, G., Valera, M. A. y Garcia, N. E. *Modelación de la materia orgánica en suelos volcánicos de la región de Teziutlán, Puebla, México*. Revista Latinoamericana de Recursos Naturales, 5(2), 148-154. ISSN 1870-0667. 2009.

[4] Chan Keb, C. A., Linares, G., Agraz, C., Valera, M. A., Pérez, R. & Villegas, M. L. *Correlaciones Canónicas en los bosques de manglar del sistema lagunar Chacahua-Pastorías, Oaxaca*. Revista Ciencia en la Frontera 9(3), 27-43. ISSN 2007-042X. 2013.

[5] Chang, B., Kruger, U., Kustra, R. & Zhang, J. *Canonical Correlation Analysis based on Hilbert-Schmidt Independence Criterion and Centered Kernel Target Alignment*. 30th International Conference on Machine Learning. Atlanta, Georgia. 2013.

[6] Coello, C. A. *Introducción a los Algoritmos Genéticos*. Soluciones Avanzadas. Tecnologías de Información y Estrategias de Negocios, 3(17), 5-11. 1995

[7] Cuadras, C.M. *Nuevos métodos de análisis multivariante*. CMC Editions. Barcelona, Espa a. 2008.

[8] González, I., Déjean, S., Martin, P. & Baccini, A. *CCA: An R Package to Extend Canonical Correlation Analysis.* Journal of Statistical Software, 23(12). 2008.

[9] Hardoon, D., Szedmak, S. & Shawe-Taylor, J. *Canonical correlation analysis: an overview with application to learning methods*. Neural Computation, 16(12), 2639-2641. 2004.

[10] Horrillo, J. M., Pedrozo, A. & Onandia, B. *Pronóstico del perfil de playas de grava usando análisis de correlación canónica*. Tecnología y Ciencias del Agua, 1(2), pp. 5-19. 2010.

[11] Hotelling, H. *Relations between two sets of variates*. Biometrika 28, pp. 321-377. 1936.

[12] Ison, M., Sitt, J, & Trevisan, M. *Algoritmos genéticos: aplicación en MATLAB. Guía de la materia Sistemas Complejos*. Disponible en www.df.uba.ar/users/mison/genetico.tar.gz. 2005.

[13] Johnson, R. A. & Wichern, D. W. *Applied Multivariate Statistical Analysis*. Sexta edición. Editorial Prentice Hall. New Jersey. 2007.

[14] Karatzoglou, A., Smola, A. & Hornik, K. *kernlab An S4 Package for Kernel Methods in R*. Journal in Statistical Software, 11(9). 2004.

[15] Kettenring, J. R. *Canonical analysis of several sets of variables*. Biometrika, 58(3):433 451, 1971.

[16] Kuss, M. & Graepel. *The geometry of kernel canonical correlation analysis*. MPI-Technical Reports. http://www.kyb.mpg.de/publication.html?publ=2233. 2003.

[17] Kreyszig, E. *Introductory Functional Analysis with Applications*. Wiley Classics Lybrary. ISBN-10: 0471504599, ISBN-13: 978-0471504597. 1989.

[18] Linares Fleites, G., Ticante, J.A. & Almaray, R. *Estudio de interacciones medicamentosas en pacientes pediátricos con técnicas multivariadas*. Memorias en extenso del 2do. Congreso Nacional de Ciencias de la Computación. Facultad de Computación, BUAP. 2004.

[19] Linares, G. *Análisis de Datos Multivariados*. Editorial BUAP. Facultad de Ciencias de la Computación. ISBN: 968 9182 15 3. 2007.

[20] Linares, G., Valera, M. A. & Castillo, M. *Análisis de datos de suelos forestales en la caldera de Teziutlán, Puebla, por componentes principales y técnicas geoestadísticas*. Memorias del XXI Foro Nacional de Estadística. Instituto Nacional de Estadística, Geografía e Informática. ISBN 978-970-13-4930-4. 2007.

[21] Linares, G., Sandoval, M. de L., Matías, B. C., Reyes, H. J., Almaray, R. & Ticante, J. A. *Interacciones Medicamentosas: un estudio de correlaciones canónicas en pacientes pediátricos*. Modelos Matemáticos para el Estudio del Medio Ambiente, Salud y Desarrollo Humano. Tomo 3. Primera edición. RIDECA. Espa a. ISBN: 978-84-617-8721-0. 2017.

[22] Luenberger, D. *Programación lineal y no lineal*. Addison-Wesley Iberoamericana, S. A. E.U.A. 1989.

[23] Mardia, K., Kent, J. & Bibby, J. *Multivariate Analysis*. Academic Press. Great Britain. 1979.

[24] Matías, B. C., Linares, G., Reyes, H. J. & Sandoval, M. de L. *Comparación de Métodos para el Análisis de Correlación Canónica*. Memorias Simposio Internacional de Estadística XXIV. Bogotá, Colombia. ISSN: 2463-0861. 2014.

[25] Matías, B. C., Reyes, H. J. & Linares, G. *Análisis de Correlación Canónica Regularizada Generalizada: Una aplicación en bosques de mangle*. Aportaciones a la estadística de los XXVII y XXVIII Foros Nacionales de Estadística. Instituto Nacional de Estadística y Geografía. ISBN: 978-607-739-623-9. 2015.

[26] Matías, B. C., Sandoval, M. de L., Linares, G. & Reyes, H. J. *Uso de las funciones kernel en el Análisis de Correlación Canónica*. La Investigación y las Aplicaciones en Ciencias de la Computación 2014-2015. Benemérita Universidad Autónoma de Puebla. Primera Edición. ISBN: 978-607-487-968-1. 2015.

[27] Matías, B. C., Linares, G., Reyes, H. J., Almaray, R., & Ticante, J. A. *Correlaciones Canónicas en el estudio de interacciones medicamentosas en pacientes pediátricos: comparación entre el enfoque clásico y el uso de kernels*. Tópicos en Probabilidad y Estadística. Benemérita Universidad Autónoma de Puebla Primera Edición. ISBN: 978-607-525-083-0. 2016.

[28] Matías, B. C., Sandoval, M. de L., Linares, G. & Reyes, H. J. *Análisis de Correlación Canónica usando Algoritmos Genéticos*. Investigaci n Operacional, 38(1), 1-13. ISSN: 0257-4306. 2017

[29] Mika, S. *Kernel Fisher Discriminants*. Tesis de Doctorado. Universidad Tecnica de Berlín. 2002.

[30] Eaton, J. W., Bateman, D., Hauberg, S. & Wehbring. *GNU OCTAVE*. Disponible en www.gnu.org/software/octave/

[31] Otopal, N. *Restricted kernel canonical correlation analysis*. Linear Algebra and Its Applications, 437, pp. 1-13. 2012.

[32] Pe a, D. *Análisis de Datos Multivariantes*. McGraw-Hill Interamericana de Espa a, Espa a. 2002.

[33] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Disponible en: http://www.R-project.org/. 2013.

[34] Schölkopf, B.,Smola, A. & Muller, K. *Nonlinear component analysis as a kernel eigenvalue problem*. Neural Computation. 10, 1299-1319. 1998.

[35] Schölkopf, B., Mika, S., Burges, C., Knirsch, P., M ller, K.,R tsch G., & Smola, A. *Input space vs. feature space in kernel-based methods*. IEEE Transactions on Neural Networks, 10(5), 1000-1017. 1999.

[36] Schölkopf, B., Smola, A. y M ller, K. R. *Kernel Principal Component Analysis*. Advances in Kernel Methods-Support Vector Learning, 327-352. 1999.

[37] Schölkopf, B., & Smola, A. *Learning with Kernels*. Cambridge. The MIT Press Cambridge. 2002.

[38] Seber, G. A. F. *Multivariate Observations*. Wiley. Primera edición. New Jersey. 1984.

[39] Shawe Taylor, J. & Cristianini, N. *Kernel Methods for Pattern Analysis*. Cambridge University Press. Primera edición. 2004.

[40] Tahtali, R., ankaya, S. & Ulutas, Z. *Canonical correlation analysis for estimation of relationships between some traits measured at birth and weaning time in karayaka lambs*. Kafkas Univ Vet Fak Derg, 18(5), 839-844. 2012.

[41] Tenenhaus, Z. *Regularized generalized canonical correlation analysis*. Psychometrika, 76(2), 257-284.2011

[42] Tikhonov, A. N. *On the stability of inverse problems*. C. R. (Doklady) Acad. Sci. URSS (N.S.), 39, 176 179. 1943

[43] Wainwright, M. *Structured Regularizers for high-dimensional problems: statistical and computational issues*. Annu. Rev. Stat. Appl., 1, 233 53. 2014.

[44] Yang, X. *Nature-Inspired Metaheuristic Algorithms*. Luniver Press. 2008.