Benemérita Universidad Autónoma de Puebla.
Facultad de Ciencias Físico Matemáticas.

Modelos de regresión para datos funcionales por la metodología de kernel reproductor en espacios de Hilbert

Tesis presentada como requisito para obtener el título de:

Maestría en Ciencias Matemáticas

Presenta:

Gabriela López Pineda

Directoras de Tesis:

Dra. Hortensia J. Reyes Cervantes
Dra. Gladys Linares Fleites

Puebla, Pue. Mayo 2017.

Índice general

Introducción

La estadística aplicada es una herramienta imprescindible cuando se desea analizar un conjunto de datos, entendiendo como *"datos"* un conjunto de observaciones las cuales se obtienen a partir de muestreos realizados en algún experimento.

Actualmente se reconoce la importancia de la estadística aplicada en el desarrollo de investigaciones en diversos campos; cada vez son más los profesionales de diferentes disciplinas que requieren de métodos estadísticos como muestreo, simulación, diseño de experimentos, modelación estadística e inferencia, para llevar a cabo análisis de datos e interpretación.

En ciencias ambientales, los métodos estadísticos son de amplio uso y el Análisis de Regresión es la técnica estadística más frecuente para modelar la relación entre variables ecológicas y medioambientales las cuales se obtienen a partir de muestreos realizados en algún área de interés, en este caso el área de interés se encuentra localizada en la Región Terrestre Prioritaria (RTP-105) ubicada en Teziutlán correspondiente al estado de Puebla.

Para explicar el comportamiento del Carbono Orgánico en el Suelo (COS), se presentan tablas de datos donde abundan problemas de multicolinealidad así como relaciones no lineales entre las variables y, que por tanto, no pueden analizarse a través de las técnicas clásicas. Por lo tanto, si los datos en el espacio original no se pueden analizar de manera satisfactoria con técnicas de análisis multivariado, se tiene la necesidad de usar los denominados *métodos Kernel* con el fin de tomar en cuenta la estructura de dichos datos.

El tema central de este proyecto es el empleo de los *"métodos Kernel"* en el Análisis de Regresión con relaciones no lineales, ya que estos procedimientos se han vuelto una alternativa eficiente en modelación, y la idea fundamental es proyectar la base de datos original dada como un espacio vectorial de dimensión finita en un espacio de Hilbert de dimensión infinita. El llamado truco Kernel consiste en transformar los datos mediante una función Kernel y posteriormente aplicar las técnicas de análisis multivariado en los datos transformados esperando que se tengan mejores resultados.

Dado el avance y desarrollo que ha presentado la Estadística Computacional en las últimas décadas, actualmente se dispone de herramientas computacionales capaces de hacer cálculos que eran prácticamente imposibles de realizar debido a la gran complejidad de los procedimientos de cálculo. En este trabajo se usa el *software R* dado que, en la práctica su manejo es de gran utilidad para el

análisis e interpretación de datos. Como lenguaje de análisis estadístico, R tiene entre sus funciones básicas rutinas habitualmente sencillas y también permite utilizar funciones que implementan técnicas mas avanzadas.

Considerando lo previamente mencionado, este proyecto de tesis tiene los siguientes objetivos:

- Explicar el comportamiento del Carbono Orgánico en el Suelo (COS) en función de diferentes propiedades y tipos de suelo.
- Aplicar los métodos Kernel en el Análisis de Regresión donde existen problemas de multicolinealidad y no linealidad.
- Comparar diferentes tipos de Kernel utilizados para el análisis.
- Comparar los resultados de RCPK y RKPLS a través del uso de R.

En este trabajo de tesis se pretende trabajar de manera conjunta con diferentes temas, por lo cual se ha estructurado de la siguiente forma:

En el Capítulo 1, se revisan algunos de los aspectos básicos sobre el Análisis de Regresión, tales como el modelo de Regresión Lineal Múltiple, las pruebas de bondad del ajuste y comprobación de las suposiciones del modelo. También se brindan algunos de los criterios más comunes para la detección y tratamiento de *outliers* o datos atípicos ya que es uno de los problemas que aparece en el análisis que se realiza posteriormente. De igual forma se desarrolla de manera breve el problema de multicolinealidad, así como sus diagnósticos y, al final del capítulo se presentan distintos criterios para la selección de modelos los cuales nos permitirán elegir el mejor modelo.

En el Capítulo 2, se describe de forma general la teoría matemática de los espacios de Hilbert con Kernel Reproductor (RKHS, siglas en inglés). En el desarrollo de este capítulo se introduce el concepto de *función Kernel*, la cual es la base para la construcción del método Kernel. De igual forma se presentan resultados importantes dentro de esta teoría como son el *Teorema de Moore-Aronszajin* cuyo resultado establece la relación que existe entre una función Kernel y su respectivo espacio de Hilbert. Otro resultado importante en este capítulo es el *Teorema de Mercer* ya que permite calcular el producto interior en el conjunto de datos proyectados al espacio de Hilbert sin necesidad de conocer implícitamente la aplicación utilizada, esto es lo que se conoce como el *Truco Kernel*. Se dedica una sección para describir la técnica de *Regresión de Componentes Principales con Kernels* (RCPK) y se presenta una serie de pasos útiles para seguir esta técnica. También se presentan las ideas que toma la *Regresión Mínimos Cuadrados Parciales con Kernels* (RKPLS), así como el algoritmo general que utiliza esta técnica.

En el Capítulo 3, se presentan los resultados que se obtienen al hacer el análisis de los datos de los cuales se dispone, mediante el uso de diferentes librerías o paquetes de R. Durante la aplicación, se trabaja en diferentes etapas: la primera etapa consiste en ajustar un modelo de Regresión Lineal Múltiple, y revisar las suposiciones básicas del modelo, cabe señalar que en esta primera etapa se decide eliminar una observación, la cual fue considerada un outlier para posteriormente ajustar un nuevo modelo de Regresión sin dicha observación. En esta nueva etapa se realizan pruebas para diagnóstico de multicolinealidad y se decide utilizar como solución a este problema las técnicas de RCPK y RKPLS, ya que en los gráficos de diagnóstico se aprecian relaciones de tipo no lineal. En base a lo anterior se comparan ambas técnicas y se observa que para el estudio del COS, la técnica de RKPLS brinda mejores resultados en cuanto a la capacidad predictiva del modelo seleccionado.

Finalmente, dentro de las Conclusiones se presentan los resultados que se obtienen después de analizar los datos del COS mediante las técnicas señaladas, es importente señalar que el uso de los paquetes Kernlab y PLS incluidos en R resultan muy eficientes, ya que mediante ellos se pueden comparar las técnicas de RKPLS y RCPK. Cabe resaltar que para este problema, se concluye que la RKPLS es superior a la RCPK..

Capítulo 1

Regresión Lineal Múltiple

El Análisis de Regresión es la técnica estadística de uso más frecuente para investigar y **modelar la relación entre variables**. Su atractivo y utilidad generalmente son el resultado del proceso conceptualmente lógico de usar una ecuación para expresar la relación entre una variable de interés (la respuesta) y un conjunto de variables predictoras relacionadas.

1.1. Modelo de Regresión Lineal Múltiple

Sean $X_1, X_2, \ldots, X_k$, k variables predictoras que se cree están relacionadas con una variable de respuesta Y.

El modelo clásico de regresión lineal indica que la variable Y se compone de una media, que depende de manera continua de las $X_i's$, y un error aleatorio ε, que representa el error de medición y los efectos de otras variables que no fueron consideradas explícitamente en el modelo.

Los valores de las variables predictoras son tratados como *fijos*, mientras que el error y la respuesta son vistos como variables aleatorias cuyo comportamiento se caracteriza por un conjunto de hipótesis de distribución.

Así el modelo clásico de regresión lineal en forma matricial puede escribirse como:

$$Y = X\beta + \varepsilon, \tag{1.1}$$

en donde,

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

En general, $\mathbf{Y}$ es una matriz de $n \times 1$ de las observaciones, $\mathbf{X}$ es una matriz de $n \times (k+1)$ de variables regresoras, β es un vector de $(k+1) \times 1$ de los coeficientes de regresión desconocidos y ε es un vector de $n \times 1$ de errores aleatorios.

A fin de poder determinar las propiedades de los estimadores que se obtienen, debemos especificar un conjunto de hipótesis sobre el modelo (Montgomery et al., 2004).

Las principales suposiciones sobre el modelo de regresión son las siguientes:

- **Linealidad**: Y está relacionada con X mediante el modelo de regresión dado por la expresión (1.1).
- **Varianza constante**: $Var(\varepsilon_i) = \sigma^2$, para $i = 1, \dots, n$.
- **Independencia**: $Cov(\varepsilon_i, \varepsilon_j) = 0$, para $i \neq j$.
- **Normalidad**: $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ son variables aleatorias **no observables**, distribuidas normalmente con $\mu = 0$ y σ^2 constante. Esto puede escribirse como:

$$\varepsilon_i \sim N(0, \sigma^2), \qquad i = 1, \cdots, n. \tag{1.2}$$

- Las variables regresoras $X_1, X_2, \cdots, X_k$ son linealmente independientes en sentido algebraico y se consideran fijas, es decir, no son variables aleatorias.

Estimar el modelo de Regresión equivale a asignar valores numéricos a los parámetros desconocidos β a partir de la información muestral disponible de las variables observables del modelo.

Únicamente consideraremos dos métodos de estimación:

- El método de mínimos cuadrados ordinarios (OLS).

$$\hat{\beta} = (X'Y)(X'X)^{-1}, \qquad \hat{\sigma}^2 = \frac{(Y - X\hat{\beta})'(Y - X\hat{\beta})}{n - (k+1)}. \tag{1.3}$$

- El método de máxima verosimilitud (MV)

$$\hat{\beta} = (X'Y)(X'X)^{-1}, \qquad \hat{\sigma}^2 = \frac{(Y - X\hat{\beta})'(Y - X\hat{\beta})}{n}. \tag{1.4}$$

El teorema de Gauss-Markov en (Montgomery et al., 2004) establece que el estimador de mínimos cuadrados para β, es el mejor estimador lineal insesgado (BLUE, en inglés). En este caso, se puede decir que $\hat{\beta}$ tiene la varianza mínima, entre la clase de todos los estimadores insesgados que son combinaciones lineales de los datos. Si además se supone que los errores ε_i tienen distribución normal, entonces el estimador de mínimos cuadrados es el mismo que el estimador de MV y también es un estimador insesgado y de varianza mínima.

1.2. Pruebas de Bondad del Ajuste

Una vez estimados los parámetros del modelo, surgen de inmediato dos preguntas:

1. ¿Qué tan bien se ajusta el modelo a los datos?

2. ¿Qué regresores específicos son importantes?

Hay varios procedimientos de pruebas de hipótesis que demuestran su utilidad para contestar esas preguntas.

La prueba de **significancia de la regresión** nos sirve para determinar si existe una relación lineal entre la respuesta Y y cualquiera de las variables regresoras $X_1, X_2, \cdots, X_k$. Este procedimiento suele considerarse como una prueba general o global del ajuste del modelo. El contraste de hipótesis pertinentes es:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \quad Vs \quad H_1 : \beta_j \neq 0, \text{ para alguna } j = 1, \cdots, k.$$

Dado un nivel de significancia α, el rechazo de la hipótesis nula implica que al menos uno de los regresores $X_1, X_2, \cdots, X_k$ contribuye al modelo en forma significativa. Por consiguiente, para probar la hipótesis $H_0 : \beta_1 = \beta_2 = \beta_k = 0$, se calcula el estadístico de prueba F_0 y se rechaza H_0 si

$$F_0 > F_{\alpha,k,n-k-1}.$$

El procedimiento de prueba normalmente puede resumirse en una **tabla de ANOVA** como en la Tabla 1.1.

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Cuadrado medio**	$\mathbf{F}_0$
Regresión	SS_R	k	MS_R	MS_R/MS_{Res}
Residuales	SS_{Res}	$n-k-1$	MS_{Res}	
Total	SS_T	$n-1$		

Tabla 1.1: Análisis de varianza para determinar el ajuste de la regresión.

Una vez determinado que al menos una de las variables regresoras es importante, la siguiente pregunta es: ¿Cuál(es) variables son importantes?

El juego de hipótesis para probar la significancia de cualquier coeficiente β_j $j = 1, 2, \cdots, k$, está dado por:

$$H_0 : \beta_j = 0 \qquad Vs \qquad H_1 : \beta_j \neq 0.$$

Si no se rechaza $H_0 : \beta_j = 0$, quiere decir que se puede eliminar el regresor x_j del modelo. El **estadístico de prueba** para esta hipótesis es,

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}, \tag{1.5}$$

donde C_{jj} es el elemento diagonal de $(X'X)^{-1}$ que corresponde a $\hat{\beta}_j$. Se rechaza la hipótesis nula $H_0 : \beta_j = 0$ si

$$|t_0| > t_{\alpha/2, n-k-1}.$$

Por otro lado existen otras maneras de evaluar el ajuste general del modelo, a saber los estadísticos R^2 y R^2_{Adj}.

La cantidad

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}, \tag{1.6}$$

se llama **coeficiente de determinación**. Nótese que R^2 indica la proporción de variabilidad explicada por la regresión. Ya que $0 \leq SS_{Res} \leq SS_T$, entonces $0 \leq R^2 \leq 1$. Los valores de R^2 cercanos a 1 implican que la mayor parte de la variabilidad de Y esta explicada por el modelo de regresión.

En general, R^2 aumenta siempre y cuando se agrega un regresor al modelo, independientemente del valor de la contribución de esa variable. En consecuencia, es difícil juzgar si un aumento de R^2 dice en realidad algo importante.

Algunos investigadores que trabajan con modelos de regresión prefieren usar el estadístico $\boldsymbol{R^2_{Adj}}$, que se define como sigue:

$$R^2_{Adj} = 1 - \frac{SS_{Res}/(n - (k+1))}{SS_T/(n-1)}. \tag{1.7}$$

En vista de que $SS_{Res}/(n-k-1)$ es el cuadrado medio de residuales, $SS_T/(n-1)$ es constante e independiente de cuántas variables hay en el modelo, de modo que R^2_{Adj} sólo aumentará al agregar una variable al modelo si esa variable reduce el cuadrado medio residual.

La R^2 ajustada penaliza el aumento de términos que no son útiles, además es un procedimiento para evaluar y comparar los posibles modelos de regresión (Montgomery et al., 2004).

1.3. Comprobación de las Suposiciones del Modelo

Siempre se debe tener en cuenta que la validez de estas premisas es dudosa y se deben hacer análisis para examinar la adecuación del modelo que se ha desarrollado en forma tentativa. Las grandes violaciones a las suposiciones pueden dar como resultado un modelo inestable, en el sentido que una muestra distinta puede conducir a un modelo totalmente diferente y por tanto obtener conclusiones opuestas. En general, no se pueden detectar desviaciones respecto a las premisas básicas examinando los estadísticos estándar de resumen (t, F o R^2). Éstas propiedades son globales del modelo y como tal no aseguran la adecuación del mismo.

La herramienta básica para el diagnóstico del modelo es el análisis de los residuos, tanto a través de gráficos, como de test que verifican la validez de las hipótesis asumidas en el ajuste del modelo lineal.

A continuación se presentan algunos métodos para diagnosticar la violación a las suposiciones básicas de la regresión. Estos métodos se basan principalmente en el estudio de los **residuales** del modelo.

Como sabemos, los residuales se definen como:

$$\varepsilon = y_i - \hat{y}_i, \quad i = 1, \cdots, n \tag{1.8}$$

siendo y_i una observación y $\hat{y}_i$ su valor ajustado correspondiente. Podemos considerar que un residual es la **desviación** entre los **datos** y el **ajuste**, también es una medida de la variabilidad de la variable de respuesta que no explica el modelo de regresión.

También es conveniente imaginar que los residuales son los valores realizados (observados), de los errores del modelo, por lo que toda desviación de las suposiciones sobre los errores se debe reflejar en los residuales.

Los residuales tienen varias propiedades importantes. Tienen media cero y su varianza promedio se estima con:

$$\frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n-(k+1)} = \frac{SS_{Res}}{n-(k+1)} = MS_{Res}. \tag{1.9}$$

Sin embargo los residuales no son independientes, ya que los n residuales sólo tienen $n - (k + 1)$ grados de libertad asociados a ellos. Se ha observado que cuando no hay independencia en los residuales se tiene poco efecto para comprobar la adecuación del modelo, siempre y cuando, n no sea pequeña en relación con la cantidad de parámetros.

El análisis de residuales es una forma muy eficaz de descubrir diversos tipos de inadecuación del modelo. Como veremos, el análisis gráfico de los residuales es una forma muy efectiva de investigar la adecuación del ajuste de un modelo de regresión y para comprobar las suposiciones básicas.

1.3.1. Gráfica de Residuales en función de los $\hat{Y}$

La heterocedasticidad, que es como se denomina el problema de varianza no constante, aparece generalmente cuando el modelo está mal especificado, bien en la relación de la respuesta con los predictores, bien en la distribución de la respuesta, o bien en ambas cuestiones.

Para poder detectar algunas inadecuaciones en nuestro modelo de regresión, es útil tener una gráfica de los residuales en función de los valores ajustados correspondientes $\hat{y}_i$.

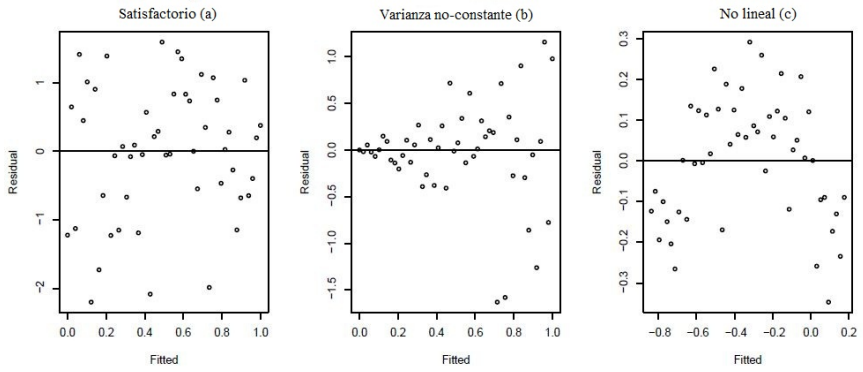

Figura 1.1: Patrones en las gráficas de residuales.

Si la gráfica que obtengamos es parecida a la Figura 1.1 del lado izquierdo, esta indica que los residuales se pueden encerrar en una banda horizontal y por tanto, no hay defectos graves en el modelo.

Las gráficas parecidas a la Figura 1.1 del centro, indican que la varianza de los errores no es constante.

El método más utilizado para manejar la varianza no común es aplicar una transformación adecuada ya sea a las variables regresoras o a la variable de respuesta.

Una gráfica en curva, como la de la Figura 1.1 del lado derecho, señala **no linealidad**. Esto podría indicar que se necesitan otras variables regresoras en el modelo (Faraway, 2014).

Dentro de (Aparicio et al., 2004) y (Faraway, 2014) se mencionan distintas pruebas o test para reconocer la heterocedasticidad (falta de homogeneidad de varianza en los residuos).

El test de Breusch-Pagan ajusta un modelo de Regresión Lineal a los residuos del modelo de Regresión ajustado, generalmente se toman las mismas variables explicativas que en el modelo de Regresión principal, y rechaza si una buena parte de la varianza es explicada por dichas variables. El estadístico de contraste de *Breusch-Pagan* (bajo homocedasticidad) sigue una distribución *Chi-cuadrado* con tantos grados de libertad como variables explicativas introducidas para justificar la falta de varianza constante (Aparicio et al., 2004).

1.3.2. Gráfica de Residuales en el tiempo

Para detectar la falta de independencia en los errores (autocorrelación) suelen ser útiles las gráficas de residuales. La presentación más adecuada es la de los residuales en función del tiempo. Si hay autocorrelación positiva, los residuales de igual signo se presentarán en grupos, indica que no hay los suficientes cambios de signo en la secuencia de los residuales. Por otra parte, si hay autocorrelación negativa, los residuales cambiarán de signo con demasiada rapidez.

Para detectar la presencia de la autocorrelación se pueden aplicar diversas pruebas estadísticas. La que desarrollaron Durbin y Watson se usa ampliamente, ya que se basa en la hipótesis de que los errores del modelo de regresión se generan en un proceso autoregresivo de primer orden, que se observa a intervalos de tiempo igualmente espaciados, esto es:

$$\varepsilon_i = \rho\varepsilon_{i-1} + \omega_i, \qquad\qquad i = 1, \cdots, n$$

donde ε_i es el término de error en el modelo, en el periodo t; $\omega_i \sim N(0, \sigma^2)$ es una variable aleatoria y ρ es el parámetro de autocorrelación.

Como la mayor parte de los problemas de regresión donde intervienen las series de tiempo tienen autocorrelación positiva, las hipótesis que se suelen considerar en la prueba de *Durbin-Watson* son:

$$H_0 : \rho_s = 0 \qquad Vs \qquad H_1 : \rho_s > 0. \tag{1.10}$$

Para resolver el test (1.10) se utiliza el estadístico de Durbin-Watson, definido por:

$$d = \frac{\sum_{i=2}^{n}(\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^{n}\varepsilon_i^2}$$

en donde las ε_i, $t = 1, 2, \ldots, n$ son los residuales de un análisis de mínimos cuadrados ordinarios aplicado a los datos (y_i, x_i). Desafortunadamente, la distribución de d depende de la matriz X, sin embargo, Durbin y Watson demostraron que d está entre dos cotas, digamos que d_L y d_U, tales que si d sale de esos límites, se puede llegar a una conclusión acerca de las hipótesis de las ecuaciones (1.10). El procedimiento de decisión es el siguiente:

Si $d < d_L$	Rechazar $H_0 : \rho = 0$.
Si $d > d_U$	No rechazar $H_0 : \rho = 0$.
Si $d_L \leq d \leq d_U$	La prueba no es concluyente.

Es claro que los valores pequeños de d implican que $H_0 : \rho = 0$ se deben rechazar, porque la autocorrelación positiva indica que los términos consecutivos de error son de magnitud parecida, y las diferencias en los residuales, $\varepsilon_i - \varepsilon_{i-1}$ serán pequeñas (Montgomery et al., 2004).

1.3.3. Gráfica de Probabilidad Normal

La hipótesis de normalidad de los errores ε_i en el modelo lineal justifica la utilización de los *tests F y t* para realizar los contrastes de hipótesis habituales y obtener conclusiones confiables a cierto nivel de confianza $1 - \alpha$ dado. Además, si los errores provienen de una distribución con colas más gruesas que la normal, el ajuste por mínimos cuadrados será sensible a un subconjunto menor de datos.

Las distribuciones de los errores con *colas* gruesas generan con frecuencia valores atípicos que *jalan* demasiado en su dirección el ajuste por mínimos cuadrados. En esos casos se deben considerar otras técnicas de estimación.

La forma habitual de diagnosticar no normalidad es a través de los gráficos de normalidad (*qq-plot*) y de tests como el de *Shapiro-Wilks*, específico para normalidad.

- *Gráficos de normalidad*: Un método muy sencillo de comprobar la suposición de normalidad es trazar una gráfica de **probabilidad normal** de los residuales. Esta es una gráfica diseñada para que se dibuje una línea recta, que representa a una normal acumulada.

  La recta se suele determinar en forma visual, con énfasis en los valores centrales (*por ejemplo, los puntos de probabilidad acumulada* 0.33 *y* 0.67)

y no en los extremos. Las diferencias apreciables en distancia respecto a la recta indican que la distribución no es normal.

A veces, las gráficas de probabilidad normal se trazan graficando el residual clasificado $\varepsilon_{[i]}$ con $i = 1, \ldots, n$, en función del *valor normal esperado*, $\phi^{-1}[(i - \frac{1}{2})/n]$, donde ϕ representa la distribución normal estándar acumulada. Esto es consecuencia de que $E(\varepsilon_{[i]}) \simeq \phi^{-1}[(i - \frac{1}{2})/n]$ (Faraway, 2014) (Montgomery et al., 2004).

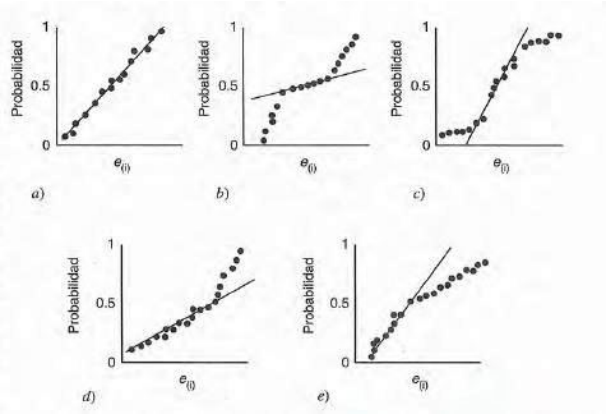

Figura 1.2: Gráficas de probabilidad normal: a) ideal; b) distribución con colas gruesas; c) distribución con colas delgadas; d) asimetría positiva; e) asimetría negativa.

- *Test de Shapiro-Wilk*: Esta prueba para normalidad se basa, más o menos, en una correlación entre los cuantiles empíricos de los residuos y los teóricos según una distribución normal. Cuanta mayor correlación, más indicios de normalidad para los residuos (Aparicio et al., 2004).

1.4. Criterios de detección de outliers.

Un outlier (valor atípico) es una observación o un conjunto de observaciones que *parecen* ser inconsistentes con el resto del conjunto de datos. Aunque son muchas las definiciones sobre el concepto, lo que caracteriza a una observación outlier es el *impacto* que produce en el ajuste de los datos. Parece evidente que la presencia de outliers en un conjunto de datos puede conducir a errores en el intento de hacer inferencias acerca de la población de la que proceden, de ahí que la presencia de outliers plantee un problema fundamental en el análisis de datos.

Los outliers se deben investigar con cuidado, para ver si se puede encontrar una razón de su comportamiento extraordinario. A veces, los valores atípicos son *"malos"* y se deben a eventos desacostumbrados, pero explicables. Entre los ejemplos están la medición o el análisis incorrecto, el registro incorrecto de los datos o bien la falla de un instrumento de medición (Montgomery et al., 2004).

Un procedimiento para la resolución de este problema consiste en encontrar reglas de decisión para detectar dichas observaciones, podemos basarnos, por ejemplo en *métodos analíticos* o bien en *métodos gráficos* (Ampanthong and Suwattee, 2009).

1.4.1. Métodos Analíticos

Hay muchos valores estadísticos calculados a partir de los datos de la muestra que se pueden utilizar para identificar la existencia de los valores atípicos. Para identificar la existencia de uno o más valores extremos se pueden utilizar las siguientes estadísticas:

- Distancia de Mahalanobis: La distancia de Mahalanobis se define de la siguiente manera:

  $$DM_i = \sqrt{(X_i - \bar{X})'\Sigma^{-1}(X_i - \bar{X})}.$$

  Para datos distribuidos normalmente, éstos valores de tienen aproximadamente una distribución chi-cuadrado con p grados de libertad. En consecuencia, aquellas observaciones con una distancia de Mahalanobis grande se indican como valores atípicos (Montgomery et al., 2004).

- Distancia de Cook's: La distancia de Cook's es una medida de cómo cambia la longitud entre los valores ajustados calculados con y sin la i-ésima observación,

  $$DC_i^2 = \frac{(\hat{\beta}_i - \beta)'(X'X)(\hat{\beta}_i - \beta)}{k\hat{\sigma}^2}.$$

  Se puede utilizar DC para identificar observaciones que tengan valores predictores poco comunes en comparación con los datos restantes y las observaciones que el modelo no ajusta adecuadamente, usualmente se examinan casos con $DC_i^2 > 1.0$.

- Potencial de un punto: El potencial (leverage) de un punto, h_{ii}, es el correspondiente elemento de la diagonal *matriz sombrero* $H = X(X'X)^{-1}X'$. Una interpretación del leverage indica que por debajo de 0.2, el leverage es bajo; valores entre 0.2 y 0.5 pueden ser peligrosos; los puntos cuyo leverage es superior a 0.5 deben ser analizados (Ampanthong and Suwattee, 2009).

- Residuos Estandarizados: Para cada residual $\varepsilon_i = y_i - \hat{y}_i$, se calcula el residual estandarizado

$$r_i = \frac{\varepsilon_i}{\sqrt{\hat{\sigma}^2_{(}1 - h_{ii})}}$$

con $i = 1, \ldots, n$. Los residuales estandarizados tienen media cero y varianza aproximadamente unitaria, en consecuencia, un residual estandarizado grande (por ejemplo $r_i > 3$) indica que se trata de un valor atípico potencial.

- Residuos Estudentizados: Los residuos estudentizados están dados por:

$$t_i = \frac{\varepsilon_i}{\sqrt{\hat{\sigma}^2_{(i)}(1 - h_{ii})}}$$

con $i = 1, \ldots, n$ donde $|t_i| > t_{(\alpha/2n),n-(k-1)}$ indica la existencia de valores atípicos. Los residuos estudentizados son usados para detectar outliers de la variable respuesta independientemente de la escala de su medición. Podemos decir que para valores absolutos de residuos estudentizados entre 2 y 3 son usualmente considerados como posibles outliers, mientras que valores mayores a 3 ya son considerados outliers o valores atípicos (Ampanthong and Suwattee, 2009).

1.4.2. Métodos Gráficos

Para los métodos gráficos, identificamos la presencia de valores atípicos por la forma del diagrama o el gráfico de los datos observados o residuos, dentro de éstos se encuentran los siguientes:

- Gráficos de dispersión (Scatter Plot): En este tipo de gráfico usualmente se observa la dispersión de los puntos de datos observados, uno o más puntos que se muestran apartados de la mayoría, indican la presencia de valores atípicos (Aparicio et al., 2004).

- Gráficos de probabilidad normal (qq-plot): Un método muy sencillo es trazar una gráfica de probabilidad normal de los residuales. Esta es una gráfica diseñada para que se dibuje una línea recta, que representa a una normal acumulada. Sean $\varepsilon_{[1]} < \varepsilon_{[2]} < \cdots < \varepsilon_{[n]}$ los residuales ordenados en orden creciente. Si se grafican $\varepsilon_{[i]}$ en función de la probabilidad acumulada $P_i = (i - \frac{1}{2})/n$, $i = 1, 2, \cdots, n$, los puntos que resulten deberían estar aproximadamente sobre una línea recta, de modo que si algunos puntos se encuentran alejados de esa recta indican que se tienen valores atípicos (Montgomery et al., 2004).

- Gráfico de cajas y bigotes (boxplot): Es un gráfico que muestra información sobre los valores mínimo y máximo, los cuartiles $Q1$, $Q2$ o *mediana* y $Q3$, la simetría de la distribución y sobre la existencia de valores atípicos (Aparicio et al., 2004).

- Gráfico de *Residuals vs Leverage*: Generalmente se utiliza este gráfico, como indicador de observaciones alejadas o influyentes en el ajuste (posibles outliers u observaciones deficientes) (Aparicio et al., 2004).

Los análisis de identificación y de seguimiento de los outliers con frecuencia dan como resultado mejoras en el proceso, o nuevos conocimientos acerca de factores cuyo efecto sobre la respuesta se desconocía antes. Si éste es el caso, el outlier se debería corregir (si es posible) o eliminar del conjunto de datos. Es claro que el eliminar valores malos es conveniente, porque los mínimos cuadrados jalan la ecuación ajustada hacia el valor atípico, ya que eso minimiza la suma de cuadrados de residuales (Ampanthong and Suwattee, 2009).

1.5. Multicolinealidad

Un problema relativamente frecuente en regresión lineal múltiple es la presencia de multicolinealidad, esto es: la violación al supuesto de regresión referido a la independencia lineal entre las variables predictoras.

Cuando los regresores no están relacionados linealmente entre sí, se dice que son *ortogonales*. Que exista multicolinealidad significa que las columnas de X son linealmente dependientes, de modo que si existiera una dependencia lineal total entre algunas de las columnas, se tendría que el rango de la matriz $X'X$ es menor a k y la matriz inversa $(X'X)^{-1}$ no existe (Vega-Vilca and Guzmán, 2011).

Este problema hace difícil cuantificar con precisión el efecto que cada variable predictora ejerce sobre la variable dependiente Y, lo cual hace que el análisis del modelo de regresión por OLS no sea adecuado. Las principales fuentes de multicolinealidad son las siguientes:

- El método de recolección de los datos que se empleó.
- Restricciones en el modelo o en la población.
- Especificación del modelo.
- Un modelo sobredefinido.

Es importante comprender las diferencias entre las fuentes de multicolinealidad, ya que los datos y la interpretación del modelo resultante dependen, en cierto grado de la causa del problema. Así mismo, la multicolinealidad tiene una gran cantidad de efectos graves sobre los estimadores de los coeficientes de regresión por OLS.

La fuerte multicolinealidad entre variables da como resultado **grandes varianzas y covarianzas** de los estimadores de coeficientes de regresión,

por otra parte también puede producir estimadores $\hat{\beta}_j$, $j = 1, \cdots, n$ que son **demasiado grandes** en valor absoluto, es decir, que la distancia de $\hat{\beta}_j$ al vector β_j, $j = 1, \cdots, n$ del parámetro real es muy grande (Montgomery et al., 2004).

1.5.1. Diagnósticos de Multicolinealidad

Existen diversos diagnósticos propuestos para detectar problemas de multicolinealidad. Las características deseables en un método de diagnóstico son que refleje el grado del problema de multicolinealidad y, por tanto que proporcione información de utilidad para determinar qué regresores están implicados.

A continuación se describirán las medidas más relevantes para el diagnóstico.

La multicolinealidad puede ser determinada mediante el cálculo del Factor de Inflación de la Varianza (*VIF*, por sus siglas en inglés) y por el número condición (η).

El *VIF* es un indicador de multicolinealidad específica de cada variable predictora del modelo, y se define como sigue:

$$VIF_j = \frac{1}{1 - R_j^2}, \quad i = 1, \cdots, n, \tag{1.11}$$

donde R_j^2 es el coeficiente de determinación obtenido cuando se hace la regresión de x_j respecto a las demás variables predictoras. Es claro que si x_j depende casi linealmente de alguno de los demás regresores, entonces R_j^2 será cercano a uno y *VIF* será grande.

Generalmente, valores *VIF* mayores a 5 o 10, es indicio de que los coeficientes asociados de regresión están mal estimados debido a la multicolinealidad.

Así mismo el número de condición también es un indicador de multicolinealidad global de las variables predictoras del modelo, tal número se define como:

$$\eta = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}, \tag{1.12}$$

donde λ_{max} y λ_{min} son los eigenvalores máximo y mínimo de la matriz de correlaciones R.

En general, si el número de condición es menor que 100, no hay problema grave de multicolinealidad. Los números de condición de 100 a 1000 implican multicolinealidad de moderada a fuerte, y si η es mayor que 1000, es indicio de una fuerte multicolinealidad (Montgomery et al., 2004), (Vega-Vilca and Guzmán, 2011).

1.6. Predicción de nuevas observaciones

Una aplicación importante del modelo de regresión es *predecir observaciones futuras* de Y que correspondan a determinados valores para las variables regresoras (Montgomery et al., 2004). Si $X_0' = [1, x_{01}, x_{02}, \cdots, x_{0n}]$, entonces un estimado puntual de la observación futura Y_0 en el punto X_0 estará dado por:

$$\widehat{Y}_0 = X_0'\widehat{\beta}.$$

Para desarrollar un *intervalo de predicción* para la observación futura Y_0, notemos que la varianza asociada a la predicción es:

$$Var(\widehat{Y}_0) = \sigma^2(1 + X_0(X'X)^{-1}X_0').$$

Así, el intervalo de confianza a nivel $100(1-\alpha)\,\%$ para esta futura observación viene dado por:

$$\widehat{Y}_0 \pm t_{n-p}^{\alpha/2}\sqrt{\sigma^2(1 + X_0(X'X)^{-1}X_0')}.$$

1.7. Comparación y selección de modelos.

En el trabajo de modelación estadística, es de primordial importancia la selección del modelo, es decir, elegir dentro de un conjunto de modelos alternativos el modelo más apropiado para el conjunto de datos que se considera.

Se dispone de diversos criterios para comparar y seleccionar el mejor modelo de entre diferentes alternativas. En ocasiones diferentes modelos darán los mismos resultados, pero esto no sucede en general, por lo que habrá de decidir qué criterio utilizar en función de sus intereses prioritarios. La selección del modelo la podemos basar entre otros en:

- El coeficiente de determinación ajustado R^2_{Adj}.
- El estadístico AIC (Akaike Information Criteria).
- El error de predicción o estadística de PRESS.

Una vez seleccionado el *mejor* modelo según el criterio elegido, habremos de continuar con la confirmación del mismo realizando la diagnosis y la validación del modelo, que puede fallar en algún paso, lo que nos llevaría nuevamente a la reformulación del modelo (y todos los pasos que le siguen), optando por aquellas correcciones y/o transformaciones de variables sugeridas en el diagnóstico. La obtención del mejor modelo será pues, un procedimiento iterativo, basado en selección y valoración de la calidad del ajuste, diagnóstico y validación.

1.7.1. El estadístico AIC

El criterio de información de Akaike está basado en la función de verosimilitud e incluye una penalización que aumenta con el número de parámetros estimados en el modelo. Premia pues, los modelos que dan un buen ajuste en términos de verosimilitud y a la vez son parsimoniosos (tienen pocos parámetros).

Este criterio *(AIC)* es una medida de la calidad relativa de un modelo estadístico, para un conjunto dado de datos. Como tal, el AIC proporciona un medio para la selección del modelo.

Si $\hat{\beta}$ es el estimador de máxima verosimilitud del modelo (1.1), de dimensión k, y $\log(n)$ denota el logaritmo (neperiano) de la verosimilitud dada por (1.4), el estadístico AIC se define por:

$$AIC = -2l(\hat{\beta}) + 2k.$$

Como se señala en (Aparicio et al., 2004), valores pequeños para este criterio identifican mejores modelos.

1.7.2. La estadística $PRESS$

Para comparar la capacidad predictiva de los modelos es útil introducir varias medidas del ajuste del modelo a los datos y de la fuerza de predicción de esos modelos.

La Suma de Cuadrados de Error de Predicción, conocida como la estadística *PRESS*, se considera una medida de lo bien que funciona un modelo de regresión para predecir nuevos datos y se define como la suma de los residuales PRESS al cuadrado que son los residuos que se obtiene entre el valor observado y el valor predicho de la i-ésima respuesta observada, basado en un ajuste de modelo con los puntos restantes de la muestra.

Definimos los residuales *PRESS* como $\varepsilon_{(i)} = y_{(i)} - \hat{y}_{(i)}$ para toda $i = 1, \cdots, n$; siendo $\hat{y}_{(i)}$ el valor predicho de la i-ésima respuesta observada, basado en un ajuste del modelo con los $n-1$ puntos de muestra.

$$\boldsymbol{PRESS} = \sum_{i=1}^{n} e_{(i)}^2 = \sum_{i=1}^{n} [y_i - \hat{y}_{(i)}]^2. \tag{1.13}$$

Se considera que la estadística de **PRESS** dada por (1.13), es una medida de lo bien que funciona un modelo de regresión para *predecir nuevos datos*. Lo deseable es tener un modelo con valor pequeño de PRESS (Montgomery et al., 2004).

Capítulo 2

Modelos de Regresión y el Método Kernel

Los modelos de regresión no lineal tienen por meta construir modelos exactos, mediante ecuaciones funcionales que permitan predecir, controlar u optimizar problemas no lineales a lo cual se conoce como Análisis de Datos Funcionales (FDA, por sus siglas en inglés, Functional Data Analisys). En (Ferraty and Vieu, 2006) y (Ramsay, 2006) se desarrolla de manera más general la teoría del FDA en espacios de Hilbert de dimensión infinita.

En este Capítulo consideraremos los Espacios de Hilbert con Kernel Reproductor (RKHS, siglas en inglés) con el propósito de recoger la naturaleza funcional de los propios datos y obtener así una representación de los mismos. Muchos de los métodos desarrollados en el FDA utilizan diferentes bases ortogonales en espacios de funciones para poder representar cada dato funcional como combinación lineal de los elementos de cierta base. (Ferraty and Vieu, 2006).

El *método de kernel* es una herramienta de modelación muy poderosa, y la idea fundamental es que los datos de entrada son transformados en un espacio de dimensiones altas que pueden ser eficientemente computadas mediante la utilización de kernels. De esta forma se permite generalizar las técnicas de análisis multivariado lineal en un RKHS.

2.1. Introducción a los Espacios de Hilbert con Kernel Reproductor

A continuación se pretende revisar de forma general algunos conceptos básicos referentes a la teoría de los Espacios de Hilbert con Kernel Reproductor.

Comenzaremos brindando algunas definiciones como la de espacio vectorial, producto interior, norma, entre otras; con el fin de establecer la condición de completez para construir un espacio de Hilbert, también se presentan algunas de las propiedades básicas del Kernel reproductor y algunos de los ejemplos más ilustrativos (Friedberg et al., 1997), (Kreyszig, 1989).

2.1.1. Espacio vectorial y Espacio con producto interior

Definición 2.1 (Espacio Vectorial) *Un espacio vectorial sobre un campo* $\mathbb{F}$ *(*$\mathbb{C}$ *o* $\mathbb{R}$*), es un conjunto de objetos (comúnmente llamados vectores), dotado de dos operaciones binarias llamadas suma y producto por un escalar, lo denotamos como la terna* $(X, +, \cdot)$*, donde las operaciones* $+ : X \times X \to \mathbb{F}$ *y* $\cdot : \mathbb{F} \times X \to X$*, satisfacen las siguientes propiedades:*

1. $\forall x, y \in X : x + y = y + x \in X$ *(conmutatividad).*
2. $\forall x, y, z \in X : (x + y) + z = x + (y + z) \in X$ *(asociatividad respecto a los vectores).*
3. $\forall x \in X : \exists e \in X$*:* $e + x = x + e = x \in X$ *(neutro aditivo).*
4. $\forall x \in X : \exists - x \in X : -x + x = x + (-x) = e \in X$ *(inverso aditivo).*
5. $\forall x \in X$ *y* $\lambda, \mu \in \mathbb{F} : (\lambda\mu)x = \lambda(\mu x) = \mu(\lambda x) \in X$ *(asociatividad respecto a los escalares).*
6. $\forall x \in X : \exists 1 \in \mathbb{F} : 1x = x \in X$ *(neutro multiplicativo).*
7. $\forall x, y \in X$ *y* $\lambda \in \mathbb{F} : \lambda(x + y) = \lambda x + \lambda y \in X$ *(distributividad respecto a los vectores).*
8. $\forall x \in X$ *y* $\lambda, \mu \in \mathbb{F} : (\lambda + \mu)x = \lambda x + \mu x \in X$ *(distributividad respecto a los escalares).*

Definición 2.2 (Subespacio Vectorial) *Sea* X *un espacio vectorial. Si* $Y \subseteq X$ *y* Y *es un espacio vectorial sobre el mismo campo escalar y con las mismas operaciones (de suma y producto por escalar) definidas para* X*, entonces decimos que* Y *es un* ***subespacio vectorial*** *de* X*.*

Otros conceptos importantes, que serán de mucha utilidad son el concepto de independencia lineal, bases y dimensión de un espacio vectorial, las cuales se enuncian enseguida.

Definición 2.3 *Decimos que un conjunto de vectores* $x_1, x_2, \cdots, x_n \in X$ *es linealmente independiente si, existen escalares* $\alpha_1, \alpha_2, \cdots, \alpha_n \in \mathbb{F}$*, tales que:*

$$\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n = 0,$$

implica que $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$*.*

Definición 2.4 (Base y dimensión) *Una base β para un espacio vectorial X es un subconjunto linealmente independiente de X que genera a X. El número de elementos de una base de X se llama dimensión de X y se denota por $dim(X)$, el cual es único. Un espacio vectorial se llama dimensionalmente finito si tiene una base que consta de un número finito de elementos, análogamente, si un espacio vectorial es de dimensión infinita se llama dimensionalmente infinito (Friedberg et al., 1997).*

Definición 2.5 (Producto interior) *Sea X un espacio vectorial sobre $\mathbb{F}$. Un* ***producto interior*** *en X es una función $\langle \cdot, \cdot \rangle : X \times X \to \mathbb{F}$, que asigna a cada par ordenado de vectores $x, y \in X$ un escalar en $\mathbb{F}$, tal que para toda x, y y $z \in X$ y toda $\alpha \in \mathbb{F}$ se cumplen las siguientes propiedades:*

- $\langle x + y, z \rangle = \langle x, y \rangle + \langle x, y \rangle$.
- $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$.
- $\langle x, y \rangle = \overline{\langle y, x \rangle}$, *donde la barra indica conjugación compleja.*
- $\langle x, x \rangle \geq 0$ *para todo* $x \in X$ *y* $\langle x, y \rangle = 0$, *si y sólo si* $y = 0$.

Cuando $\mathbb{F} = \mathbb{C}$ la barra de la tercera propiedad denota la conjugación compleja, y en el caso en que $\mathbb{F} = \mathbb{R}$ esta barra no afecta el producto (Friedberg et al., 1997).

Proposición 2.1 *Consideremos un espacio vectorial X con producto interior definido. Entonces,*

- $\langle x, \alpha y + \beta z \rangle = \bar{\alpha} \langle x, y \rangle + \bar{\beta} \langle x, z \rangle$ *para todo par* $\alpha, \beta \in \mathbb{F}(\mathbb{C}$ *o* $\mathbb{R})$ *y para todo* $x, y, z \in X$.
- $\langle x, y \rangle = 0$ *para toda* $x \in X$ *si y sólo si* $y = 0$.

En algunos casos, dependiendo de las características que se tienen en ciertos espacios, es fácil definir un producto interior, como el siguiente ejemplo.

Ejemplo 2.1 *Consideremos $X = \mathbb{C}^n$ (el n-espacio euclidiano complejo) y $\mathbb{F} = \mathbb{C}$. Para un par de vectores $x = (\alpha_1, \cdots \alpha_n)$ y $y = (\beta_1, \cdots \beta_n)$, definimos el siguiente producto*

$$\langle x, y \rangle = \sum_{i=1}^{n} \alpha_i \bar{\beta}_i. \tag{2.1}$$

Sean x, y y z elementos de $\mathbb{C}^n$, y $a, b \in \mathbb{C}$. Es claro que $ax + by = (a\alpha_1 + b\beta_1, \cdots, a\alpha_n + b\beta_n)$ y si $z = (\xi_1, \cdots \xi_n)$, entonces,

$$\begin{aligned}\langle ax+by,z\rangle &= \sum_{i=1}^{n}(a\alpha_i+b\beta_i)\bar{\xi}_i \\ &= a\sum_{i=1}^{n}\alpha_i\bar{\xi}_i + b\sum_{i=1}^{n}\beta_i\bar{\xi}_i \\ &= a\langle x,z\rangle + b\langle y,z\rangle.\end{aligned}$$

Ahora bien, $\langle x,y\rangle = \sum_{i=1}^{n}\alpha_i\bar{\beta}_i = \overline{\sum_{i=1}^{n}\bar{\alpha}_i\beta_i} = \overline{\langle y,x\rangle}$. *Por último* $\langle x,x\rangle = \sum_{i=1}^{n}\alpha_i\bar{\alpha}_i = \sum_{i=1}^{n}|\alpha_i|^2 \geq 0$. *De este modo, el producto 2.1 es un producto interior sobre* $\boldsymbol{X} = \mathbb{C}^n$.

Ejemplo 2.2 *Sea* $X = \mathcal{C}([a,b],\mathbb{C})$ *el conjunto de las funciones complejas continuas sobre el intervalo cerrado* $[a,b]$. *Definimos la suma puntual de funciones y el producto puntual por escalares como sigue* $(f+g)(x) := f(x)+g(x)$ *y* $(\alpha f)(x) := \alpha f(x)$, *entonces el producto*

$$\langle f,g\rangle = \int_a^b f(x)\overline{g(x)}dx, \tag{2.2}$$

define un producto interior sobre el espacio X. *La linealidad y la antisimétria, se siguen de forma inmediata por cómo está definido el producto. Ahora, notemos que si* $\langle f,f\rangle = 0$, *entonces* $f = 0$, *ya que* f *es continua; el reciproco es inmediato. Así la expresión (2.2) define un producto interior.*

Ejemplo 2.3 *Sea* $L^2(\mathbb{R})$ *el conjunto de todas las funciones medibles de* $\mathbb{R}$ *en* $\mathbb{C}$ *tales que se cumplen la siguiente propiedad: Si* $f(x) \in L^2(\mathbb{R})$, *entonces*

$$\int_{-\infty}^{\infty}|f(x)|^2 d\mu$$

$L^2(\mathbb{R})$ *es un espacio vectorial con el producto interior*

$$\langle f,g\rangle = \int_a^b f(x)\overline{g(x)}d\mu,$$

Observación 2.1 *En particular si* $a,b \in \mathbb{R}$ *y* $a < b$, *consideremos el espacio* $L^2([a,b],\mathfrak{B}([a,b]),\nu)$ *donde* $\mathfrak{B}([a,b])$ *son los Borelianos de* $[a,b]$ *y* ν *es la medida de Lebesgue en* $[a,b]$. *Estos espacios los denotaremos por* $L^2[a,b]$ *y los dotaremos del producto interior que esta dado por*

$$\langle f,g\rangle = \int_a^b |f(x)|^2 d\nu.$$

2.1.2. Espacios normados

Como sabemos el producto interior induce un tipo de función real llamada *norma* en X, la cual está dada por (2.3) y tiene diferentes propiedades en dichos espacios.

$$\|x\| = \langle x, x\rangle^{1/2}, x \in X. \tag{2.3}$$

Definición 2.6 (Norma) *Sea X un espacio vectorial sobre $\mathbb{F}$. Una norma sobre X es una función $\|\cdot\| : X \to \mathbb{R}^+$ tal que para todo $x, y \in X$ y $\alpha \in \mathbb{F}$ se cumplen las siguientes propiedades:*

- $\|x\| \geq 0$.
- $\|x\| = 0$ *si y sólo si* $x = 0$.
- $\|\alpha x\| = |\alpha|\|x\|$.
- *Desigualdad triangular:* $\|x + y\| \leq \|x\| + \|y\|$.

Este tipo de espacios se denominan ***espacios normados*** *(Kreyszig, 1989).*

Algunas propiedades importantes dentro de los espacios con producto interior y en particular dentro de los espacios normados son las siguientes.

Proposición 2.2 (Identidad del paralelogramo) *Sea X un espacio vectorial con producto interior, entonces para toda $x, y \in X$ se cumple lo siguiente:*

$$\|x + y\|^2 + \|x - y\|^2 = 2\left(\|x\|^2 + \|y\|^2\right),$$

donde $\|x\|^2 = \langle x, x\rangle$.
***Demostración**: Sean $x, y \in X$*

$$\begin{aligned}
\|x + y\|^2 + \|x - y\|^2 &= \langle x + y, x + y\rangle + \langle x - y, x - y\rangle \\
&= \langle x, x + y\rangle + \langle y, x + y\rangle + \langle x, x - y\rangle + \langle y, x - y\rangle \\
&= \langle x, x\rangle + \langle x, y\rangle + \langle y, x\rangle + \langle y, y\rangle + \langle x, x\rangle \\
&\quad - \langle x, y\rangle - \langle y, x\rangle + \langle y, x\rangle \\
&= 2\langle x, x\rangle + 2\langle y, y\rangle \\
&= 2\left(\|x\|^2 + \|y\|^2\right).
\end{aligned}$$

Proposición 2.3 (Desigualdad de Cauchy-Schwartz) *Sea X un espacio vectorial con producto interior. Entonces para todo $x, y \in X$ se cumple:*

$$|\langle x, y\rangle| \leq \|x\|\|y\|.$$

***Demostración**: Para toda $a \in \mathbb{F}$, y para todo $x, y \in X$, se tiene*

$$0 \leq \langle x + ay, x + ay\rangle = \|x\|^2 + \bar{a}\langle x, y\rangle + a\langle x, y\rangle + |a|^2\|y\|^2. \tag{2.4}$$

Supongamos que $\langle y, y\rangle = 0$*, entonces* $\|y\| = 0$ *y* $\langle x, y\rangle = 0$*, lo cual implicaría que* $|\langle x, y\rangle| = 0$*, de modo que*

$$0 = |\langle x, y\rangle| \leq \|x\|\|y\| = 0.$$

Ahora supongamos que $\langle y, y\rangle > 0$*, entonces* $\|y\| > 0$*, sustituyendo* $a = -\frac{\langle x,y\rangle}{\langle y,y\rangle}$ *en (2.4), se tiene*

$$\begin{aligned} 0 &\leq \|x\|^2 - 2\frac{|\langle x, y\rangle|^2}{\|y\|^2} + \frac{|\langle x, y\rangle|^2}{\|y\|^2} \\ &= \|x\|^2 - \frac{|\langle x, y\rangle|^2}{\|y\|^2}, \end{aligned}$$

de donde se obtiene la desigualdad deseada.

Una de las características del producto interior es la continuidad, es decir el producto interior es una función continua lo cual se muestra en el siguiente resultado (Friedberg et al., 1997), (Kreyszig, 1989).

Proposición 2.4 (Continuidad del Producto interior) *Sea* X *un espacio con producto interior y sean* $\{x_n\}$ *y* $\{y_n\}$ *sucesiones en* $\mathcal{H}$ *tales que* $x_n \to x$ *y* $y_n \to y$ *con* $x, y \in \mathcal{H}$*. Entonces*

$$\langle x_n, y_n\rangle \to \langle x, y\rangle.$$

Demostración*: Sean* $\{x_n\}$ *y* $\{y_n\}$ *sucesiones en* $\mathcal{H}$*, si sumamos y restamos el término* $\langle x_n, y\rangle$ *a la expresión* $|\langle x_n, y_n\rangle - \langle x, y\rangle|$*, después usamos la desigualdad del triángulo y la desigualdad de Cauchy-Schwarz, obtenemos:*

$$\begin{aligned} |\langle x_n, y_n\rangle - \langle x, y\rangle| &= |\langle x_n, y_n\rangle - \langle x_n, y\rangle + \langle x_n, y\rangle - \langle x, y\rangle| \\ &\leq |\langle x_n, y_n - y\rangle| + ||\langle x_n - x, y\rangle| \\ &\leq \|x_n\|\|y_n - y\| + \|x_n - x\|\|y\|. \end{aligned}$$

Como la sucesión $\{x_n\}$ *es convergente, entonces existe una constante* $M > 0$ *tal que* $\|x_n\| \leq M$ *y como* $\|x_n - x\| \to 0$ *y* $\|y_n - y\| \to 0$*, se concluye que:*

$$\langle x_n, y_n\rangle \to \langle x, y\rangle.$$

Definición 2.7 *Decimos que una sucesión* $\{x_n\}_{n\geq 1}$ *de elementos de un espacio vectorial normado* X *es una* ***sucesión de Cauchy****, si para cada* $\epsilon > 0$ *existe* $N \geq \mathbb{N}$ *tal que* $\|x_m - x_n\| < \epsilon$*, siempre que* $n, m > N$*. Decimos que la sucesión* $\{x_n\}_{n\geq 1}$ *converge a* x*, cuando* $\|x_n - x\| \to 0$*, siempre que* $n \to \infty$*, además si* $x \in X$*, entonces, decimos que la sucesión es convergente en* X*.*

Se puede verificar fácilmente que toda sucesión convergente en X es una sucesión de Cauchy. Sin embargo el recíproco no siempre es cierto, en el caso en que éste se cumpla el espacio X es llamado *espacio Completo* (Kreyszig, 1989).

Definición 2.8 *Decimos que un espacio vectorial normado es* ***completo****, si toda sucesión de Cauchy es convergente a un elemento en* X*.*

Definición 2.9 *Un espacio normado completo se denomina* ***espacio de Banach****.*

Podemos observar que una norma en X define una métrica d en X la cual está dada por:

$$d(x, y) = \|x - y\| \qquad x, y \in X, \tag{2.5}$$

y es llamada la métrica inducida por la norma, de modo que un espacio vectorial normado se denota por $(X, \|\cdot\|)$.

Los espacios de Hilbert permiten generalizar los espacios euclidianos empleados usualmente, para poder incluir espacios de dimensión infinita. Los espacios de Hilbert se definen en (Kreyszig, 1989) del siguiente modo .

Definición 2.10 (Espacio de Hilbert) *Un espacio de Hilbert* $\mathcal{H}$ *es un espacio vectorial sobre un campo* $\mathbb{F}$ *con un producto interior* $\langle\cdot,\cdot\rangle$ *tal que* $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ *es un espacio completo.*

Definición 2.11 (Subespacio de Hilbert) *Sea* $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ *un espacio de Hilbert. Decimos que* $\mathcal{H}_1$ *es un subespacio de Hilbert si*

1. $\mathcal{H}_1 \subseteq \mathcal{H}$*, y*
2. $(\mathcal{H}_1, \langle\cdot,\cdot\rangle)$ *es espacio de Hilbert.*

Notemos que en ambos espacios se tiene el mismo producto interior.

2.1.3. Ortonormalidad y Proyecciones

Como hemos dicho un espacio de Hilbert es un espacio normado completo (bajo la norma inducida por el producto interior), entonces debemos tener en cuenta que aquí se verifican todos los conceptos y propiedades topológicas de los espacios normados (Kreyszig, 1989). A continuación se definen conceptos importantes dentro de la teoría de los espacios de Hilbert como el concepto de ortogonalidad de vectores, ortonormalidad y proyección ortogonal.

Definición 2.12 (Ortogonalidad) *Sea* $\mathcal{H}$ *un espacio de Hilbert. Decimos que un elemento* $x \in \mathcal{H}$ *es* ***ortogonal*** *a* $y \in \mathcal{H}$ *si* $\langle x, y\rangle = 0$ *y lo denotamos por* $x \perp y$*. De igual forma, si* $S \subseteq \mathcal{H}$ *es tal que* $\langle x, y\rangle = 0$ *para todo* $x, y \in S \subseteq \mathcal{H}$*, decimos que* S *es un subconjunto ortogonal.*

Definición 2.13 (Ortonormalidad) *Sea* $\mathcal{H}$ *un espacio de Hilbert ortogonal. Si para toda* $x \in \mathcal{H}$ *se tiene que* $\|x\| = 1$*, entonces decimos que* $\mathcal{H}$ *es* ***ortonormal****.*

Cabe señalar que, si $S_1 \subseteq \mathcal{H}$ y $S_2 \subseteq \mathcal{H}$ son subconjuntos tales que para todo $x \in S_1$ y $y \in S_2$ se tiene que $x \perp y$, lo denotamos por $S_1 \perp S_2$ (Friedberg et al., 1997).

Una colección de vectores no necesariamente es un subespacio vectorial, pero si es posible encontrar un subespacio que contenga dicha colección de vectores como subconjunto. Algunos de los resultados sobre los espacios normados (y en particular los espacios de Hilbert), requieren la utilización de algunas propiedades referidas al campo escalar sobre el cual están definidos.

Proposición 2.5 *Sea U una colección de vectores en el espacio vectorial $\mathcal{H}$. Entonces existe un único subespacio B tal que*

(i) $U \subseteq B$

(ii) Si existe $B' \subseteq \mathcal{H}$ subespacio tal que $U \subseteq B'$, entonces $B \subseteq B'$.

***Demostración**: Sea $U \subseteq \mathcal{H}$.*

***Existencia**: Sea $\mathcal{D} = \{D \subseteq \mathcal{H} | D$ es subespacio vectorial y $U \subseteq D\}$ Consideremos el subconjunto $B = \bigcap_{D \in \mathcal{D}}$, entonces, tenemos que B es subespacio vectorial de $\mathcal{H}$ y además $U \subseteq B$, con lo cual se verifica la parte (i). Ahora si suponemos que B' es un subespacio tal que $U \subseteq B'$, entonces se tiene $U \subseteq \mathcal{D}$ y por tanto, $\bigcap_{D \in \mathcal{D}} \subseteq B'$, es decir, $B \subseteq B'$.*

***Unicidad**: Supongamos que existe un subespacio B^* que satisface las propiedades (i) y (ii), entonces, como B también satisface ambas propiedades, se tiene que $B^* \subseteq B \subseteq B*$, con lo cual $B = B^*$.*

Definición 2.14 (Espacio generado) *El subespacio dado por la proposición anterior se denomina **espacio generado** por U y se denota por $Gen[U]$.*

Observación 2.2 *Notemos que $Gen[U]$ se puede ver como el siguiente conjunto.*

$$Gen[U] = \left\{ \sum_{i \in I} \alpha_i u_i : \alpha_i \in \mathbb{F}, u_i \in U, i \in I \right\}.$$

De lo anterior se puede deducir que para un espacio de Hilbert $\mathcal{H}$ (y en general para cualquier espacio vectorial), con una base B de vectores se tiene que $Gen[B] = \mathcal{H}$. Ahora, también es fácil deducir que todo subconjunto ortonormal de vectores diferentes de cero dentro de un espacio de Hilbert, es también un subconjunto linealmente independiente. El siguiente resultado nos ayudará a encontrar una base ortonormal para los espacios de Hilbert de dimensión finita.

Teorema 2.1 (Proceso de Gram-Schmidt) *Sea $\mathcal{H}$ un espacio de Hilbert y sea $Y = \{y_1, \cdots, y_n\}$ un subconjunto de $\mathcal{H}$ de vectores linealmente independientes. Definamos $Y' = \{x_1, \cdots, x_n\}$ donde $x_1 = y_1$ y*

$$x_k = y_k - \sum_{i=1}^{k-1} \frac{(y_k, x_i)}{\|x_i\|^2} x_i, \qquad \text{para } 2 \leq k \leq n. \tag{2.6}$$

Entonces Y' es un conjunto ortogonal de vectores no nulos tales que $Gen[Y'] = Gen[Y]$.

***Demostración**: La demostración de este Teorema se hace por inducción sobre el número de vectores y se puede encontrar en (Friedberg et al., 1997).*

La construcción de $\{x_1, \cdots, x_n\}$ usando la ecuación (2.6) se llama *proceso de ortogonalización de Gram-Schmidt.* Este algoritmo se usa para obtener un conjunto ortogonal β de vectores no nulos con $Gen[\beta] = \mathcal{H}$. Dividiendo cada vector de β entre su norma se obtiene un conjunto ortonormal β' que genera a $\mathcal{H}$, al cual llamaremos base ortonormal para $\mathcal{H}$ (Friedberg et al., 1997).

Definición 2.15 (Complemento ortogonal) *Sea $\mathcal{H}$ un espacio de Hilbert y S un subconjunto de $\mathcal{H}$. La colección de vectores*

$$S^{\perp} = \{y \in \mathcal{H} | y \perp x, \text{para cada } x \in S\},$$

se denomina complemento ortogonal de S. Lo cual se denota como $S \perp S^{\perp}$.

Es fácil demostrar que $S^{\perp}$ es un subespacio de $\mathcal{H}$ para cualquier subconjunto S de $\mathcal{H}$.

Teorema 2.2 *Sea $\mathcal{H}$ un espacio de Hilbert y S un subespacio de $\mathcal{H}$ de dimensión finita. Entonces $\mathcal{H} = S \bigoplus S^{\perp}$.*

Teorema 2.3 (Teorema de proyección) *Sea $\mathcal{H}$ un espacio de Hilbert, S un subespacio vectorial cerrado de $\mathcal{H}$. Entonces para cada $x \in \mathcal{H}$ existe un único elemento $y_0 \in S$ tal que*

$$\|x - y_0\| = \text{ínf}\{\|x - y\| : y \in S\}.$$

El vector $y_0 \in S$ se conoce como la proyección de x sobre el subespacio S.

Sobre un subespacio S podemos denotar como $P_S(x)$ la proyección de x sobre S, para cada x en el espacio $\mathcal{H}$. Definimos entonces el operador proyección como $P_S : \mathcal{H} \to S$, tal que $x \mapsto P_S(x)$ para toda $x \in \mathcal{H}$.

2.1.4. Operadores lineales

Ahora introduciremos un nuevo tipo de funciones denominadas operadores lineales, es decir, funciones definidas sobre espacios vectoriales y normados.

Definición 2.16 *Un operador lineal T es una función con dominio $\mathcal{D}(T)$ y rango $\mathcal{R}(T)$, ambos son espacios vectoriales definidos sobre el mismo cuerpo $\mathbb{F}$ que satisface:*

- $T(x+y) = T_x + T_y$ *para todo* $x, y \in \mathcal{D}(T)$.
- $T(\alpha x) = \alpha T_x$ *para todo* $\alpha \in \mathbb{F}$ *y todo* $x \in \mathcal{D}(T)$.

Definición 2.17 *Sea $T : \mathcal{D}(T) \subset X \to Y$ un operador lineal. T es acotado si existe $c > 0$ tal que*

$$\|T_x\| \leq c\|x\| \tag{2.7}$$

para todo $x \in \mathcal{D}(T)$. En otras palabras, T envía conjuntos acotados en conjuntos acotados.

En (2.7) la norma de la izquierda es la norma en el espacio Y y la norma de la derecha es la asociada al espacio X. Por simplicidad hemos denotado ambas normas por el mismo símbolo $\|\cdot\|$, sin peligro de confusión.

Definición 2.18 *Un funcional lineal es un operador lineal con dominio un espacio vectorial y con rango en el espacio de los escalares, es decir, es una transformación que aplica vectores en escalares.*

$$f : \mathcal{D}(f) \to \mathbb{F}.$$

Definición 2.19 *Si $x \neq 0$, entonces $\frac{\|T_x\|}{\|x\|} \leq c$ para todo $x \neq 0$, $x \in \mathcal{D}(T)$. Definimos*

$$\|T\| := \sup\left\{\frac{\|T_x\|}{\|x\|} : X \neq 0, x \in \mathcal{D}(T)\right\}.$$

donde $\|T\|$ es llamada la ***norma del operador T****.*

Observación 2.3 *Si T es un operador lineal acotado y $\|T\| = c$, entonces*

$$\|T_x\| \leq \|T\|\|x\|.$$

Teorema 2.4 *Sea X un espacio normado, si $Dim(X) < \infty$. Entonces todo operador lineal es acotado.*

***Demostración**: Sea $\{e_1, \cdots, e_n\}$ una base de X. Sea $x \in X$, $x = \alpha_1 e_1, \cdots, \alpha_n e_n$. Sea $T : X \to X$ un operador lineal, entonces*

$$\begin{aligned}\|T_x\| &= \|T(\alpha_1 e_1, \cdots, \alpha_n e_n)\| \\ &= \|\alpha_1 T e_1, \cdots, \alpha_n T e_n\| \\ &\leq |\alpha_1|\|Te_1\| + \cdots + |\alpha_n|\|Te_n\| \\ &\leq \max_{1\leq i\leq n} \|Te_i\|(|\alpha_1| + \cdots + |\alpha_n|) \\ &\leq c^{-1} \max_{1\leq i\leq n} \|Te_i\|\|x\|\end{aligned}$$

donde $c^{-1}\,\text{máx}_{1\leq i\leq n}\,\|Te_i\|$ es constante, así T es acotado.

2.2. Espacios de Hilbert con Kernel reproductor (RKHS)

Los espacios de Hilbert con Kernel reproductor (RKHS, por sus siglas en inglés, Reproducing kernel Hilbert spaces) representan un marco conceptual para poder abordar diferentes problemas (Aronszajn, 1950), (Preda, 2007). En esta tesis este tipo de espacios se utilizarán para representar funciones las cuales provienen de un conjunto de datos de naturaleza discreta.

A partir de ahora denotaremos por $\mathcal{H}$ al espacio de Hilbert de funciones $f : X \to \mathbb{F}$, donde X generalmente es un subconjunto de $\mathbb{R}$ o $\mathbb{C}$ dotado de un producto interior.

Para cada $x \in X$, el funcional lineal

$$\begin{aligned}E_x :&\mathcal{H} \to \mathbb{F} \\ &f \mapsto E_x(f) = f(x),\end{aligned}$$

es continuo y se le llama funcional evaluación en el punto x.

Definición 2.20 (RKHS) *Decimos que un espacio de Hilbert $\mathcal{H}$ de funciones definidas en un conjunto compacto X es un (RKHS) si todos los funcionales evaluación son acotados en $\mathcal{H}$. Es decir, para cada $x \in X$ existe una constante M_x tal que*

$$|F_x| = |f(x)| \leq M_x\|f\| \quad \text{para cada función } f \in RKHS \text{ con } x \in X, \tag{2.8}$$

donde $\|\cdot\|$ es la norma asociada en el espacio de Hilbert (Aronszajn, 1950).

Observación 2.4 *Si para cada $x \in X$, se cumple que todos lo funcionales evaluación son acotados, por el Teorema de Representación de Riez que se encuentra en (Kreyszig, 1989) se tiene que existe un único elemento $K_x \in \mathcal{H}$ tal que $f(x) = \langle k_x, f\rangle$ para todo $f \in \mathcal{H}$ y además $\|E_y\| = \|K_x\|$.*

Lo anterior nos lleva a una nueva definición.

Definición 2.21 (Kernel reproductor) *Sea $\mathcal{H}$ un espacio de Hilbert de funciones definidas en $X \subset \mathbb{R}^n$ con producto interior $\langle \cdot, \cdot \rangle$. Una función $K : X \times X \to \mathbb{F}$ continua y simétrica definida por*

$$K(x, y) := \langle K_x, K_y \rangle = K_x(y), \qquad (x, y) \in X \times X,$$

se denomina Kernel reproductor de $\mathcal{H}$ si satisface las siguientes condiciones:

- *Para cada $x \in X$ fijo, la función $K(x, \cdot) = K_x(\cdot) \in \mathcal{H}$.*
- ***Propiedad reproductiva**: Para cada $x \in X$ y $f \in \mathcal{H}$*

$$f(x) = \langle K_x, f \rangle.$$

Observemos que si la función Kernel K es real; es decir, si $K(x, y) \in \mathbb{R}$ para toda $x, y \in \mathbb{F}$, entonces $K(x, y) = K(y, x)$, así $K(y, \cdot) \in \mathcal{H}$ para todo $x, y \in X$. De modo que la propiedad reproductiva se puede escribir como

$$f(x) = \langle f(\cdot), K(x, \cdot) \rangle.$$

Veamos ahora algunas propiedades de las funciones Kernel.

Proposición 2.6 *El Kernel reproductor de un espacio RKHS $\mathcal{H}$ es único.*
***Demostración**: Supongamos que $K(x, y) = K_y(x)$ y $K'(x, y) = K'_y(x)$ son dos Kernel reproductores para $\mathcal{H}$, entonce para $x, y \in X$ se tiene*

$$K'_y(x) = \langle K'_y, K_x \rangle = \overline{\langle K_x, K'_y \rangle} = \overline{K_x(y)} = \langle K_y, K_x \rangle = K_y(x),$$

de aquí tenemos que $K' = K$.

Ahora si conocemos una base ortonormal $\{e_n(x)\}_{n=1}^{\infty}$ en $\mathcal{H}$ es fácil encontrar una expresión para K, lo cual se muestra en la siguiente proposición.

Proposición 2.7 *Sea $\{e_n(x)\}_{n=1}^{\infty}$ una base ortonormal de $\mathcal{H}$. Para cada $x, y \in X$ se tiene la siguiente expresión*

$$K(x, y) = \sum_{n=1}^{\infty} e_n(x)\overline{e_n(y)}, \qquad x, y \in \mathbb{R}.$$

***Demostración**: Si desarrollamos K_x y K_y en la base ortonormal $\{e_n(x)\}_{n=1}^{\infty}$ de $\mathcal{H}$ se obtiene que $K_x = \sum_{n=1}^{\infty} \langle K_x, e_n \rangle e_n$ y $K_y = \sum_{n=1}^{\infty} \langle K_y, e_n \rangle e_n$ donde se puede escribir*

$$K(x, y) = \langle K_y, K_x \rangle = \sum_{n=1}^{\infty} \langle K_y, e_n \rangle \overline{\langle K_y, e_n \rangle} = \sum_{n=1}^{\infty} \overline{e_n(y)} e_n(x).$$

Definición 2.22 *Dado $K : X \times X \to \mathbb{F}$ y entradas $x_1, x_2, \cdots, x_n \in X$, la matriz K con elementos $K_{ij} := K(x_i, x_j)$ con $i, j = 1, \cdots, n$, se denomina* ***Matriz Gram*** *de K (o matriz de kernels) respecto a $x_1, x_2, \cdots, x_n \in X$.*

Definición 2.23 *Una matriz $K_{n\times n}$ compleja, que satisface*

$$\sum_{i,j} c_i \bar{c}_j K_{ij} \geq 0, \tag{2.9}$$

para todo c_i en $\mathbb{C}$ se denomina Matriz Definida Positiva. De manera similar, una matriz $K_{n\times n}$ real y simétrica, que cumple con 2.9 para todo $c_i \in \mathbb{R}$ se denomina Definida Positiva (Kreyszig, 1989).

Definición 2.24 *Sea X no vacío. Una función K en $X \times X$ tal que para todo $i \in \mathbb{N}$ y para todo $x_i \in X$, da lugar a una matriz definida positiva, se denomina* ***Kernel Definido Positivo****, o simplemente Kernel.*

Proposición 2.8 *Si K es un Kernel definido positivo y $x_i, x_j \in X$, entonces*

(i) $K(x_i, x_i) \geq 0$*, para todo $x_i \in X$.*

(ii) $K(x_i, x_j) = \overline{K(x_j, x_i)}$.

(iii) $|K(x_i, x_j)|^2 \leq K(x_i, x_i)K(x_j, x_j)$.

Demostración*:*
(i) Tenemos

$$K(x_i, x_i) = \langle K(\cdot, x_i), K(\cdot, x_i)\rangle = \|K(\cdot, x_i)\| \geq 0.$$

(ii) Sean $x_i, x_j \in$

$$K(x_i, x_j) = \langle K(\cdot, x_i), K(\cdot, x_j)\rangle = \overline{\langle K(\cdot, x_j), K(\cdot, x_i)\rangle} = \overline{K(x_j, x_i)}.$$

(iii) Sean $x_i, x_j \in X$ y construyamos su correspondiente matriz de kernels

$$\begin{bmatrix} K(x_i, x_i) & K(x_i, x_j) \\ K(x_j, x_i) & K(x_j, x_j) \end{bmatrix}.$$

Se sabe que los determinantes de una matriz definida positiva son positivos, de modo que

$$\begin{aligned} K(x_i, x_i)K(x_j, x_j) - K(x_i, x_j)K(x_j, x_i) &\geq 0. \\ K(x_i, x_i)K(x_j, x_j) - K(x_i, x_j)\overline{K(x_i, x_j)} &\geq 0. \\ K(x_i, x_i)K(x_j, x_j) - |K(x_i, x_j)|^2 &\geq 0. \end{aligned}$$

Así tenemos:

$$|K(x_i, x_j)|^2 \leq K(x_i, x_i)K(x_j, x_j).$$

Ejemplo 2.4 (Kernels reproductivos) *Sea $X \subset X$. Los kernels más comunes en estadística multivariada son los siguientes:*

(i) ***Kernel Lineal****: $K(x,y) = \langle x, y\rangle$.*

(ii) ***Kernel Polinomial****: $K(x,y) = \langle x, y\rangle^d$.*

(iii) ***Kernel Laplaciano (Exponencial)****: $K(x,y) = \exp\left(\frac{-\|x-y\|^2}{\sigma}\right)$.*

(iv) ***Kernel Gaussiano****: $K(x,y) = \exp\left(\frac{-\|x-y\|^2}{2\sigma^2}\right)$.*

Una propiedad importante de los RKHS que relaciona la convergencia en norma con la convergencia puntual en X, se da en el siguiente resultado.

Proposición 2.9 *En un RKHS $\mathcal{H}$ la convergencia en norma implica convergencia puntual en X, la cual es uniforme en subconjuntos de X en donde la función $x \mapsto K(x,x)$ está acotada.*
Demostración*: Sea $\{f_n\}$ una sucesión en $\mathcal{H}$ tal que $f_n \to f$ cuando $n \to \infty$. Como $\mathcal{H}$ es un RKHS satisface la propiedad de reproducción con la función $f_n - f \in \mathcal{H}$, con lo cual obtenemos $f_n(x) - f(x) = \langle f_n - f, K(x,\cdot)\rangle$. Finalmente, la desigualdad de Cauchy-Schwartz nos permite escribir*

$$|f_n(x) - f(x)| \leq \|f_n - f\|\|K(x,\cdot)\| = \sqrt{K(x,x)}\|f_n - f\| \to 0, \text{ cuando } n \to \infty.$$

Además la convergencia puntual será uniforme en subconjuntos de X donde $\sqrt{K(x,x)}$ está acotado.

Supongamos que nuestro RKHS $\mathcal{H}$ es un subespacio (cerrado) de un espacio de Hilbert $\widetilde{\mathcal{H}}$ (no necesariamente con Kernel reproductor). En este caso, tenemos el siguiente resultado:

Proposición 2.10 *Si el RKHS $\mathcal{H}$ es un subespacio cerrado de un espacio de Hilbert $\widetilde{\mathcal{H}}$. Entonces*

$$\langle f, K(x,\cdot)\rangle = P_{\mathcal{H}} f(x), \text{ para toda función } f \in \widetilde{\mathcal{H}},$$

donde $P_{\mathcal{H}}$ denota la proyección ortogonal sobre $\mathcal{H}$.
Demostración*: Dada $f \in \widetilde{\mathcal{H}}$ tenemos $f = f_1 + f_2$ con $f_1 \in \mathcal{H}$ y $f_2 \in \mathcal{H}^{\perp}$, de donde*

$$\langle f, K(x,\cdot)\rangle = \langle f_1 + f_2, K(x,\cdot)\rangle = \langle f_1, K(x,\cdot)\rangle + \langle f_2, K(x,\cdot)\rangle = f_1(x) = P_{\mathcal{H}} f(x),$$

ya que $f_2 \perp K(x,\cdot)$ y $f_1 \in \mathcal{H}$.

Supongamos que existe una sucesión $\{x_n\}_{n=1}^{\infty} \in X$ tal que $\{K(x_n,\cdot)\}_{n=1}^{\infty}$ es una base ortogonal de $\mathcal{H}$. Existe una fórmula en $\mathcal{H}$ que nos permite recuperar cada función $f \in \mathcal{H}$ a partir de la sucesión de sus muestras $\{f(x_n)\}_{n=1}^{\infty}$.

2.2.1. Teorema de Moore-Aronszajn

Como hemos visto un espacio de Hilbet que admite un Kernel reproductor se denomina RKHS. El *Teorema de Moore-Aronszajn* en (Aronszajn, 1950) establece una relación biunívoca entre kernels y espacios RKHS. Para cada espacio RKHS de funciones en X existe un único Kernel reproductor K definido positivo. De manera recíproca si K es un Kernel simétrico y definido positivo, entonces genera un único RKHS en el cual el Kernel dado actúa como Kernel reproductor (Preda, 2007).

Teorema 2.5 (Teorema de Moore-Aronszajn)
Sea K un Kernel reproductor definido positivo y simétrico sobre un conjunto X. Entonces, existe un único espacio de Hilbert de funciones en X para el cual K es un Kernel reproductor, es decir, un Kernel reproductor definido positivo y simétrico determina un único espacio de Hilbert con Kernel reproductor.
***Demostración**: Para cada $x \in X$ definimos $K_x = K(x, \cdot)$. Sea $\mathcal{H}_0$ el espacio generado por $\{K_x : x \in X\}$. Definimos un producto interno sobre $\mathcal{H}_0$ tal que*

$$\left\langle \sum_{i=1}^{n} \beta_i K_{y_i}, \sum_{j=1}^{m} \alpha_j K_{x_j} \right\rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} \bar{\alpha}_j \beta_i K(y_i, x_j).$$

Ahora como este producto interior es simétrico obtenemos la simetría de K. Sea $\mathcal{H}$ la compleción de $\mathcal{H}_0$ respecto al producto interior definido, entonces las funciones de $\mathcal{H}$ son de la forma

$$f(x) = \sum_{i=1}^{\infty} \alpha_i K(x_i, \cdot),$$

donde $\sum_{i=1}^{\infty} \alpha_i^2 K(x_i, x_i) < \infty$ por la desigualdad de Cauchy-Schwarz.

Veamos que se cumple la propiedad reproductiva.

$$\begin{aligned} \langle f, K_x \rangle &= \left\langle \sum_{i=1}^{\infty} \alpha_i K_{x_i}, K_x \right\rangle \\ &= \sum_{i=1}^{\infty} \alpha_i K(x_i, x) \\ &= f(x). \end{aligned}$$

Ahora veamos que $\mathcal{H}$ es único. Supongamos que existe $\widetilde{\mathcal{H}}$ otro espacio de Hilbert de funciones para el cual K es un Kernel reproductor. Para todo $x, y \in X$, por la propiedad reproductiva se tiene que

$$\langle K_x, K_y \rangle_{\mathcal{H}} = K(x, y) = \langle K_x, K_y \rangle_{\widetilde{\mathcal{H}}}.$$

Ya que el producto interior es lineal $\langle \cdot, \cdot \rangle_{\mathcal{H}} = \langle \cdot, \cdot \rangle_{\widetilde{\mathcal{H}}}$ en el espacio generado por $\{K_x : x \in X\}$. Así tenemos que $\mathcal{H} = \widetilde{\mathcal{H}}$ por la unicidad de la compleción.

El recíproco de este Teorema también se cumple, es decir, dado un RKHS existe un único Kernel que lo caracteriza (Preda, 2007).

Observación 2.5 (Generación de $\mathcal{H}_K$ a partir de K) *Sea $\mathcal{H}'$ el conjunto de todas las combinaciones lineales finitas de la forma $f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$ donde $n \in \mathbb{N}$, $x_i \in X$ y $\alpha_i \in \mathbb{R}$ dotado con el producto interior*

$$\langle f, g \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \beta_j K(x_i, x_j).$$

donde $f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$ y $g(x) = \sum_{j=1}^{n} \beta_j K(x_j, x)$. Entonces $\mathcal{H}_\mathcal{K}$ es la compleción de $\mathcal{H}'$ con el producto interior asociado, es decir, se deben añadir al conjunto $\mathcal{H}'$ los limites de todas las sucesiones de Cauchy.

A continuación definiremos otra caracterización de los RKHS basada en funciones propias de operadores lineales

Definición 2.25 (Operador integral) *Sea $L_\mu^2(X)$ el espacio de las funciones cuadrado integrables en X, donde μ es una medida de Borel. Sea $K : X \times X \to \mathbb{R}$ una función continua. Definimos la aplicación lineal $L_K : L_\mu^2(X) \to \mathcal{C}(X)$ donde $\mathcal{C}(X)$ es el conjunto de las aplicaciones continuas en X por*

$$L_K(f)(x) = \int_X K(x, t) f(t) d_\mu(t), \tag{2.10}$$

esta función está bien definida y la función K de denomina Kernel del operador L_K.

Definición 2.26 (Funciones y valores propios de L_K) *Diremos que ϕ_j es una función propia de L_K y el valor λ_j es su correspondiente valor propio si cumple*

$$L_K(\phi_j) = \lambda_j \phi_j \tag{2.11}$$

es decir,

$$\phi_j(x) = \frac{1}{\lambda_j} \int_X K(x, t) \phi_j(t) d_\mu(t). \tag{2.12}$$

2.3. Métodos Kernel

La construcción que se ha venido desarrollando muestra que cualquier Kernel definido positivo se puede considerar como una función que evalúa los productos internos en otro espacio, es decir $K(x, x') = \langle \phi(x), \phi(x') \rangle$. Por lo tanto, el espacio de producto interior construido de esta manera es un espacio de características asociado a un Kernel K.

Supóngase ahora que se parte de forma inversa, es decir, a partir de un mapeo del espacio original a un espacio de producto interno. En este caso se obtiene un Kernel definido positivo vía $K(x, x') = \langle \phi(x), \phi(x') \rangle$, esta aplicación

sólo se conoce implícitamente y permite asignar datos de un espacio finito $X \subset \mathbb{R}^p$ a un espacio de dimensión infinita denominado el *feature space*(espacio de características) el cual denotaremos por $\mathfrak{F}$.

Realizar particionamientos lineales en el feature space nos conduce a particionamientos no lineales en el espacio de datos originales. El siguiente Teorema nos permitirá calcular el producto interno en el conjunto de puntos proyectados en el feature space sin necesidad de conocer explícitamente la aplicación ϕ, este hecho se conoce como el *Truco Kernel* (Gibaja Martíns, 2010).

Teorema 2.6 (Teorema de Mercer) *Sea X un conjunto compacto y μ una medida de Borel en X y $K : X \times X \to \mathbb{R}$ un Kernel definido positivo y simétrico. Sea $\{\phi_i\}_{i\geq 1}$ y $\{\lambda_j\}_{j\geq 1}$ respectivamente las correspondientes funciones y valores propios de L_K. Entonces para todo $x, y \in X$*

$$K(x,y) = \Sigma_{j=1}^{\infty} \lambda_j \phi_j(x)\phi_j(y) = \langle \phi(x), \phi(y) \rangle, \tag{2.13}$$

donde las series convergen absolutamente para cada $(x,y) \in X \times X$ y uniformemente en $X \times X$.

Este resultado nos permite calcular distancias en $\mathfrak{F}$ mediante el llamado truco Kernel,

$$\begin{aligned} \|\phi(x) - \phi(y)\|^2 &= \langle \phi(x) - \phi(y), \phi(x) - \phi(y) \rangle \\ &= \langle \phi(x), \phi(x) \rangle + \langle \phi(y), \phi(y) \rangle - 2\langle \phi(x), \phi(y) \rangle \\ &= K(x,x) + K(y,y) + 2K(x,y), \end{aligned}$$

como una función de los datos originales (Gibaja Martíns, 2010).

2.4. Aplicación de los Métodos Kernel

Los métodos estadísticos tradicionales (métodos lineales) fallan tan pronto como tratamos los métodos no lineales con datos funcionales. De hecho, si, por ejemplo, consideramos una muestra de curvas finamente discretizadas, aparecen dos problemas estadísticos cruciales.

- El primero, proviene de la relación entre el tamaño de la muestra y el número de variables.
- El segundo, se debe a la existencia de correlaciones (multicolinealidad) entre las variables y se convierte en un problema mal condicionado en el contexto del modelo lineal multivariante.

Por lo tanto, existe la necesidad de desarrollar nuevos métodos (modelos) estadísticos con el fin de tener en cuenta la estructura de este tipo de datos.
.

Los métodos no lineales en estadística permiten explorar la estructura de los datos y la mayoría de estos métodos asumen que los datos se pueden representar como puntos en el espacio euclidiano $\mathbb{R}^n$. En este trabajo se supone que los datos que se tienen son funciones $f : X \subset \mathbb{R}^k \to \mathbb{R}$ y que la información que se tiene para cada una de ellas es un muestreo en un conjunto finito de puntos.

La precisión en los resultados obtenidos con las técnicas de aprendizaje basado en la teoría matemática de los Espacios de Hilbert con Kernel Reproductor (RKHS), han demostrado ser métodos no lineales robustos, y han despertado un considerable interés en extender esta metodología para generalizar las técnicas de análisis multivariado lineal, como los métodos de Análisis de Componentes Principales (PCA), Análisis de Correlación Canónica (CCA), Mínimos Cuadrados Parciales (PLS), entre otros (Montano Rivas, 2013), (Schölkopf et al., 1998).

Los métodos basados en transformaciones implícitas (llamados métodos kernel) se han vuelto muy populares para analizar patrones no lineales en conjuntos de datos provenientes de diversos campos de estudio. Más aún, la introducción de funciones kernel se ha vuelto una alternativa eficiente para obtener medidas de similaridad entre objetos que no tienen una representación vectorial natural. Aunque la aplicación más conocida de métodos kernel es en Máquinas de Soporte Vectorial (SVM), se ha mostrado últimamente que cualquier algoritmo de aprendizaje basado en distancias entre objetos puede formularse en términos de funciones kernel, aplicando el llamado "truco del kernel"(Gibaja Martíns, 2010).

El método kernel se puede aplicar en todos los algoritmos de análisis de datos cuyas entradas se pueden expresar en términos de productos punto. Si los datos en el espacio original no se pueden analizar satisfactoriamente con las técnicas de análisis multivariado, la estrategia para extenderlo a modelos no lineales usando métodos kernel, se basa en la idea aparentemente paradójica de transformar los datos, mediante una función no lineal, hacia un espacio de mayor dimensión al espacio en el que se encuentran los datos y realizar el análisis multivariado en los datos transformados.

Al conjunto de elementos original lo llamaremos *Espacio de Entrada* (Input Space) y lo denotaremos por X. No es necesario que este conjunto esté dotado de una estructura algebraica particular, por tanto, puede ser un espacio vectorial (en el que los elementos de la base de datos aparecen representados en función de los valores que toman en un conjunto de variables de naturaleza cuantitativa) como un conjunto sin dicha estructura. Como veremos, la clave del enfoque kernel consiste en definir una función que a cada pareja de elementos de este espacio X se le haga corresponder un valor real.

El proceso para construir un *espacio de Características* (Feature Space), denotado por $\mathcal{H}$ asociado al mapeo ϕ, considera los siguientes pasos:

- Convertir la imagen de ϕ en un espacio vectorial.
- Definir un producto interno $\langle \cdot, \cdot \rangle$ en $\mathcal{H}$, el cual debe satisfacer:

$$K(x, x') = \langle \phi(x), \phi(x') \rangle.$$

  Es decir debemos definir un kernel reproductor.
- Con el Kernel definido en el punto anterior, se construye el espacio de características el cual, será un RKHS asociado al kernel K elegido.

Observación 2.6 (Truco Kernel) *El truco kernel permite realizar las operaciones algebraicas en el espacio de los datos transformados de manera eficiente y sin conocer a la transformación ϕ. Así, en principio, cualquier técnica de análisis multivariado para datos en $X \subset \mathbb{R}^k$ que se pueda formular en un algoritmo computacional en términos de productos punto se puede generalizar a los datos transformados utilizando el truco Kernel (Montano Rivas, 2013), (Schölkopf et al., 1998).*

La Kernelización de un algoritmo consiste en su reformulación, de manera que la determinación de una pauta o regularidad lineal en los datos pueda llevarse a cabo a partir, exclusivamente, de la información recogida en los productos escalares calculados para todas las parejas de elementos del espacio (Gibaja Martíns, 2010).

2.5. Técnica de Componentes Principales

Las técnicas tradicionales del análisis multivariado funcionan adecuadamente bajo ciertas condiciones de los datos. En la Regresión con Componentes principales (RCP) se busca representar a los datos en un subespacio vectorial de dimensión menor al espacio original mediante la construcción de variables que son combinaciones lineales de las variables originales, llamadas componentes principales, con la restricción de que estas nuevas variables tengan varianza máxima y estén incorrelacionadas.

La regresión con Componentes Principales con Kernels (RCPK) es una extensión de la regresión con Componentes Principales (RCP) que consiste en enviar los datos mediante una transformación no lineal, a otro espacio, llamado espacio de características, y realizar el análisis en este espacio. La clave del éxito de la RCPK está en lograr la extracción de direcciones de máxima variabilidad en el espacio de las características y luego identificar estas direcciones con las direcciones (no lineales) de variabilidad de los datos en el espacio original. Sin embargo, existen situaciones donde el ACPK no es suficiente para detectar estas direcciones no lineales de máxima variabilidad.

2.5.1. Regresión con Componentes Principales

Para estudiar las relaciones que se presentan entre variables correlacionadas (que miden información común) se puede transformar el conjunto original de variables en otro conjunto de nuevas variables *no* correlacionadas entre sí (que no tenga repetición o redundancia en la información) llamado conjunto de *componentes principales* (Montgomery et al., 2004).

La regresión de componentes principales es un método matemático que aplica mínimos cuadrados sobre un nuevo conjunto de variables, obtenidas a partir de la matriz de correlaciones R.

La forma canónica del modelo es:

$$Y = Z\alpha + \varepsilon, \tag{2.14}$$

donde:

$$Z = XT, \;\; \alpha = T'\beta, \;\; T'X'XT = Z'Z = \Lambda. \tag{2.15}$$

Recordemos que $\Lambda = diag(\lambda_1, \lambda_2, \cdots, \lambda_k)$ es una matriz diagonal de $k \times k$ de los eigenvalores de R y T es una matriz ortogonal cuyas columnas son los eigenvectores asociados con $\lambda_1, \lambda_2, \cdots, \lambda_k$.

Las columnas de Z definen un nuevo conjunto de regresores ortogonales, $Z = [Z_1, Z_2, \cdots, Z_k]$ llamados *componentes principales*.

En general, se extraen k componentes (variables)$Z_1, Z_2, \cdots, Z_k$ no correlacionadas y cuyas varianzas son igual a la varianza total, es decir $tr(\Lambda) = \sum_j \lambda_j$.

De estas componentes se espera que las primeras (usualmente) expliquen una alta proporción de la varianza total, es decir se espera que la proporción de varianza explicada sea alta. Uno de los problemas que resuelve la *RCP*, es que reduce la dimensionalidad del problema, para lo cual debemos poner nuestra atención sobre las primeras componentes principales, pero sin perder mucha variabilidad.

Después del diagnóstico de multicolinealidad señalado en el Capítulo 1, deben realizarse dos pasos en el procedimiento de *RCP* (Vega-Vilca and Guzmán, 2011):

- Realizar un Análisis de Componentes Principales con las variables predictoras del modelo.
- Hacer un nuevo Análisis de Regresión Múltiple utilizando las componentes más importantes como variables explicativas y la misma variable respuesta.

2.5.2. Regresión con Componentes Principales con Kernels

En el campo de la estadística multivariada, la regresión de componentes principales con Kernel (RCPK) es una extensión de la RCP utilizando técnicas de métodos kernel. Es decir, debemos hacer un análisis de componentes principales con kernel(ACPK)y posteriormente hacer un nuevo análisis de regresión, usando ahora las componentes obtenidas con el ACPK. El uso de un kernel, hace que las operaciones originalmente lineales de RCP se realicen en un RKHS (Gibaja Martíns, 2010).

La regresión con kernels (RCPK) tiene como objetivo extraer los patrones de máxima variabilidad no lineales de un conjunto de datos transformándolos mediante una función no lineal y haciendo un ACP, mediante el truco kernel, en el espacio de los datos transformados (Montano Rivas, 2013), (Schölkopf et al., 1998). Los detalles del ACPK son los siguientes:

Supongamos que se tiene una muestra de observaciones , centradas, es decir, $\Sigma_{i=1}^{n} X_i = 0$ donde $X_i \in X \subset \mathbb{R}^k$. Sea $\phi : \mathbb{R}^k \to \mathcal{H}$ un mapeo no lineal, donde $\mathcal{H}$ es un espacio de Hilbert de dimensión mayor que k. El ACP se hace con los datos transformados $\{\phi(X_i)\}$. De igual modo supongamos que los datos transformados están centrados, es decir, $\Sigma_{i=1}^{n} \phi(X_i) = 0$. En el ACPK se debe diagonalizar la matriz de varianzas y covarianzas de los datos trasformados, usando la matriz de covarianza

$$C = \frac{1}{k} \sum_{i=1}^{k} \phi(X_i)\phi(X_i)'. \tag{2.16}$$

Ahora se procede a encontrar los eigenvalores $\lambda \geq 0$ y los eigenvectores $V \in \mathcal{H} \setminus \{0\}$ de C, que satisfacen

$$\lambda V = CV.$$

Notemos que las soluciones de V con $\lambda \neq 0$ caen en el espacio generado por $\{\phi(X_1), \cdots, \phi(X_k)\}$, lo cual tiene dos consecuencias útiles. Primero se puede considerar la ecuación equivalente

$$\lambda(\phi(X_i) \cdot V) = (\phi(X_i) \cdot CV), \text{ para todo } i = 1, \cdots, k, \tag{2.17}$$

y segundo, como los eigenvectores se pueden expresar como una combinación lineal de los $\{\phi(X_1), \cdots, \phi(X_k)\}$, tenemos que existen coeficientes $\alpha = (\alpha_1, \cdots, \alpha_k)$ tales que

$$V = \sum_{i=1}^{k} \alpha_i \phi(X_i). \tag{2.18}$$

Combinando las ecuaciones (2.17) y (2.18), tenemos

$$\lambda \sum_{i=1}^{k} \alpha_i(\phi(X_n) \cdot \phi(X_i)) = \frac{1}{k} \sum_{i=1}^{k} \alpha_i \left(\phi(X_n) \cdot \sum_{j=1}^{k} \phi(X_j) \right) (\phi(X_j) \cdot \phi(X_i)),$$

para todo $n = 1, \cdots, k$.

Definimos la matriz $\boldsymbol{K}_{(k\times k)}$ cuyos elementos están dados por

$$\boldsymbol{K}_{ij} = (\phi(X_i) \cdot \phi(X_i)).$$

La matriz $\boldsymbol{K}$ se llama matriz Gram la cual se definió en la Sección 2.2. Sustituyendo en (2.16),(2.17) y (2.18) se tiene

$$k\lambda \boldsymbol{K}\alpha = \boldsymbol{K}^2\alpha. \tag{2.19}$$

Como $\boldsymbol{K}$ es una matriz simétrica, sus eigenvectores generan todo el espacio y por lo tanto proporcionan todas las soluciones α de la ecuación (2.19), donde α denota los vectores columna con entradas $\alpha_1, \cdots, \alpha_k$ y se resuelve el problema del eigenvalor

$$k\lambda\alpha = \boldsymbol{K}\alpha, \tag{2.20}$$

para eigenvalores diferentes de cero.

Sean $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$ los eigenvalores de $\boldsymbol{K}$ (es decir, las soluciones $k\lambda$ de la ecuación (2.20)) y $\alpha^1, \cdots, \alpha^k$ corresponden al conjunto completo de eigenvectores, con λ_k eigenvalores siendo el primero no nulo.

Ahora se deben normalizar $\alpha^1, \cdots, \alpha^k$, para ello requerimos que los correspondientes vectores en $\mathcal{H}$ sean normalizados, es decir $(V^q, V^q) = 1$ para todo $q = 1, \cdots, k$.

En virtud de las ecuaciones (2.18) y (2.20), esto se traduce en una condición de normalización para $\alpha^1, \cdots, \alpha^k$:

$$\begin{aligned} 1 &= \sum_{i,j=1}^{k} \alpha_i^n \alpha_j^n (\phi(X_i) \cdot \phi(X_j)) = \sum_{i,j=1}^{k} \alpha_i^n \alpha_j^n K_{ij} \\ &= \alpha^n \cdot K\alpha^n = \lambda_n(\alpha^n \cdot \alpha^n). \end{aligned}$$

Después de extraer los componentes principales en ACPK, se procede a calcular las proyecciones sobre los eigenvectores V^q en $\mathcal{H}$. Si tenemos un punto de prueba x, con imagen $\phi(x)$ en $\mathcal{H}$, entonces

$$(V^q \cdot \phi(x)) = \sum_{i=1}^{k} \alpha_i^q (\phi(X_i) \cdot \phi(x)) = \sum_{i=1}^{k} \alpha_i^q K(X_i, x),$$

es la proyección ortogonal de $\phi(x)$ sobre V^q.

En resumen, para realizar el ACPK se siguen los siguientes pasos:

1. Seleccionar la función kernel que representa el producto punto de los datos transformados $\langle \phi(x) \cdot \phi(y) \rangle$.

2. Calcular la matriz $\boldsymbol{K}$ con elementos $k(x_i, x_j)$, $i, j = 1, \cdots, n$.

3. Diagonalizar $\boldsymbol{K}$ (encontrar $\boldsymbol{K} = U\Lambda U'$, donde U es la matriz formada por los eigenvectores de $\boldsymbol{K}$ y Λ es la matriz con los correspondientes eigenvalores en la diagonal principal).

4. Calcular los primeros m eigenvectores $\alpha_j = \frac{u_j}{\lambda_j}$, $j = 1, \cdots, m$.

5. Calcular las componentes principales en el espacio de características

$$z_{ij} = P_{V_j}(\phi(x_i)) = \Sigma_{i=1}^{n} \alpha_{ij} k(x_i, x_j), \qquad j = 1, \cdots, m; i = 1, \cdots, n.$$

Posterior al procedimiento señalado se deberá realizar nuevamente un análisis de regresión.

La idea básica del ACPK se muestra en la Figura 2.1. En algún espacio de características de alta dimensionalidad F (abajo a la derecha), estamos realizando ACP lineal, al igual que en el espacio de entrada (parte superior). Dado que F está relacionado de manera no lineal con el espacio de entrada, se muestran las líneas de contorno de las Proyecciones constantes en el vector propio principal (dibujado como una flecha). Debemos tener en cuenta que no se puede extraer la imagen del vector propio en el espacio de entrada, porque puede probablemente no exista.

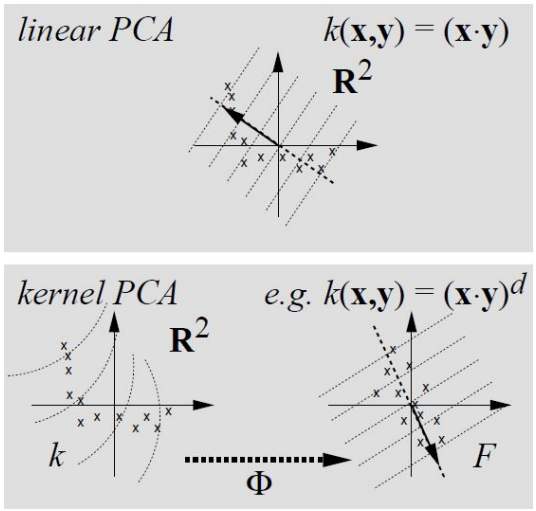

Figura 2.1: Descripción gráfica del ACPK (Schölkopf et al., 1998).

2.6. Técnica de Mínimos Cuadrados Parciales

La Regresión de Mínimos Cuadrados Parciales (PLSR, por sus siglas en inglés), también denominado proyección sobre estructuras latentes, es una

técnica que combina y generaliza características del Análisis de Componentes Principales (ACP) y la regresión lineal múltiple. Su objetivo es predecir un conjunto de variables dependientes a partir de un conjunto de variables independientes o predictores. Esta predicción se logra mediante la extracción de un conjunto de factores ortogonales llamados *variables latentes* que tengan la mejor capacidad de predicción (Martínez and Barrios, 2016).

2.6.1. Regresión Mínimos Cuadrados Parciales (PLSR)

La regresión PLS se utiliza generalmente en dos situaciones: cuando se tiene un gran número de variables predictoras con problemas de multicolinealidad, o bien cuando el número de variables independientes puede ser incluso mayor al número de observaciones, o cuando existe multicolinealidad entre las variables predictoras (Carrascal, 2015).

En general, la regresión PLS consta de dos pasos fundamentales (Vega-Vilca and Guzmán, 2011):

- Transformar la matriz X, con la ayuda del vector de respuestas Y, en una matriz de componentes ortogonales, $T = (T_1, T_2, \cdots, T_k)$ llamadas componentes *PLS*.
- Realizar un nuevo Análisis de Regresión Múltiple utilizando la misma variable de respuesta Y, y como variables predictoras las componentes PLS.

Para llevar a cabo la regresión de Y con las variables $X_1, X_2, \cdots, X_k$, PLS trata de encontrar nuevos factores que juegan el mismo papel que las $X's$. Estos nuevos factores se llaman variables latentes o componentes. Análogamente como en el método de RCP, cada componente es una combinación lineal de $X_1, X_2, \cdots, X_k$, pero mientras RCP usa sólo la variación de X para construir los nuevos factores, PLS usa tanto la variación de X como de Y para construir los nuevos factores que se usarán como variables explicatorias del modelo (López Pineda, 2013).

Generalmente, se supone que existen q variables dependientes $Y_1, Y_2, \cdots, Y_q$ y k variables independientes $X_1, X_2, \cdots, X_k$. Se dispone de n observaciones y se desea ajustar un modelo de regresión. Los datos se resumen en forma matricial: $Y_{(n\times q)}$ y $X_{(n\times k)}$, respectivamente.

El modelo de regresión PLS se puede escribir en forma de matriz como:

$$Y = XB + F,$$

La idea básica que se presenta en (Rodríguez et al., 2009) y (Medina et al., 2016) es hallar una descomposición en factores latentes T tales que:

$$\begin{aligned} X &= TP' + E, \\ Y &= UC' + F. \end{aligned} \tag{2.21}$$

Donde T y U son matrices de tamaño $n \times c$, que contiene las componentes latentes (*scores*) de las n observaciones. Por su parte, P de tamaño $k \times c$ es la matriz de *loadings* de X; C de $q \times c$ es la matriz de *loadings* de Y. Por su parte E y F, de dimensiones $n \times k$ y $n \times q$ respectivamente, son matrices de errores aleatorios (Rodríguez et al., 2009).

La regresión PLS es un método que sugiere construir componentes ortogonales (vectores latentes o scores) en X e Y de la forma $t = Xw$ y $u = Yc$, respectivamente, donde w y c son los vectores de pesos (loadings) adecuados de norma 1. Cada vector columna t genera la matriz T, de manera análoga cada vector columna u genera la matriz U, permitiendo la descomposición de X y Y, en la forma (2.21) como se muestra en (Medina et al., 2016).

La gran virtud de la técnica PLS es considerar el problema explicativo entre X e Y, a partir de las nuevas variables latentes representativas $t = Xw$ y $u = Yc$. La técnica considera como un modelo de regresión adecuado, aquel que además de reducir la dimención garantice la relación explicativa entre estas nuevas variables.

Encontrando los w y c adecuados, los cuales formarán las matrices W y C, y generando las matrices T y U, obtenemos la matriz de coeficientes B, donde

$$B = W(P'W)^{-1}C'$$

La solución al modelo PLS dada por el espacio de variables latentes también puede ser representado como un modelo de regresión consistente en los términos de las variables originales

$$Y = XB_{PLS} + F = XW(P'W)^{-1}C' + F.$$

2.6.2. Algoritmos PLS

Entre la literatura como (Medina et al., 2016) y (Mevik et al., 2016), se mencionan varios algoritmos PLSR, y el *paquete pls de R* implementa actualmente tres de ellos, a saber: El algoritmo Kernel para matrices (muchas observaciones, pocas variables), el algoritmo clásico de puntuación ortogonal (algoritmo NIPALS) y el algoritmo SIMPLS. Los algoritmos Kernel y NIPALS producen los mismos resultados (el algoritmo del kernel es el más rápido de ellos para la mayoría problemas). Mientras que el algoritmo SIMPLS produce el mismo ajuste para modelos de respuesta única, pero ligeramente diferente en

resultados para los modelos de respuesta múltiple.

En este caso, la matriz de datos puede ser escrita como $X = [X_1, \cdots, X_k]$, donde $X_1, \cdots, X_k$ son las columnas de la matriz X y el vector respuesta $Y_{(n\times 1)}$, los cuales han sido estandarizadas por columnas. Aquí, PLS puede considerarse como una transformación de las variables independientes, teniendo en cuenta su relación con la dependiente. Esta es precisamente la gran diferencia con el RCP en el que la transformación se aplica sólo a la matriz X (Rodríguez et al., 2009), (Rodríguez et al., 2012) y (Vega-Vilca and Guzmán, 2011).

1. Entrada de los datos: $X(0)_{n\times k}$, $Y(0)_{n\times 1}$
2. **Inicio**: para $i = 1, \ldots, k$.
3. $W = Cov(Y, X)$, $\mathcal{W} = \frac{W}{\|W\|}$
4. $T_i = X\mathcal{W}$
5. $v = (T_i'T_i)^{-1}(T_i'Y)$
6. $b = (T_i'T_i)^{-1}(T_i'X)$
7. $X(i) = X(i-1) - T_ib = X - \hat{X}$
8. $Y(i) = Y(i-1) - T_iv = Y - \hat{Y}$
9. **Fin** i

Tabla 2.1: Algoritmo simplificado para PLS (Vega-Vilca and Guzmán, 2011).

En los pasos 3 y 4 que se muestran en la Tabla 2.1 está uno de los puntos esenciales del método propuesto: la variable latente se conforma a partir de la covarianza entre las variables independientes y, en este caso, la dependiente. En los pasos 5 y 6 es fácil ver que v no es más que el coeficiente de regresión simple de Y sobre T, y b es un vector de dimensión k cuyas componentes no son más que los coeficientes de regresión simple de cada variable independiente, X_i, sobre T. Los pasos 7 y 8 del algoritmo son los de actualización de los valores.

El algoritmo que se presenta es iterativo y en cada proceso se calcula una componente latente. La idea fundamental es maximizar el cuadrado de la covarianza entre la componente $T = X\mathcal{W}$, y la variable respuesta Y, es decir:

$$\text{máx}[cov(Xw, Yc)]^2, \qquad \text{Sujeto a } \|w\| = \|c\| = 1.$$

Esto lleva a la aplicación de los multiplicadores de Lagrange y la solución no es más que el vector de covarianzas normalizado.

La reducción de la dimensionalidad puede ser aplicada directamente sobre las componentes ya que estas son ortogonales. El número de componentes necesarios para el análisis de regresión debe ser mucho menor que el número de predictoras.

2.6.3. Regresión Mínimos Cuadrados Parciales con Kernel (RKPLS)

Siguiendo las líneas de la técnica PLS, podemos asumir diferentes maneras de modelar relaciones de datos no lineales a partir de la técnica PLSR (Rosipal

and Trejo, 2001). Una estrategia consiste en usar métodos Kernel en espacios de Hilbert Reproductor del Kernel (RKHS) (Rodríguez et al., 2009). La idea básica consiste en asumir una transformación no lineal de las variables de entrada $\{X_i\}_{i=1}^n$ en un espacio de características $\mathcal{F}$, esto es, considerar una aplicación

$$\Phi : X_i \in \mathbb{R}^n \longrightarrow \Phi(X_i) \in \mathcal{F}$$

y luego construir un modelo RPLS en $\mathcal{F}$ (Medina et al., 2016).

La técnica RKPLS se basa en mapeos del espacio original de datos X a un espacio de alta dimensión $\mathcal{F}$, luego con la aplicación del truco Kernel la estimación de PLS en el nuevo espacio se reduce a cálculos de algebra lineal tan simples como en PLS lineal (Rosipal and Trejo, 2001). Como es sabido el truco Kernel permite conocer el producto punto entre dos elementos $\Phi(x)$, $\Phi(x')$ en $\mathcal{F}$ y construir la matriz Kernel K sin necesidad de conocer quién es en realidad Φ, es decir

$$K(x, x') = \langle \phi(x), \phi(x') \rangle, \qquad \forall x, x' \in X.$$

Se define así la matriz K de productos puntos entre todos los mapeos de los puntos de datos $K = \Phi'\Phi$, donde Φ denota la matriz de mapeo de los elementos o datos del espacio $\{\Phi(X_i)\}_{i=1}^n \in \mathcal{F}$.

Ahora, de acuerdo con (Medina et al., 2016) y (Rosipal and Trejo, 2001) del algoritmo PLS presentado en la Tabla 2.1, se puede derivar un nuevo algoritmo para el modelo no lineal, simplemente utilizando la matriz transformada $\boldsymbol{K}$ como datos de entrada.

Todas las técnicas Kernel multivariantes no lineales: Análisis de Componentes Principales con Kernel (ACPK), Análisis de Correlación Canónica con Kernel (KCCA), Mínimos Cuadrados Parciales con Kernel (KPLS), etc, se basan en aplicar los esquemas lineales a la matriz de entrada transformada ó matriz Gramiana $\boldsymbol{K}$ y a las respuestas. La meta de estas técnicas multivariantes no lineales se fundamenta en encontrar las direcciones de nuevas bases sobre las que se proyectan las variables originales (Rodríguez et al., 2009).

Capítulo 3

Aplicación a un estudio sobre secuestro de carbono en la Caldera de Teziutlán, Puebla

En (López Pineda, 2013) se desarrolla el estudio del Carbono Orgánico en el Suelo, en la zona de la Caldera de Teziutlán, Puebla., mediante diferentes técnicas de Regresión multivariada se realizó un estudio básico encontrando graves problemas de multicolinealidad, por lo cual se decidió utilizar las técnicas de Regresión con Componentes Principales (RCP) y Regresión de Mínimos Cuadrados Parciales (PLSR).

En este Capítulo, el objetivo es determinar un modelo de Regresión que muestre de manera clara la relación que existe entre el Carbono Orgánico del Suelo (COS) con diferentes propiedades del mismo. Para ello se trabaja en diferentes etapas, llegando a la aplicación de las técnicas: Regresión de Componentes Principales con Kernels (RCPK) y Regresión Mínimos Cuadrados Parciales con Kernels.

3.1. Planteamiento del problema

(Castillo-Morales et al., 2009) define el secuestro de carbono por el suelo como el proceso de transformación del carbono del aire (dióxido de carbono o CO_2) en carbono almacenado en suelo. También recibe el nombre de sumidero (transferencia neta de CO_2 atmosférico a la vegetación y al suelo para su almacenamiento). El dióxido de carbono es absorbido por las plantas a través del proceso de fotosíntesis y es incorporado al tejido vegetal. Cuando las plantas mueren, el carbono de las hojas, tallos y raíces se descompone en el suelo y se convierte en materia orgánica.

A través del secuestro de carbono, los niveles de CO_2 atmosférico son reducidos con el incremento de los niveles de carbono del suelo. Así el carbono, si no es perturbado, puede permanecer por muchos años como materia orgánica estable. Este proceso, como mencionamos antes, puede atenuar el cambio climático global o el calentamiento de la atmósfera. La relevancia que tiene en la actualidad la modelación del Carbono Orgánico del Suelo (COS), se encuentra íntimamente ligado a la propiedad de incidir directamente en el almacenamiento del carbono por medio del suelo y, por ende, en la mitigación de las emisiones de CO_2, que es uno de los Gases de Efecto Invernadero (GEI) de mayor importancia (Castillo-Morales et al., 2009).

A pesar de la importancia del secuestro de carbono para la mitigación del cambio climático, su evaluación se encuentra muy limitada en muchas zonas del mundo, en particular, en los suelos de la Sierra Norte de Puebla, México. Aún son escasas las investigaciones en el país que tratan de explicar el almacenamiento de carbono en suelos a través de propiedades de los mismos (Linares et al., 2014). Las muestras de suelo en las que se determinaron los diferentes contenidos extraíbles corresponden a diferentes perfiles de suelo, que fueron caracterizados previamente en el *Departamento de Investigación en Ciencias Agrícolas* (DICA) del *Instituto de Ciencias* de la Benemérita Universidad Autónoma de Puebla (*BUAP*).

El estudio se llevó a cabo en los suelos de la Región de Teziutlán, situada en la porción nororiental del estado de Puebla y que comprende una porción de la Región Terrestre Prioritaria para la Conservación en México (*RPT-105*), la cual se encuentra entre los paralelos $19°43'30''$ y $20°14'54''$ de latitud norte y los meridianos $97°07'42''$ y $97°43'30''$ de longitud occidental. Los suelos son derivados de material piroclástico, los cuales se presentan cubriendo una superficie aproximada de $846 Km^2$.

En la actualidad existe un gran interés científico por establecer la relación que guarda el COS con otras propiedades del mismo. En (López Pineda, 2013) se hizo un análisis estadístico en el cual se utilizó la base de datos que se encuentra en el **Apéndice A**, la cual está conformada por 42 observaciones correspondientes a 22 variables predictoras y una variable respuesta.

La variable respuesta es $Y = COS$ y las 22 variables predictoras corresponden a diferentes características que componen el suelo: la *Densidad Aparente* $(DA), \%Arena, \%Limo, \%Arcilla$, *Contenidos de Nitrógeno Total* (%N), *Relación C/N*, contenidos en *Al y Fe extraibles* (%Al y %Fe), *Retención de fosfatos*, *pH en NaF 1N*, *pH en* H_2O, *pH en* KCl, Delta pH, *Capacidad de Intercambio Catiónico Total* (CIC), *Porcentaje de saturación en bases* (%V), *Calcio* (Ca), *Magnesio* (Mg), *Sodio* (Na) y *Potasio* (K) Intercambiables (Castillo-Morales et al., 2009).

3.2. Modelo de Regresión Lineal Múltiple

La Tabla 3.1 muestra los coeficientes del modelo de regresión obtenidos con el método de MV, los errores estándar para cada coeficiente, así como el estadístico *t de Student* con sus correspondientes `p_valores` , donde podemos observar que en este caso los coeficientes obtenidos no contribuyen en forma significativa al modelo, pues los `p_ valores` son muy altos.

También puede apreciarse que el modelo de regresión lineal obtenido no tiene un buen ajuste, dado que el coeficiente de determinación es $R^2 = 0.6369$, es decir; el 63.69 % de la variabilidad de los datos queda explicado, por otra parte el coeficiente de determinación ajustado sólo alcanza el 21.65 %; además, la prueba F no es significativa ya que el `p-value` empírico es mucho mayor que algún nivel de significancia $\alpha = 0.05$ aceptable.

```
 Coefficients:      Estimate   Std. Error  t value  Pr(>|t|)P

(Intercept)        29525.578  35869.754   0.823    0.4206
DensidadA          258.725    202.826     1.276    0.2175
Arena              -297.959   358.251     -0.832   0.4159
Limo               -298.419   357.540     -0.835   0.4143
Arcilla            -300.224   357.850     -0.839   0.4119
M.O.               249.785    433.939     0.576    0.5716
C.Org.             -413.366   746.559     -0.554   0.5862
N                  -17.682    24.206      -0.730   0.4740
C.N                -4.162     7.950       -0.524   0.6066
RetenciónFosfatos  1.512      1.983       0.762    0.4552
Al.Extraible       199.403    115.208     1.731    0.0997·
Fe.Extraible       167.360    92.932      1.801    0.0876·
Al.Fe.extr         -167.623   114.712     -1.461   0.1603
ph.NaF.N           -1.660     2.317       -0.716   0.4825
pH.H2O.S           38.963     51.766      0.753    0.4609
pH.KCl.S           -42.120    59.934      -0.703   0.4907
Delta.pH           68.535     72.808      0.941    0.3584
CIC                2.011      3.905       0.515    0.6125
V                  4.533      3.890       1.165    0.2583
Ca                 -28.954    25.528      -1.134   0.2708
Mg                 -5.464     11.245      -0.486   0.6326
Na                 -18.502    41.602      -0.445   0.6615
K                  -15.067    41.250      -0.365   0.7190

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 81.7 on 19 degrees of freedom
Multiple R-squared: 0.6369, Adjusted R-squared: 0.2165
F-statistic: 1.515 on 22 and 19 DF, p-value: 0.1821
```

Tabla 3.1: Modelo de regresión: Método MV.

Cabe señalar que en (López Pineda, 2013) no se realizó un estudio completo en cuanto al análisis sobre los supuestos del modelo, por lo cual se describirán en forma detallada los resultados que se obtienen.

En la Fig 3.1 se muestran los diferentes gráficos de diagnóstico para el modelo ajustado, en el gráfico de la parte superior izquierda, se puede observar que en este caso parece no haber homogeneidad de varianza. En cuanto a la normalidad, el gráfico de la parte superior derecha muestra que de igual forma no se tiene normalidad en este modelo.

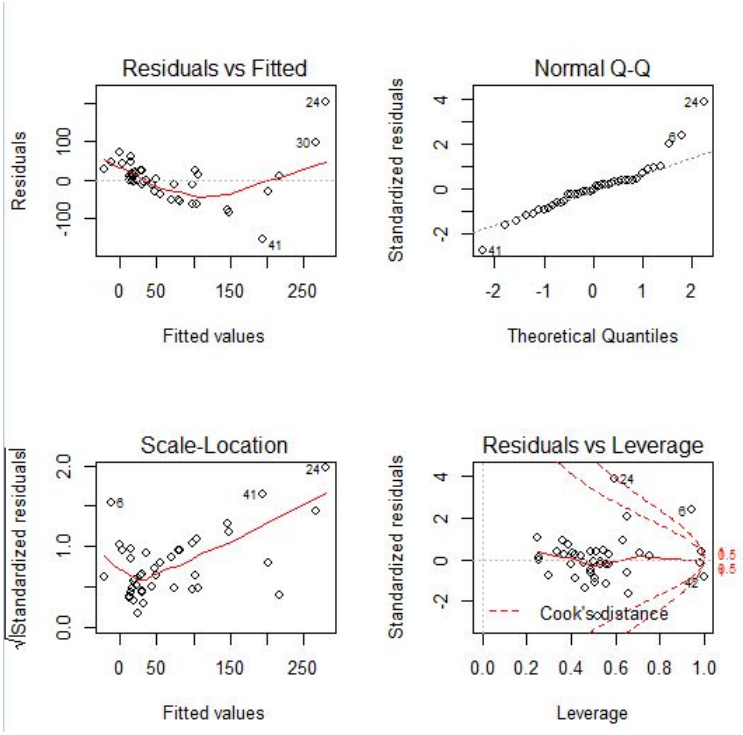

Figura 3.1: Gráficos de diagnóstico.

Por otra parte, con el software R se realizaron las correspondientes pruebas de *Breusch-Pagan* y *Shapiro-Wilks* para homogeneidad de varianza y normalidad respectivamente, obteniendo como resultado que en ambos casos no se cumplen estos supuestos. En cuanto al supuesto de independencia en los errores se utilizó la prueba de *Durbin-Watson* obteniendo que en este caso si se tiene independencia en los errores.

```
> library(lmtest)
> dwtest(ajuste, data = prop)
        Durbin-Watson test
data:  ajuste
DW = 1.9082, p-value = 0.0975
alternative hypothesis: true autocorrelation is greater than 0
```

En este caso tenemos:

$$p\_value = 0.0975 > 0.05 \qquad \Rightarrow \qquad \text{No rechazamos } H_0.$$

por lo tanto tenemos $\rho = 0$ es decir los residuos no están correlacionados.

3.3. Diagnóstico y eliminación de *Outliers*

En la Figura 3.1 se puede apreciar el gráfico qq-plot que se obtiene al hacer el ajuste de regresión considerando las 42 observaciones, en el se encuentran marcadas las observaciones 6, 24 y 41 como posibles outliers, en cambio en la Figura 3.3 de *Residuals Vs Leverage* se señalan las observaciones 6, 24 y 42.

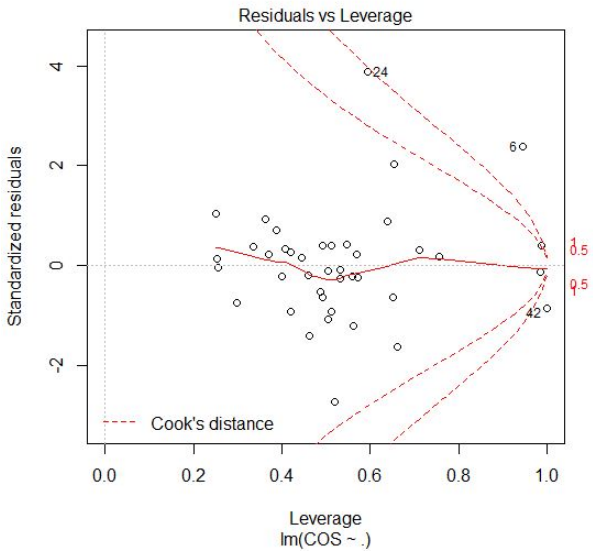

Figura 3.2: Gráfico de Residuals Vs Leverage.

En la Figura 3.3 se muestra el gráfico boxplot correspondiente a los datos ajustados, en el cual se observa la presencia de observaciones atípicas.

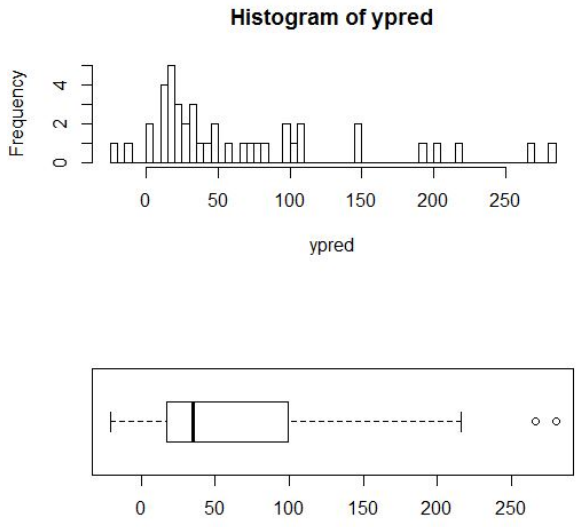

Figura 3.3: Gráfico boxplot para los valores ajustados.

En la Tabla A.4 se presentan las diferentes medidas de influencia, las cuales corresponden a las Distancias de Mahalanobis, Distancias de Cook's, residuos estandarizados y estudentizados y los valores de h_{ii}, se puede observar que aparecen marcadas con $*$ las observaciones que bajo los diferentes criterios se consideran outliers. Después de hacer algunas observaciones en los gráficos obtenidos y en las medidas de influencia que se tienen, se detecto la posible presencia de varios outlier. Para comprobar si alguno de estos puntos corresponde efectivamente a una observación atípica, haremos uso del Software R y el paquete *outliers* (Faraway, 2014).

Con esta línea obtendremos la observación que corresponde al outlier, la cual será localizada entre los valores ajustados (marcados como *predichos*), para su eliminación.

```
> outlier(ypred)
[1] 280.0194
```

Después de detectar y localizar el outlier, lo que se hizo fue eliminar el renglón correspondiente a esa observación y realizar nuevamente un Análisis de Regresión Múltiple. Una vez que se han hecho los ajustes correspondientes a la base de datos con y sin la observación 24 que se considero un outlier, lo que haremos será comparar ambos modelos y elegir el mejor en base a diferentes criterios de selección de modelos.

	R^2_{Adj}	AIC
Modelo (Original)	21.65 %	503.733
Modelo (Sin outlier)	66.11 %	429.567

Tabla 3.2: Selección del modelo.

En base a la Tabla 3.2 podemos observar que el Modelo (Sin outlier) tiene mejor capacidad de ajuste y además el estadístico AIC es menor al del Modelo (Original).

Podemos observar en la Tabla 3.3 que para el nuevo Modelo de regresión (Sin outlier), las pruebas para los coeficientes individuales muestran que para la variable *Na* (Sodio) el `p_valor=0.0133<0.05` por lo cual podemos decir que esta variable es de importancia en este nuevo ajuste.

Por otra parte el coeficiente de determinación $R^2 = 0.8475$ aumento con respecto al modelo original, mientras que el coeficiente de determinación ajustado alcanza el 66.11 % que también es más alto que el anterior. Por otra parte, la prueba F es significativa, ya que el `p_valor=0.0009531<0.05` y por tanto se rechaza la hipótesis nula, aceptándose que algún coeficiente es diferente de cero.

En la Fig 3.4 se muestran los gráficos de diagnóstico para el nuevo modelo ajustado, en el gráfico de la parte superior izquierda, se puede observar que nuevamente parece no haber homogeneidad de varianza por lo que se sugiere considerar un modelo de tipo no lineal. En cuanto a la normalidad, el gráfico de la parte superior derecha muestra que de igual forma no se tiene normalidad en este modelo.

De nueva cuenta se realizarón las correspondientes prubas de *Breusch-Pagan* y *Shapiro-Wilks* para homogeneidad de varianza y normalidad respectivamente, obteniendo como resultado en este caso si se tiene homogeneidad de varianza. Sin embargo la prueba de Shapiro-Wilks da como resultado que el supuesto de normalidad sigue sin cumplirse. En cuanto al supuesto de independencia en los errores se utilizó la prueba de *Durbin-Watson* obteniendo la independencia en los errores.

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)P
(Intercept)	22929.1525	16841.9101	1.361	0.1902
DensidadA	22.1056	99.3365	0.223	0.8264
Arena	-231.0559	168.2153	-1.374	0.1864
Limo	-231.4488	167.8826	-1.379	0.1849
Arcilla	-231.1602	168.0405	-1.376	0.1858
M.O.	178.3605	203.7018	0.876	0.3928
C.Org.	-299.9699	350.4063	-0.856	0.4032
N	3.2858	11.6326	0.282	0.7808
C.N	0.4182	3.7693	0.111	0.9129
RetenciónFosfatos	0.4277	0.9393	0.455	0.6543
Al.Extraible	-3.1572	59.32628	-0.053	0.9581
Fe.Extraible	-74.4238	52.4845	-1.418	0.1733
Al.Fe.extr	36.2322	59.1794	0.612	0.5480
ph.NaF.N	-0.2701	1.0996	-0.246	0.8088
pH.H2O.S	17.5354	24.4164	0.718	0.4819
pH.KCl.S	-6.0253	28.4461	-0.212	0.8346
Delta.pH	-20.3296	35.7984	-0.568	0.5771
CIC	0.9167	1.8364	0.499	0.6237
V	0.7095	1.8821	0.377	0.7106
Ca	-10.7501	12.1736	-0.883	0.3888
Mg	-0.2558	5.3112	-0.048	0.9621
Na	59.5182	21.6730	2.746	0.0133 *
K	-22.3574	19.3666	-1.154	0.2634

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38.32 on 18 degrees of freedom
Multiple R-squared: 0.8475, Adjusted R-squared: 0.6611
F-statistic: 4.547 on 22 and 18 DF, p-value: 0.0009531
```

Tabla 3.3: Modelo de regresión: Método de MV. (Sin obs. 24).

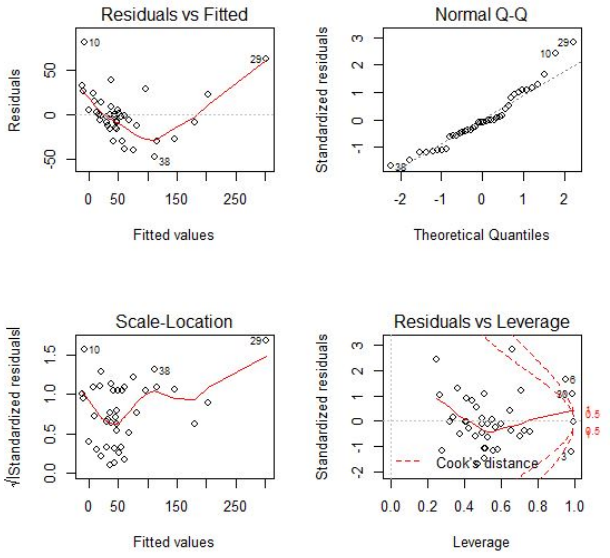

Figura 3.4: Gráficos de diagnóstico.

3.4. Diagnóstico y tratamiento de Multicolinealidad

Ahora que se ha elegido uno de los dos modelos lo que se requiere es indagar en la existencia de multicolinealidad, como se puede observar en la Tabla 3.4, se detectaron problemas de multicolinealidad: pues los valores VIF, en **12** de las **22** variables predictoras, están en el rango de **10.75** a **188048.5**, de igual forma el alto valor del número condición, $\boldsymbol{\kappa = 7056.70}$ confirma la existencia de multicolinealidad en un grado bastante alto.

DensidadA	Arena	Limo	Arcilla
2.720676	**188048.5**	**78042.19**	**67232.84**
M.O.	C.Org.	N	C.N
22757.68	**22738.51**	1.574550	3.724395
RetenciónFosfatos	Al.Extraible	Fe.Extraible	Al.Fe.extr
3.826883	**111.5773**	**13.01004**	**144.8779**
pH.NaF.N	pH.H2O.S	pH.KCl.S	Delta.pH
10.88465	7.401780	7.817162	4.808106
CIC	V	Ca	Mg
13.24558	**23.20940**	**13.41509**	1.88083
Na	K		
4.886702	4.896007		

Tabla 3.4: Valores VIF para detectar multicolinealidad.

Es importante mencionar que previanente en (López Pineda, 2013) se trabajó de manera similar considerando como solución al problema de multicolinealidad las técnicas de Regresión de Componentes Prinipales (RCP) y Regresión de Mínimos Cuadrádos Parciales (PLSR), en el caso multivariado.

3.5. Métodos Kernel

Después del diagnóstico realizado en cuanto a los supuestos básicos del modelo de regresión lineal múltiple, se puede observar que el Modelo (Sin Outlier), cumple con los supuestos de homogeneidad de varianzas e independencia en los errores; por otra parte la suposición de normalidad no se cumple y de igual forma la suposición de independencia en las variables regresoras no se cumple debido a la existencia de *multicolinealidad*.

Como se explica en el Capítulo 2, existen diferentes enfoques estadísticos, en este caso el hecho de no tener normalidad en los residuos y tener problemas de multicolinealidad, da entrada para considerar nuevas técnicas para poder corregir el problema de multicolinealidad.

Dentro de las técnicas que se tienen para tratar la multicolinealidad con relaciones no lineales, se encuentran la regresión de Componentes Principales mediante el uso de Kernels (RCPK) y la regresión de Mínimos Cuadrados Parciales con Kernels (RKPLS).

3.5.1. Componentes Principales con Kernels

Como se indica en la sección 2.5.2, el procedimiento para utilizar la regresión de componentes principales mediante kernels, consiste en seguir una serie de pasos para alcanzar los resultados esperados. En primer lugar, una vez elegida la función Kernel que representa el producto punto de los datos transformados, lo siguiente es hallar la matriz Kernel asociada la cual servirá como base para realizar el análisis de componentes principales mediante kernels y posteriormente aplicar el análisis de regresión a estas nuevas componentes.

Para esta técnica se utiliza el paquete Kernlab de R, ya que dicho paquete brinda las funciones y rutinas necesarias para desarrollar el análisis. En (Karatzoglou et al., 2016) y (Zeileis et al., 2004) se describen de manera más detallada las funciones que implementa este paquete.

El primer Kernel que se presenta es el Kernel *Gaussiano*, el cual está dado por:

$$K(x, y) = \exp\left(\frac{-\|x - y\|^2}{2\sigma^2}\right),$$

donde σ es el parámetro que determina la amplitud de la función Gaussiana. Para el problema tratado se utilizan los siguientes valores para $\sigma = 0.001$, $\sigma = 0.01$ y $\sigma = 0.05$. Después de fijar los valores para σ se calcularon las correspondientes matrices Kernel, para posteriormente realizar el análisis de componentes principales.

Cabe resaltar que solo se trabajó con las primeras 10 componentes principales ya que estas cubren la mayor parte de variabilidad de los datos. Los resultados obtenidos después de realizar nuevos modelos de regresión se encuentran resumidos en la Tabla 3.5, en la cual se comparan los valores R^2 y R^2_{Adj} para los diferentes valores de σ.

Como se puede observar en la Tabla 3.5 el Kernel que brinda valores más altos para R^2 y R^2_{Adj} es el que tiene el valor $\sigma = 0.001$.

En la Tablas A.5, A.6 y A.7 del Apéndice A, se muestran los valores de R^2, R^2_{Adj}, AIC y $PRESS$, para los modelos de regresión con Componentes Principales mediante el uso del Kernel Gaussiano con $\sigma = 0.001$, $\sigma = 0.01$ y $\sigma = 0.05$. Comparando todos estos resultados se decide elegir el Kernel Gaussiano con $\sigma = 0.001$.

	$\sigma = 0.001$		$\sigma = 0.01$		$\sigma = 0.05$	
N. Comp	R^2	R^2_{Adj}	R^2	R^2_{Adj}	R^2	R^2_{Adj}
(10)	0.3499	0.1332	0.2356	−0.0191	0.1869	−0.0814
(9)	0.3201	0.1227	0.2131	−0.0154	0.1849	−0.0517
(8)	0.3200	0.1500	0.2048	0.0060	0.1826	−0.0217
(7)	0.3191	0.1746	0.2021	0.0325	0.1825	0.0090
(6)	0.3024	0.1793	0.0869	−0.0741	0.0899	−0.0707
(5)	0.2090	0.0959	0.0534	−0.0818	0.0667	−0.0666
(4)	0.2032	0.1146	0.0533	−0.0517	0.0588	−0.0456
(3)	0.1882	0.1224	0.0437	−0.0374	0.0459	−0.0314
(2)	0.1879	0.1452	0.0137	−0.0381	0.0191	−0.0325
(1)	0.1606	0.1391	0.0007	−0.0248	0.0007	−0.0249

Tabla 3.5: Comparación de R^2 y R^2_{Adj} para diferentes valores de σ.

El segundo Kernel utilizado fue el de tipo *Polinomial*, el cual está dado por:

$$K(x, y) = (\langle x, y \rangle + 1)^{\alpha},$$

donde α es el parámetro que determina el grado del polinomio considerado, en este caso se utilizaron valores $\alpha = 1, 2,$ y 3. Después de fijar estos valores nuevamente se calcularon las correspondientes matrices Kernel y se procedió de manera análoga al Kernel Gaussiano.

En la Tabla 3.6, se comparan los valores R^2 y R^2_{Adj} para los diferentes valores de α.

	$\alpha = 1$		$\alpha = 2$		$\alpha = 3$	
N. Comp	R^2	R^2_{Adj}	R^2	R^2_{Adj}	R^2	R^2_{Adj}
(10)	0.4744	0.2993	0.5363	0.3818	0.6057	0.4773
(9)	0.4522	0.2932	0.4851	0.3356	0.5207	0.3816
(8)	0.4422	0.3027	0.4830	0.3537	0.5061	0.3827
(7)	0.4051	0.2790	0.4469	0.3295	0.4924	0.3847
(6)	0.3659	0.2540	0.4160	0.3129	0.4664	0.3722
(5)	0.3579	0.2662	0.4098	0.3254	0.4594	0.3822
(4)	0.3144	0.2382	0.3647	0.2941	0.4100	0.3444
(3)	0.3142	0.2586	0.3642	0.3127	0.4086	0.3607
(2)	0.2518	0.2124	0.3135	0.2773	0.3665	0.3332
(1)	0.2282	0.2084	0.2788	0.2603	0.3198	0.3024

Tabla 3.6: Comparación de R^2 y R^2_{Adj} para diferentes valores de α.

Como se puede observar el Kernel Polinomial que brinda valores más altos para R^2 y R^2_{Adj} es el que tiene grado $\alpha = 3$.

En las Tablas A.8, A.9 y A.10 del Apéndice A, se pueden observar los valores de R^2, R^2_{Adj}, AIC y $PRESS$, para los modelos de regresión con Componentes Principales mediante el uso del Kernel Polinomial con $\alpha = 1, 2$ y 3. Comparando estos resultados se decide elegir el Kernel Polinomial con grado $\alpha = 3$, ya que esté es el mejor modelo según los diferentes criterios.

El último Kernel que se utilizó fue el *Laplaciano o Exponencial*, el cual se encuentra dado por:

$$K(x,y) = \exp\left(\frac{-\|x-y\|^2}{\sigma}\right),$$

donde nuevamente σ es el parámetro que determina la amplitud de la función. En este caso, de nueva cuenta se utilizaron los valores $\sigma = 0.001$, $\sigma = 0.01$ y $\sigma = 0.05$. Después de fijar los valores para σ se procedió de manera análoga a los Kernels anteriores, se calculo la matriz Kernel y se realizó el análisis de Componentes Principales.

Los resultados que se obtuvieron al realizar los nuevos modelos de regresión se encuentran resumidos en la Tabla 3.7, en la cual se comparan los valores R^2 y R^2_{Adj} para los diferentes valores de σ.

	$\sigma = 0.001$		$\sigma = 0.01$		$\sigma = 0.05$	
N. Comp	R^2	R^2_{Adj}	R^2	R^2_{Adj}	R^2	R^2_{Adj}
(10)	0.4206	0.2275	0.3107	0.0800	0.3579	0.1439
(9)	0.4080	0.2362	0.3075	0.1065	0.3407	0.1493
(8)	0.4080	0.2600	0.3062	0.1327	0.3397	0.1746
(7)	0.3931	0.2644	0.3016	0.1535	0.3380	0.1976
(6)	0.3512	0.2368	0.2004	0.0592	0.3345	0.2171
(5)	0.3307	0.2351	0.1484	0.0267	0.2015	0.0873
(4)	0.2840	0.2045	0.1379	0.0421	0.2009	0.1121
(3)	0.2814	0.2232	0.1366	0.0666	0.1921	0.1266
(2)	0.2258	0.1851	0.1331	0.0874	0.1918	0.1493
(1)	0.2227	0.2028	0.1313	0.1090	0.1707	0.1494

Tabla 3.7: Comparación de R^2 y R^2_{Adj} para diferentes valores de σ.

Como podemos observar el valor $\sigma = 0.001$ es el que brinda valores más altos para R^2 y R^2_{Adj}. Por otra parte en las Tablas A.11, A.12 y A.13 del Apéncie A, se pueden observar valores de R^2, R^2_{Adj}, AIC y $PRESS$, para los modelos de regresión con Componentes Principales mediante el uso del Kernel Laplaciano con $\sigma = 0.001$, $\sigma = 0.01$ y $\sigma = 0.05$. Comparando todos estos resultados se decide elegir el Kernel Laplaciano con $\sigma = 0.001$.

De acuerdo con las observaciones, hasta el momento se puede observar que el Kernel que brindo mejores resultados fue el Kernel Polinomial con grado $\alpha = 3$. Este modelo incluye tres componentes principales, las cuales se obtuvieron mediante la técnica señalada anteriormente.

```
Coefficients:  Coef        SE Coef      T     P
Constant       5.732e+01   8.219e+00   6.974   3.07e-08 ***
Cp1            1.506e+08   3.367e+07   4.473   7.08e-05 ***
Cp2            3.549e+07   2.077e+07   1.709   0.0958 .
Cp3            2.3073+07   1.421e+07   1.623   0.1130

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 52.63 on 37 degrees of freedom
Multiple R-squared: 0.4086, Adjusted R-squared: 0.3607
F-statistic: 8.522 on 3 and 37 DF, p-value: 0.0001973
```

Tabla 3.8: Coeficientes de RCPK, con Kernel Polinomial de grado 3.

La Tabla 3.8 muestra los coeficientes del modelo de regresión obtenidos con el método de MV, los errores estándar para cada coeficiente, así como el estadístico *t de Student* con sus correspondientes `p_valores`, donde se puede observar que los coeficientes de este modelo contribuyen en forma significativa al modelo, ya que los `p_ valores` son relativamente bajos.

También se puede apreciar que este modelo tiene un coeficiente de determinación $R^2 = 0.6369$; es decir que aproximadamente el 63 % de los datos queda explicado, por otro lado el coeficiente de determinación ajustado es $R^2_{Adj} = 0.3607$; además, la prueba F es significativa ya que el `p-value` empírico es $p-value = 0.0001973 < 0.05 = \alpha$, con esto se rechaza la hipótesis nula dando por hecho que algún coeficiente es diferente de cero.

3.5.2. Mínimos Cuadrados Parciales con Kernels

Como se ha mencionado anteriormente el *paquete PLS de R* implementa los algoritmos: Kernel PLS, NIPALS y SIMPLS (Mevik et al., 2007). Para el problema tratado se utilizó el algoritmo Kernel PLS para obtener las componentes latentes del modelo. A modo de comparación con la metodología RCPK se decidió determinar 10 componentes PLS y posteriormente comparar los valores R^2, R^2_{Adj}, AIC y $PRESS$ para cada modelo ajustado.

Cabe mencionar que en un estudio adicional se realizó la técnica RPLS mediante el algoritmo SIMPLS, obteniendo los mismos resultados que brinda el algoritmo Kernel PLS.

Como podemos ver en la Figura 3.5 la primera componente aporta un 24.4 % de la varianza explicada y a su vez es la que aporta mayor información al modelo, de modo que la varianza explicada con cinco componentes es del 61.23 %.

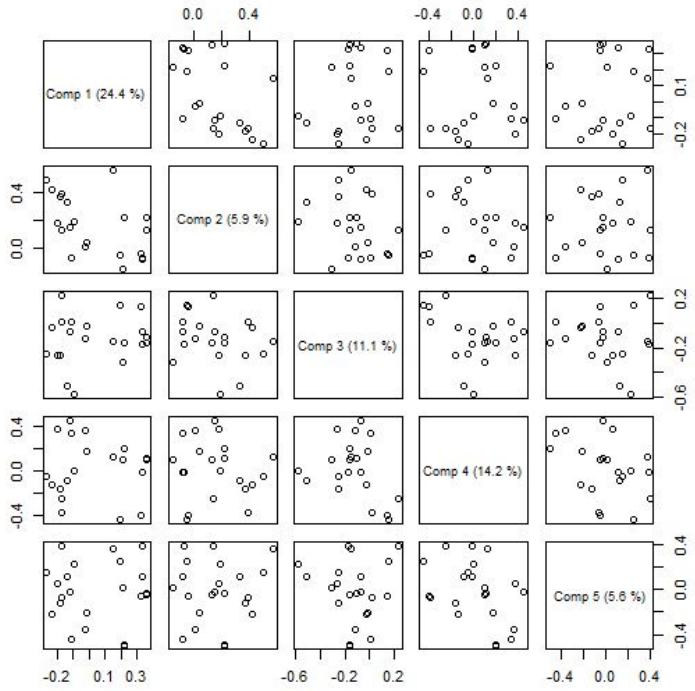

Figura 3.5: Gráficas de dispersión para las componentes seleccionadas.

En la Tabla 3.9 se muestran los valores obtenidos para diferente número de componentes, como se puede apreciar los valores más pequeños de AIC y PRESS corresponden al modelo ajustado con cinco componentes. De igual forma este modelo nos brinda el valor R^2_{Adj} más alto, de modo que este es el modelo seleccionado.

$\sigma = 0.001$

N. Comp	R^2	R^2_{Adj}	AIC	$PRESS$
(10)	0.8192	0.7589	412.55	63375
(9)	0.8186	0.7660	410.67	60240
(8)	0.8173	0.7717	408.96	58467
(7)	0.8163	0.7774	407.19	53946
(6)	0.8145	0.7817	405.60	52593
(5)	0.8097	0.7826	403.76	52340
(4)	0.8021	0.7801	404.24	52475
(3)	0.7946	0.7780	404.76	52547
(2)	0.7448	0.7314	410.67	64125
(1)	0.4963	0.4833	436.55	115124

Tabla 3.9: Comparación de R^2_{Adj}, $PRESS$ y AIC.

En la Tabla 3.10 se puede observar que el modelo ajustado con 5 componentes KPLS es adecuado, pues el `p_valor` para la prueba F es $1.11e-11$, menor que el nivel de significancia $\alpha = 0.05$. En este caso el $R^2 = 0.8097$ lo que significa que el modelo explica aproximadamente el 80 % de la variabilidad de los datos, para $R^2_{Adj} = 0.7826$ indica que el modelo se ajusta un 78 %. En cuanto a los coeficientes, tres de los `p_valores` son menores que $\alpha = 0.05$, es decir tres de los cinco coeficientes son significativos.

```
Coefficients:  Estimate  Std. Error  t value  Pr(>|t|)
Constant       57.321    4.749       11.958   6.47e-14 ***
Comp1          20.907    2.188       9.554    2.76e-11 ***
Comp2          34.922    5.165       6.762    7.78e-08 ***
Comp3          10.140    3.349       3.028    0.0046 **
Comp4          3.747     3.191       1.174    0.2483
Comp5          6.417     5.426       1.183    0.2449

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30.69 on 35 degrees of freedom
Multiple R-squared: 0.8097, Adjusted R-squared: 0.7826
F-statistic: 29.79 on 5 and 35 DF, p-value: 1.11e-11
```

Tabla 3.10: Coeficientes de RKPLS, con 5 componentes.

Se debe resaltar que con esta nueva regresión se elimina el problema de multicolinealidad, pues los valores VIF son igual a 1.

En la Tabla 3.11 se muestran los *loadings (cargas)* del modelo que se obtuvo con la técnica RKPLS. Como puede observarse, para cada componente PLS se presentan los valores más altos que aporta cada variable predictora al modelo determinado. Estas cargas oscilan en el rango $-1 \setminus +1$, denotando el signo y la intensidad del efecto.

Además de este modo podemos asignar a las componentes una proporción de la información que contienen de cada variable, teniendo en cuenta cómo *pesan* en ellas las distintas variables predictoras.

Loadings:

	Comp 1	Comp 2	Comp 3	Comp 4	Comp5
DensidadA	-0.171	0.134	0.231	-0.247	0.390
Arena	0.191		0.151	-0.440	0.251
Limo	-0.110			0.335	-0.445
Arcilla	-0.200	0.177	-0.263	0.374	
M.O.	0.221	0.218	-0.163	0.197	-0.500
C.Org.	0.220	0.220	-0.163	0.196	-0.497
N				0.171	-0.206
C.N	-0.118	0.149		0.453	
RetenciónFosfatos	0.328		-0.172		0.383
Al.Extraible	0.359	0.222	-0.112	0.112	
Fe.Extraible	0.211	-0.158	-0.312	0.106	
Al.Fe.extr	0.348	0.132	-0.162	0.102	
ph.NaF.N	0.318		0.138	-0.391	
pH.H2O.S	-0.171	0.393		-0.378	
pH.KCl.S	-0.237	0.422		-0.129	-0.225
Delta.pH			-0.123	0.368	-0.360
CIC	0.334				0.114
V	-0.264	0.499	-0.253		0.148
Ca		0.191	-0.581		0.222
Mg	-0.182	0.375	-0.260	-0.162	-0.122
Na	0.148	0.571	-0.152	0.128	0.367
K	-0.135	0.330	-0.513		0.118

Tabla 3.11: Loadings del modelo con 5 componentes.

La interpretación de las cargas en este caso es sencilla. Por ejemplo la componente uno (sus cargas están en Comp 1 de la Tabla 3.11) se asocia de modo positivo e intenso con las variables M.O(Materia orgánica), C.Org(Carbono orgánico), RetenciónFosfatos (Retención de Fosfatos), Al.Extraible (aluminio extraíble), Al.Fe.extr (aluminio más fierro extraíble) y CIC(Capacidad de Intercambio Catiónico).

La segunda componente (Comp 2 de la Tabla 3.11) se asocia de manera positiva con las variables pH.H2O (pH en agua), pH.KCl (pH en cloruro de potasio), V (porcentaje de saturación en bases)y Na(Sodio). Estas observaciones se pueden hacer para cada una de las componentes, es decir; podemos describir de manera clara la información que aporta cada variable predictora al modelo determinado por las nuevas variables (componentes).

3.5.3. Predicción para COS

Con el fin de evaluar y comparar la predicción para los modelos de regresión estimados, se seleccionaron aleatoriamente 5 observaciones: 2, 8, 11, 21, 40 de la base de datos utilizados para la aplicación. En las Tablas 3.12 y 3.13 se pueden apreciar las predicciones para las observaciones determinadas de $Y = COS$ y sus respectivos intervalos de predicción.

Obs	Valor real	Predicción	Intervalo 95 %
2	17.47	−9.1058	(−104.8771, 86.6654)
8	38.64	47.0549	(−49.0614, 143.1713)
11	22.19	60.6187	(−30.3618, 151.5994)
21	70.00	82.3819	(−22.6794, 187.4432)
40	40.33	48.0970	(−57.8283, 154.0225)

Tabla 3.12: Predicciones para el Modelo (Sin outlier).

Obs	Valor real	Predicción	Intervalo 95 %
2	17.47	−3.7506	(−69.4821, 61.9807)
8	38.64	32.7530	(−31.9447.97.4508)
11	22.19	72.9613	(7.4926, 138.4300)
21	70.00	82.2182	(12.4538, 151.9825)
40	40.33	65.8152	(−52.8304, 83.2603)

Tabla 3.13: Predicciones para el Modelo RKPLS con 5 componentes.

Para el problema de predicción del Carbono Orgánico en Suelos COS, la Tablas 3.12 y 3.13 presentan las predicciones para el Modelo (Sin Outlier) y las predicciones para el Modelo RKPLS respectivamente. Por ejemplo para la observación 2, con el Modelo (Sin Outlier) tenemos un intervalo de predicción del 95 %, lo cual significa que si se espera tener un valor real de $COS = 17.47$, el valor predicho estaría entre $(-104.8771, 86.6654)$ aproximadamente, mientras que con la técnica RKPLS el valor predicho oscila entre $(-69.4821, 61.9807)$, lo cual genera una mayor precisión. Sin embargo ambos modelos son adecuados para efectos predictivos.

3.6. Resultados y discusiones

Esta sección tiene como objetivo evaluar y comparar el desempeño de las metodologías utilizadas para tratar la modelación del Carbono Orgánico en Suelos (COS) en función de las variables predictoras descritas anteriormente. Cabe mencionar que esta investigación se centró principalmente en dar solución al problema de la violación al supuesto de independencia en las variables predictoras del modelo y dadas las observaciones en los gráficos de diagnóstico se decidió hacer uso de técnicas para modelado no lineal.

En principio se consideró la Regresión de Componentes Principales con Kernels, en esta parte de la investigación se trabajo con diferentes Kernels y para cada Kernel considerado se tomaron valores distintos para poder comparar los resultados que arrojaba cada modelo. Los Kernels utilizados fueron: Kernel Gaussiano ($\sigma = 0.001$, $\sigma = 0.01$ y $\sigma = 0.005$), Kernel Polinomial (Grados $\alpha = 1$, $\alpha = 2$ y $\alpha = 3$) y por último el Kernel Laplaciano ($\sigma = 0.001$, $\sigma = 0.01$ y $\sigma = 0.005$).

De acuerdo con las observaciones anteriores, se sugiere que para la RCPK el Kernel que brinda mejores resultados es el de tipo Polinomial con grado $\alpha = 3$.

Por otra parte, en cuanto a la Regresión de Mínimos Cuadrados Parciales con Kernels (RKPLS), el software R no permite usar diferentes Kernels, de modo que los resultados obtenidos para esta técnica solo se compararon para diferente número de componentes latentes, dando como resultado que el modelo ajustado con cinco componentes KPLS es el que presenta mejores propiedades.

En la Tabla 3.14 se puede observar, que para los diferentes criterios de selección de modelos, en todos los casos se decide elegir el modelo obtenido con la técnica KPLS con cinco componentes, en cuanto a R^2 este explica aproximadamente el 80 % de la variación total del modelo, se tiene un $R^2_{Adj} = 78.26\,\%$, el AIC es menor para RKPLS, de modo que se tiene mejor calidad en este modelo y por último el valor $PRESS$ es menor en este modelo, lo cual indica que tiene mejor capacidad de predicción.

Criterio	RCPK(3 Comp)	RKPLS(5 Comp)
R^2	0.4086	0.8097
R^2_{Adj}	0.3607	0.7826
AIC	447.13	403.76
$PRESS$	150113	52340

Tabla 3.14: Comparación de R^2_{Adj}, $PRESS$ y AIC para RCPK y RKPLS.

Como se puede observar en la Tabla 3.13, las predicciones con la técnica RKPLS generan intervalos de predicción más pequeños, lo cual representa una ganancia en predicción.

Conclusiones

El objetivo de esta tesis fue desarrollar diferentes técnicas para modelación estadística, sin embargo, fue necesario hacer una revisión exhaustiva de diferentes bibliografías tanto del Análisis Funcional como de Inferencia Estadística, para poder desarrollar las técnicas que se deseaban utilizar.

En la investigación realizada para la modelación de COS, se presentaron diferentes problemas dentro del proceso de obtención del *mejor* modelo. El proceso señalado se llevó a cabo en diferentes etapas y el primer problema con el que se trabajó fue la detección y tratamiento de outliers, en esta etapa se decidió eliminar del conjunto de observaciones aquella que, mediante diferentes criterios y pruebas se considero un outlier.

Después de eliminar el outlier, se llevó a cabo un nuevo ajuste de regresión, el siguiente paso fue comparar ambos modelos y en base a los valores R^2, R^2_{Adj} y AIC, se decidió trabajar con el modelo sin el outlier, ya que al remover dicha observación se logró una mejor descripción de la variabilidad de los datos.

Para este nuevo modelo, se realizaron las correspondientes pruebas de bondad de ajuste y comprobación a las suposiciones básicas del modelo. Para estas últimas, los supuestos de normalidad e independencia en las variables predictoras no se cumplen. En este trabajo se decidió utilizar las técnicas de RCPK y RKPLS, las cuales tienen múltiples aplicaciones en problemas de modelación no lineal.

La técnica de RCPK se realizó con diferentes Kernels y con diferentes valores para los parámetros de cada Kernel. En esta etapa se utilizó el paquete Kernlab de R, el cual brinda las herramientas necesarias para poder hacer los diferentes cálculos requeridos para la comparación de los resultados.

Para la RCPK se decidió utilizar el Kernel de tipo Polinomial con grado 3 y se eligió el modelo que considera tres componentes principales ya que éste es el más adecuado, a comparación de los otros modelos obtenidos con esta técnica.

En cuanto a la RKPLS, se utilizó el paquete PLS de R que a diferencia del anterior, el algoritmo Kernel PLS está predeterminado en el paquete, es decir, no se pueden hacer cambios en cuanto al kernel utilizado.

Para la técnica RKPLS, se utilizó diferente número de componentes y se eligió el modelo que considera cinco componentes KPLS, ya que este modelo posee mejores características. Con este modelo se pudieron

obtener los valores que aporta cada variable predictora a la modelación del COS con esta metodología, destacando por ejemplo, que en la Componente 1 las variables que se relacionan de manera positiva con el COS son: M.O(Materia orgánica), C.Org(Carbono orgánico), RetenciónFosfatos (Retención de Fosfatos), Al.Extraible (aluminio extraíble) y CIC(Capacidad de Intercambio Catiónico).

El modelo de RKPLS con cinco componentes KPLS tiene un $R^2 = 0.8097$ los cual indica aproximadamente el $80.97\,\%$ de la variación del modelo queda explicada, además el $R^2_{Adj} = 78.26\,\%$. En este caso el valor AIC es menor y de igual forma para el $PRESS$ se tiene que, el modelo con cinco componentes tiene mejor capacidad predictiva.

De acuerdo con las observaciones anteriores, se puede afirmar que para el estudio del COS, la técnica RKPLS es superior a RCPK, ya que en la mayoría de los casos, los valores obtenidos para los criterios de selección de modelos fueron mejores para RKPLS, confirmando la eficiencia de la aplicación de dicha metodología. Además con esta técnica se tiene mayor precisión predictiva, ya que los intervalos de predicción son más estrechos para RKPLS.

Se puede concluir que el aprendizaje de nuevas metodologías ha brindado resultados muy útiles para la modelación del COS. De igual forma el manejo del software R y sus diferentes paqueterías resultan una herramienta de gran utilidad para el estudio de datos en temas ambientales.

Apéndice A

Tablas de Resultados

A.1. Base de datos utilizada

$COS(ton/ha)$	$Densidad\,Aparente(g/cm3)$	$\%Arena$	$\%Limo$	$\%Arcilla$	$\%M.O.$	$\%C.Org.$	$\%N.Tot.$
19.22	0.61	30.10	42.70	27.20	18.10	10.50	0.81
17.47	0.78	33.30	41.60	25.10	1.20	0.70	0.06
12.48	0.78	34.80	39.10	26.10	6.90	4.00	4.27
24.57	0.78	43.20	35.60	21.20	1.60	0.90	0.07
19.34	0.62	31.00	39.30	29.70	9.00	5.20	0.32
34.56	0.80	40.50	32.90	26.60	3.10	1.80	0.14
16.33	0.71	28.20	40.20	31.60	7.90	4.60	0.33
38.64	0.69	35.80	42.30	21.90	2.80	1.60	0.12
22.20	0.80	48.30	25.60	26.10	6.40	3.70	0.25
73.94	0.79	43.80	33.60	22.60	4.50	2.60	0.20
22.19	0.87	59.10	25.40	15.50	2.90	1.70	0.11
12.28	0.89	70.70	13.60	15.70	0.50	0.30	0.02
46.90	0.70	35.10	48.60	16.30	11.50	6.70	0.39
31.25	0.63	47.90	26.60	25.50	2.80	1.60	0.13
40.37	0.69	39.10	49.10	11.80	6.80	3.90	0.26
5.14	0.79	53.70	33.90	12.40	0.40	0.20	0.01
18.82	0.71	45.40	35.60	19.00	9.10	5.30	0.41
42.24	0.66	39.00	33.50	27.50	5.50	3.20	0.23
76.73	0.63	42.10	25.60	32.30	10.00	5.80	0.41
51.87	0.57	41.60	38.30	20.10	2.30	1.30	0.07
70.00	0.50	45.70	32.30	22.00	12.00	7.00	0.54
16.86	0.77	30.20	48.50	21.30	0.50	0.30	0.04
171.05	0.69	27.80	34.70	37.50	12.70	7.40	0.46
481.90	0.72	20.30	45.50	34.20	11.90	6.90	0.53
22.13	0.75	35.30	31.80	32.90	10.10	5.90	0.45
28.00	0.64	37.90	22.60	39.50	4.40	2.50	0.19
62.44	0.80	73.80	16.00	10.20	7.69	4.46	0.46
60.06	0.84	61.10	26.00	12.90	2.25	1.30	0.13
225.72	0.76	45.84	37.64	16.52	13.67	7.92	0.60
364.78	0.69	67.84	25.64	6.52	10.25	5.94	0.50
53.12	0.53	45.80	41.64	12.88	13.30	7.71	0.00
117.95	0.66	73.48	23.64	2.88	5.99	3.47	0.00
30.19	0.86	46.20	32.40	21.40	4.60	2.70	0.36
27.92	0.94	66.20	22.70	11.10	1.50	0.90	0.20
52.89	0.60	89.32	9.00	1.68	11.26	6.53	0.45
125.52	0.67	92.80	5.40	1.75	8.50	4.93	0.50
87.00	0.75	53.00	34.64	12.36	13.80	8.00	0.16
32.04	0.89	56.00	28.64	15.36	1.30	0.80	0.09
64.13	0.75	45.22	27.28	27.50	9.90	5.70	0.35
36.30	0.61	50.22	38.28	11.50	5.90	3.40	0.25
40.33	0.78	45.12	30.00	24.88	8.10	4.70	0.35
35.18	0.75	47.12	44.00	8.88	2.50	1.40	0.15

Tabla A.1: Propiedades obtenidas en la muestra.

C/N	$\%Ret.Fosfatos$	$\%Al\,Ext.$	$\%Fe\,Ext.$	$Al+1/2Fe\,extr.(\%)$	$pH(NaF\,1N)$	$pH(2:1\,H2O:S)$
13.00	46.00	1.17	0.35	1.34	9.20	6.00
11.00	50.40	1.51	0.53	1.77	9.40	6.90
15.00	51.50	1.93	0.60	2.23	9.40	5.10
12.00	63.20	2.05	0.60	2.35	10.50	7.00
16.00	60.90	1.58	0.44	1.80	10.30	6.50
13.00	57.00	1.32	0.32	1.48	10.80	7.00
14.00	50.50	1.42	0.28	1.56	9.80	5.30
13.00	44.00	2.13	0.25	2.26	10.00	5.90
15.00	61.80	1.05	0.37	1.23	9.50	5.00
13.00	65.30	1.23	0.34	1.40	9.80	6.00
15.00	60.20	1.52	0.30	1.67	9.00	6.00
17.00	58.00	1.41	0.22	1.52	9.20	6.00
17.00	34.90	1.93	0.35	2.10	10.20	5.90
12.00	56.70	2.35	0.66	2.68	11.20	6.10
15.00	72.30	1.88	0.65	2.20	10.50	5.00
13.00	63.40	1.89	0.21	1.99	10.00	6.00
13.00	64.70	1.80	0.49	2.04	10.60	5.40
14.00	70.40	2.06	0.27	2.20	10.80	5.60
14.00	70.00	2.03	0.34	2.20	10.80	5.30
18.00	74.00	2.96	0.66	3.30	9.50	6.20
13.00	86.90	3.42	1.80	4.32	11.50	4.80
16.00	70.50	2.09	0.91	2.54	10.8	4.40
16.00	80.60	3.09	1.25	3.72	10.30	5.10
13.00	87.20	3.06	1.96	4.04	10.80	5.40
13.00	76.20	2.98	0.94	3.45	10.70	6.00
13.00	80.90	3.42	1.00	3.92	10.70	6.30
9.69	68.10	1.08	1.02	1.60	48.75	5.40
10.00	76.50	1.01	0.45	1.23	36.75	5.30
13.20	82.20	3.72	0.50	3.97	35.50	5.10
11.88	85.90	5.87	0.90	6.32	58.75	6.30
7.38	87.90	3.58	0.67	3.92	49.50	5.00
4.35	72.5	3.82	0.35	4.00	67.75	5.00
7.38	65.90	1.71	0.25	1.83	46.50	6.80
4.35	67.00	0.86	0.15	0.93	40.25	7.40
14.50	63.00	2.28	0.52	2.54	53.25	5.30
9.86	72.40	2.73	0.40	2.93	54.50	5.90
13.30	76.30	2.15	0.36	2.33	29.50	6.50
8.89	78.10	1.82	0.83	2.24	44.00	6.00
14.72	80.70	4.11	0.55	4.38	23.00	5.50
13.60	79.60	3.85	0.55	4.125	43.25	5.70
13.42	81.20	3.74	1.60	4.54	34.50	5.80
9.33	83.60	3.13	1.90	5.08	41.25	6.10

Tabla A.2: Propiedades obtenidas en la muestra.

$pH(2:1\,KCl:S)$	$Delta\,pH$	$CIC(cmol(+)/Kg\,S)$	$\%V$	Ca	Mg	Na	K
5.20	-0.80	19.40	29.40	3.30	1.60	0.20	0.60
6.20	-0.70	15.10	40.40	2.50	1.40	0.70	1.50
4.40	-0.70	13.55	36.90	2.90	1.10	0.40	0.60
6.50	-0.50	14.85	36.40	3.30	1.40	0.30	0.40
5.80	-0.70	19.25	54.50	5.40	2.80	0.90	1.40
6.20	-0.80	13.25	45.28	2.80	10.00	0.80	1.40
4.50	-0.80	13.20	47.00	3.00	2.30	0.40	0.50
5.20	-0.70	12.80	36.70	2.40	1.50	0.60	0.20
4.40	-0.60	16.40	28.40	3.40	0.60	0.40	0.50
5.30	-0.70	15.00	29.80	4.00	1.00	0.40	0.50
5.10	-0.90	21.40	31.30	3.40	0.90	1.40	1.00
5.20	-0.80	14.30	29.30	2.00	0.60	0.90	0.60
5.30	-0.60	33.00	10.60	2.30	0.50	0.40	0.30
5.40	-0.70	30.70	10.20	1.50	1.00	0.50	0.10
4.20	-0.80	21.30	30.00	3.00	2.00	0.80	0.60
5.20	-0.80	25.00	32.80	5.00	2.00	0.60	1.60
4.70	-0.70	19.40	19.80	2.00	1.00	0.35	0.50
4.90	-0.70	14.30	22.40	1.50	1.00	0.40	0.30
4.80	-0.50	34.60	14.20	3.50	1.00	0.20	0.10
5.70	-0.50	29.00	6.70	1.30	0.20	0.30	0.10
4.00	-0.80	35.80	27.40	5.60	1.40	2.00	0.80
5.40	-1.00	10.50	40.90	2.00	0.80	1.20	0.30
4.50	-0.60	16.80	50.60	4.10	1.30	2.80	0.30
4.70	-0.70	13.80	38.30	4.30	0.70	0.20	0.20
5.40	-0.60	34.20	55.70	9.70	3.10	2.20	4.00
5.60	-0.70	33.70	30.10	6.50	1.20	0.70	1.70
4.40	-1.00	30.50	10.59	1.06	0.21	1.60	0.30
4.40	-0.90	31.80	8.89	1.00	0.06	1.40	0.40
4.80	-0.30	41.10	17.00	1.90	0.20	1.67	0.30
5.00	-1.30	53.30	8.33	0.28	0.28	1.76	0.30
4.10	-0.90	29.10	4.73	0.84	0.21	0.62	0.08
4.40	-0.60	38.50	5.11	0.42	0.22	1.10	0.23
5.30	-1.50	11.05	43.50	2.70	1.30	0.40	0.31
5.50	-1.90	10.35	47.50	1.80	1.70	0.56	0.30
4.40	-0.90	29.00	32.70	5.00	3.00	0.60	0.60
4.70	-1.20	18.50	48.30	5.60	1.50	0.70	0.45
4.70	-1.80	48.40	9.30	3.00	1.20	0.18	0.14
4.30	-1.70	53.20	8.80	3.00	1.00	0.31	0.38
5.10	-0.40	44.20	12.00	4.00	1.00	0.20	0.13
5.60	-0.10	39.90	10.20	2.10	1.60	0.20	0.16
4.80	-1.00	27.30	13.47	2.00	1.00	0.37	0.31
5.20	-0.90	29.00	10.30	1.60	0.70	0.37	0.31

Tabla A.3: Propiedades obtenidas en la muestra.

A.2. Medidas de influencia para detección de outliers

En la Tabla A.4 se muestran las diferentes medidas de influencia asociadas al modelo de regresión lineal múltiple considerando las 42 observaciones. Las medidas que se encuentran marcadas por $*$ son aquellas que de acuerdo con los diferentes criterios se consideran outliers. Las medidas que se presentan son: Distancias de Mahalanobis (DM), Distancias de Cook's ($cook.d$), Residuos Estandarizados (r_i), Residuos Estudentizados (t_i) y Potencial de un punto (leverage, h_i).

Obs.	DM	$cook.d$	r_i	t_i	h_i	*
1	18.927911	1.62e-02	-0.62001357	-0.60967603	0.492	
2	15.635646	1.21e-03	-0.20434943	-0.19911807	0.400	*
3	38.453281	4.99e-02	-0.12972175	-0.12631783	0.986	*
4	17.857480	2.65e-02	-0.91821483	-0.91423938	0.419	
5	17.429171	1.25e-03	-0.18392451	-0.17917855	0.460	*
6	37.228310	4.20e+00	2.37430940	2.75566988	0.945	*
7	18.980753	1.14e-02	-0.52491408	-0.51465923	0.488	*
8	16.044348	2.30e-03	0.27053865	0.26383163	0.419	*
9	9.755683	1.36e-05	-0.03012892	-0.02932604	0.256	*
10	9.0873500	1.57e-02	1.03658052	1.03873233	0.251	
11	10.113711	2.74e-04	0.13584088	0.13228205	0.254	
12	19.433358	4.64e-04	-0.10236248	-0.09965981	0.505	*
13	25.014190	6.15e-02	0.89662243	0.89177865	0.638	
14	14.151868	1.40e-03	0.23487657	0.22894468	0.369	*
15	23.025842	8.00e-02	-1.20016908	-1.21512638	0.561	
16	12.584505	3.35e-03	0.39134573	0.38245248	0.335	
17	11.927463	1.03e-02	-0.74596045	-0.73693618	0.298	
18	15.531302	3.59e-03	0.34612576	0.33796125	0.408	*
19	13.744568	2.19e-02	0.94056130	0.93756214	0.363	
20	18.943781	6.60e-03	0.39651602	0.38754716	0.491	*
21	27.151185	2.23e-01	-1.62034892	-1.69887403	0.662	
22	39.024390	NaN	NaN	NaN	1.000	
23	25.250398	3.24e-02	-0.63256049	-0.62227649	0.651	*
24	29.343232	9.57e-01	3.87822876	8.26909888	0.594	*
25	20.284035	4.30e-03	0.17874368	0.17412278	0.756	*
26	21.865927	3.78e-04	-0.08758507	-0.08526626	0.531	*
27	15.393434	3.09e-03	-0.23134235	-0.22548991	0.571	*
28	16.890961	1.41e-02	0.71562064	0.70611505	0.387	
29	25.626448	8.39e-04	0.15521419	0.15117027	0.445	*
30	38.548009	3.41e-01	2.04293686	2.25098815	0.652	*
31	22.420662	6.20e-01	0.41455908	0.40533953	0.988	*
32	20.268725	3.08e-03	0.23143876	0.22558415	0.569	*
33	19.656741	3.07e-03	-0.24979073	-0.24352864	0.531	*
34	21.100810	3.74e-02	-0.90821998	-0.90383356	0.511	
35	19.511811	9.50e-03	0.42517977	0.41582253	0.547	*
36	21.508567	7.84e-03	0.41455908	0.40533953	0.512	*
37	27.538050	2.74e-03	-0.22306687	-0.21740221	0.559	*
38	18.022710	1.09e-02	0.31823228	0.31057335	0.712	*
39	19.436155	7.27e-02	-1.39595228	-1.43425990	0.462	
40	28.276437	5.15e-02	-1.07687863	-1.08168683	0.505	
41	39.010790	3.45e-01	-2.71397693	-3.37576130	0.519	*
42		8.68e+01	-0.84376859	-0.83709713	1.000	*

Tabla A.4: Medidas de influencia.

A.3. Resultados de RCPK mediante el Kernel Gaussiano

Las siguientes Tablas A.5, A.6 y A.7, muestran los valores para R^2, R^2_{Adj}, AIC y $PRESS$, los cuales se obtuvieron en los modelos de regresión con Componentes Principales mediante el uso del Kernel Gaussiano con $\sigma = 0.001$, $\sigma = 0.01$ y $\sigma = 0.05$.

$\sigma = 0.001$

N. Comp	R^2	R^2_{Adj}	AIC	$PRESS$
(10)	0.3499	0.1332	465	221136
(9)	0.3201	0.1227	464	217510
(8)	0.3200	0.1500	462	212112
(7)	0.3191	0.1746	460	196639
(6)	0.3024	0.1793	459	197770
(5)	0.2090	0.0959	463	183020
(4)	0.2032	0.1146	461	179233
(3)	0.1882	0.1224	460	173204
(2)	0.1879	0.1452	458	166257
(1)	0.1606	0.1391	457	162877

Tabla A.5: Comparación de R^2_{Adj}, $PRESS$ y AIC.

$\sigma = 0.01$

N. Comp	R^2	R^2_{Adj}	AIC	$PRESS$
(10)	0.2356	−0.0191	457	295375
(9)	0.2131	−0.0154	470	278748
(8)	0.2048	0.0060	469	267925
(7)	0.2021	0.0325	467	178525
(6)	0.0869	−0.0741	470	176346
(5)	0.0534	−0.0818	470	184851
(4)	0.0533	−0.0517	468	180079
(3)	0.0437	−0.0374	466	181022
(2)	0.0137	−0.0381	466	184886
(1)	0.0007	−0.0248	464	185996

Tabla A.6: Comparación de R^2_{Adj}, $PRESS$ y AIC.

$\boldsymbol{\sigma = 0.05}$

N. Comp	R^2	R^2_{Adj}	AIC	$PRESS$
(10)	0.1869	−0.0814	474	270505
(9)	0.1849	−0.0517	472	205508
(8)	0.1826	−0.0217	470	201613
(7)	0.1825	0.0090	468	191613
(6)	0.0899	−0.0707	470	182397
(5)	0.0667	−0.0666	469	188216
(4)	0.0588	−0.0456	468	179489
(3)	0.0459	−0.0314	466	180202
(2)	0.0191	−0.0325	465	184130
(1)	0.0007	−0.0249	464	185961

Tabla A.7: Comparación de R^2_{Adj}, $PRESS$ y AIC.

A.4. Resultados de RCPK mediante el Kernel Polinomial

Las siguientes Tablas A.8, A.9 y A.10, muestran los valores para R^2, R^2_{Adj}, AIC y $PRESS$, los cuales se obtuvieron en los modelos de regresión con Componentes Principales mediante el uso del Kernel Polinomial con $\alpha = 1$, $\alpha = 2$ y $\alpha = 3$.

$\boldsymbol{\alpha = 1}$

N. Comp	R^2	R^2_{Adj}	AIC	$PRESS$
(10)	0.4744	0.2993	456.29	201866
(9)	0.4522	0.2932	455.99	206899
(8)	0.4422	0.3027	454.73	170813
(7)	0.4051	0.2790	455.37	166532
(6)	0.3659	0.2540	455.99	166072
(5)	0.3579	0.2662	454.50	161654
(4)	0.3144	0.2382	455.19	163217
(3)	0.3142	0.2586	453.20	159018
(2)	0.2518	0.2124	454.77	162873
(1)	0.2282	0.2084	454.05	160495

Tabla A.8: Comparación de R^2_{Adj}, $PRESS$ y AIC.

$\boldsymbol{\alpha = 2}$				
N. Comp	R^2	R^2_{Adj}	AIC	$PRESS$
(10)	0.5363	0.3818	451.16	198424
(9)	0.4851	0.3356	453.45	189479
(8)	0.4830	0.3537	451.62	176407
(7)	0.4469	0.3295	452.39	172375
(6)	0.4160	0.3129	452.62	165636
(5)	0.4098	0.3254	451.05	160282
(4)	0.3647	0.2941	452.07	158170
(3)	0.3642	0.3127	450.10	154343
(2)	0.3135	0.2773	451.25	156705
(1)	0.2788	0.2603	451.27	154482

Tabla A.9: Comparación de R^2_{Adj}, $PRESS$ y AIC.

$\boldsymbol{\alpha = 3}$				
N. Comp	R^2	R^2_{Adj}	AIC	$PRESS$
(10)	0.6057	0.4743	444.51	196743
(9)	0.5207	0.3816	450.51	202661
(8)	0.5061	0.3827	449.74	187898
(7)	0.4924	0.3847	448.87	177444
(6)	0.4664	0.3722	448.92	168803
(5)	0.4594	0.3822	447.45	162844
(4)	0.4100	0.3444	449.04	154319
(3)	0.4086	0.3607	447.13	150113
(2)	0.3665	0.3332	447.95	151472
(1)	0.3198	0.3024	448.87	149668

Tabla A.10: Comparación de R^2_{Adj}, $PRESS$ y AIC.

A.5. Resultados de RCPK mediante el Kernel Lapaciano

Las siguientes Tablas A.11, A.12 y A.13, muestran los valores para R^2, R^2_{Adj}, AIC y $PRESS$, los cuales se obtuvieron en los modelos de regresión con Componentes Principales mediante el uso del Kernel Laplaciano con $\sigma = 0.001$, $\sigma = 0.01$ y $\sigma = 0.05$.

$\sigma = 0.001$

N. Comp	R^2	R^2_{Adj}	AIC	$PRESS$
(10)	0.4606	0.2275	460.29	213585
(9)	0.4080	0.2362	459.17	201250
(8)	0.4080	0.2600	457.17	187127
(7)	0.3931	0.2644	456.19	181133
(6)	0.3512	0.2368	456.93	187711
(5)	0.3307	0.2351	456.20	164289
(4)	0.2840	0.2045	456.97	166788
(3)	0.2814	0.2232	455.12	160533
(2)	0.2258	0.1851	456.17	163719
(1)	0.2227	0.2028	454.34	159164

Tabla A.11: Comparación de R^2_{Adj}, $PRESS$ y AIC.

$\sigma = 0.01$

N. Comp	R^2	R^2_{Adj}	AIC	$PRESS$
(10)	0.3107	0.0800	467.41	219086
(9)	0.3075	0.1065	465.60	216645
(8)	0.3062	0.1327	463.68	209762
(7)	0.3016	0.1535	461.95	197866
(6)	0.2004	0.0592	465.50	195235
(5)	0.1484	0.0267	466.08	191811
(4)	0.1379	0.0421	464.58	186611
(3)	0.1366	0.0666	462.64	179754
(2)	0.1331	0.0874	460.81	167734
(1)	0.1313	0.1090	458.90	165218

Tabla A.12: Comparación de R^2_{Adj}, $PRESS$ y AIC.

$\sigma = 0.05$

N. Comp	R^2	R^2_{Adj}	AIC	$PRESS$
(10)	0.3579	0.1439	464.50	215741
(9)	0.3407	0.1493	463.59	212253
(8)	0.3397	0.1746	461.65	202116
(7)	0.3380	0.1976	459.75	194748
(6)	0.3345	0.2171	457.97	191426
(5)	0.2015	0.0873	463.44	185094
(4)	0.2009	0.1121	461.47	179380
(3)	0.1921	0.1266	459.92	171817
(2)	0.1918	0.1493	457.94	164431
(1)	0.1707	0.1494	456.99	161595

Tabla A.13: Comparación de R^2_{Adj}, $PRESS$ y AIC.

Apéndice B

Software disponible en R

El programa R es un entorno de análisis y programación estadístico que forma parte del proyecto de software libre GNU *General Public Licence*. R está disponible en la dirección `http://www.r-project.org`. Como programa de análisis estadístico, R permite realizar tareas estadísticas sencillas habituales y además permite extensiones que implementan técnicas estadísticas avanzadas. De este modo cubre las necesidades de cualquier analista, tanto en el ámbito de la estadística profesional como en el de la investigación estadística.

R consta de un sistema base pero la mayoría de las funciones estadísticas vienen agrupadas en distintas librerías (packages) que se incorporan de forma opcional según la necesidad del analista. En este apartado se hace una descripción casi exhaustiva de algunas de las funciones y paqueterías utilizadas para el análisis desarrollado.

Lo primero que debemos hacer es crear la base de datos que vamos a analizar.

```
base<-read.table(file=''clipboard'', head=T)
data.frame(base)
```

Con esta indicación se pueden leer los datos desde Excel, solo seleccionando el conjunto deseado. Después se crea un macro (conjunto) de datos y los muestra en pantalla.

Para realizar un análisis básico de regresión R proporciona un resumen del ajuste del modelo y un diagnóstico gráfico del modelo ajustado. En cuanto a las estadísticas para la selección de modelos, es necesario utilizar el *paquete CombMSC* para el cálculo del PRESS y en el caso del AIC está en las funciones básicas del programa.

```
fit <- lm(formula, base)
summary(fit)
par(mfrow=c(2,2))
plot(fit)
library(CombMSC)
AIC(fit)
PRESS(fit)
```

En cuanto al análisis para detectar observaciones outlier, se requiere hallar las medidas de influencia y posteriormente hacer pruebas para verificación de los resultados, en este caso se hace uso del paquete *outliers*.

```
#Medidas de influencias presentadas en el modelo
influence.measures(fit)
#Gráficos para el diagnóstico
plot(fit)
boxplot(fit)
library(outliers)
outlier(x, opposite = FALSE, logical = FALSE)
outlier(y)
chisq.out.test(x, variance=var(x), opposite = FALSE)
chisq.out.test (y)
```

Ahora para detectar problemas de multicolinealidad se hace un análisis del eigensistema asociado al modelo. Para obtener los valores VIF se requiere del paquete *faraway*.

```
#Crea la matriz x que corresponde a las variables de respuesta.
x<-model.matrix(g)[,-1]
#Hace un análisis del eigensistema de la matriz (x'x)
e<-eigen(t(x)%*% x)
#Muestra los eigenvalores correspondientes a (x'x).
e$val
#Calcula los números de condición asociados a (x'x) sqrt(e$val[1]/e$val)
library(faraway)
vif(x)
```

La siguiente etapa es el análisis de Componentes Principales con Kernels, para ello se requiere del paquete *kernlab*. Este paquete permite elegir la función Kernel que se quiera utilizar, en este caso los Kernels que se utilizaron son:

`rbfdot(sigma = 1)`: Kernel Gaussiano
`polydot(degree = 1, scale = 1, offset = 1)`: Kernel Polinomial
`laplacedot(sigma = 1)`: Kernel Laplaciano

donde los parámetros están dados por:

`sigma`: amplitud de la función.
`degree`: grado del polinomio.
`scale`: parámetro de escala.

`offset`: parámetro de desplazamiento.

El análisis de componentes principales en este caso se hace mediante el uso de la matriz Kernel, para ello primero se debe asignar una función Kernel al igual que los parámetros que se requieren, así mismo se debe elegir el número de componentes que se desea encontrar.

```
library(kernlab)
#Asigna la matriz de respuestas
matriz<-as.matrix(base)
#Asigna la función Kernel y el parámetro
rbf <- rbfdot(sigma = 0.05)
#Calcula la matriz Kernel
kernelMatrix(rbf, matriz)
#Realiza el ACP usando la matriz Kernel calculando las componentes
kpc<-kpca(mg, features = 1, th = 1e-4)
#Regresa las componentes principales y las asigna a un elemento
comp<-pcv(kpc)
#Regresa los eigenvalores
eig(kpc)
#Regresa los datos proyectados en el espacio de CP (kernel)
rotated(kpc)
#Regresa el tipo de Kernel utilizado
kernelf(kpc)
```

La última técnica es la Regresión de Mínimos Cuadrados Parciales, como se menciona en (Mevik et al., 2007) el *paquete pls* de R, implementa diferentes algoritmos para obtener las componentes latentes, mismas que sirven como base para el nuevo ajuste de regresión.

```
library(pls)
#Método KernelPLS
plsK <- plsr(formula , ncomp=1, method = ''kernelpls'', data = base,
validation=''LOO'', scale=T)
#Resumen de resultados
summary(plsK)
#Muestra los Scores (puntuaciones) para X
scores(plsK)
#Muestra los Loadings (Cargas) para cada componente
loadings(plsK)
#Grafica las cargas para cada componente
loadingplot(plsK, comps = 1:5)
#Grafica las correlaciones para cada componente
corrplot(plsK, comps = 1:5)
```

Es conveniente señalar que a comparación de la técnica de Componentes Principales, la técnica PLS no permite asignar de manera automática las componentes latentes a un nuevo elemento para su posterior uso, de modo que este procedimiento se tiene que hacer manualmente, es decir, debemos crear una nueva base de datos con las componentes que se eligen para realizar la regresión.

Bibliografía

Ampanthong, P. and Suwattee, P. (2009). A comparative study of outlier detection procedures in multiple linear regression. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 1.

Aparicio, J., Martinez, M., and Morales, J. (2004). Modelos lineales aplicados en R. *Dto. Estadística, Matemáticas e Informática*.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404.

Carrascal, L. M. (2015). Teoría y Praxis de modelos generalizados:"Infiriendo patrones con el paquete estadístico R".

Castillo-Morales, M., Linares Fleites, G., Valera Pérez, M., García-Calderón, N., and Acevedo-Sandoval, O. (2009). Modelación de la materia orgánica en suelos volcánicos de la región de Teziutlán, Puebla, México. *Revista Latinoamericana de Recursos Naturales*, 5(2):148–154.

Faraway, J. J. (2014). *Linear models with R*. CRC press.

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media.

Friedberg, S. H., Insel, A. J., and Spence, L. E. (1997). *Linear algebra*, volume 1. Prentice Hall Inc., Upper Saddle River, NJ.

Gibaja Martíns, J. J. (2010). Aprendizaje estadístico con funciones kernel.

Karatzoglou, A., Smola, A., Hornik, K., and Karatzoglou, M. A. (2016). Package kernlab.

Kreyszig, E. (1989). *Introductory functional analysis with applications*, volume 1. Wiley New York.

Linares, G., Valera, M. A., and Castillo, M. (2014). Modelación espacial de los contenidos de carbono orgánico en suelos volcánicos de Teziutlán, Puebla, México. *Ciencia en la frontera: revista de ciencia y tecnología de la UACJ*, pages 55–63.

López Pineda, G. (2013). *Análisis de Regresión para la Estimación del Secuestro de Carbono Orgánico en Suelos*.

Martínez, J. L. and Barrios, H. (2016). Selección del número de factores latentes apropiados en PLSR con capacidad predictiva. *Memorias en extenso en: XXVI Simposio Internacional de Estadística. Sincelejo, Sucre, Colombia*.

Medina, C. D. A., Duque, G. A. C., and Flórez, J. A. F. (2016). Modelos de regresión pls aplicados a variables educativas. *Scientia et technica*, 21(3):254–263.

Mevik, B.-H., Wehrens, R., et al. (2007). The pls package: principal component and partial least squares regression in r. *Journal of Statistical software*, 18(2):1–24.

Mevik, B.-H., Wehrens, R., Liland, K. H., Mevik, M. B.-H., and Suggests, M. (2016). Package pls.

Montano Rivas, J. A. (2013). *Análisis de componentes principales con kernels: una propuesta de mejora del kernel*. PhD thesis.

Montgomery, D. C. D. C., Peck, E. A., and Vining, G. G. (2004). *Introducción al análisis de regresión lineal*.

Preda, C. (2007). Regression models for functional data by reproducing kernel hilbert spaces methods. *Journal of Statistical Planning and Inference*, 137(3):829–840.

Ramsay, J. O. (2006). *Functional data analysis*. Wiley Online Library.

Rodríguez, E., Vinante, C., and Leal, M. (2009). Enfoque óptimo del método kernel cuadraos mínimos parciales. *SABER*, 21(2).

Rodríguez, E., Vinante, C., and Leal, M. (2012). Utilización combinada de métodos exploratorios y confirmatorios para el análisis de la actividad antibacteriana de la cefalosporina (Parte II). *Revista investigación operacional*, 33(3).

Rosipal, R. and Trejo, L. J. (2001). Kernel partial least squares regression in reproducing kernel hilbert space. *Journal of machine learning research*, 2(10):97–123.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319.

Vega-Vilca, J. C. and Guzmán, J. (2011). Regresión pls y pca como solución al problema de multicolinealidad en regresión múltiple. *Revista de Matemática Teoría y Aplicaciones*, 18(1):09–20.

Zeileis, A., Hornik, K., Smola, A., and Karatzoglou, A. (2004). kernlab-an S4 package for kernel methods in R. *Journal of statistical software*, 11(9):1–20.