COLEGIO DE POSTGRADUADOS

INSTITUCIÓN DE ENSEÑANZA E INVESTIGACIÓN EN CIENCIAS AGRÍCOLAS

CAMPUS MONTECILLO

POSTGRADO EN SOCIOECONOMÍA-ESTADÍSTICA E INFORMÁTICA-ESTADÍSTICA

Regresión de Mínimos Cuadrados Parciales para Datos Variedad-Valuados

Raúl Alberto Pérez Agámez

T E S I S

PRESENTADA COMO REQUISITO PARCIAL PARA OBTENER EL GRADO DE:

DOCTOR EN CIENCIAS

MONTECILLO,TEXCOCO, EDO. DE MÉXICO

Diciembre de 2012

La presente tesis titulada: **Regresión de Mínimos Cuadrados Parciales para Datos Variedad-Valuados**, realizada por el alumno: **Raúl Alberto Pérez Agámez**, bajo la dirección del Consejo Particular indicado ha sido aprobada por el mismo y aceptada como requisito parcial para obtener el grado de:

DOCTOR EN CIENCIAS

SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA
ESTADÍSTICA

CONSEJO PARTICULAR

CONSEJERO

Dr. José A. Villaseñor Alva

DIRECTOR DE TESIS

Dra. Graciela González Farias

ASESOR

Dr. Filemón Ramiréz Pérez

ASESOR

Dra. Elizabeth González Estrada

ASESOR

Dr. Barry C. Arnold

Montecillo, Texcoco, Estado de México, Diciembre de 2012

III

Regresión de Mínimos Cuadrados Parciales para Datos Variedad-Valuados

Raúl Alberto Pérez Agámez

Colegio de Postgraduados, 2012

En esta tesis se desarrolla la metodología de regresión de mínimos cuadrados parciales (PLS) para datos variedad-valuados. Primero se realiza una revisión sobre algunos métodos modernos de regresión no-lineal, luego se continúa con una exploración de la metodología de regresión para datos variedad-valuados y finalmente se desarrolla una nueva metodología de regresión para datos variedad-valuados, para la cual se demuestra que en ciertas situaciones especiales, produce mejores resultados que las técnicas tradicionales de regresión que son aplicables a este tipo de datos.

Aunque la metodología desarrollada es aplicable a variedades generales, esta se ilustra en aplicaciones de conjuntos de datos de ciertas estructuras orgánicas que son representadas geométricamente utilizando la representación medial axial (m-rep) de objetos geométricos y a matrices simétricas positivas definidas (PD) que se obtienen a partir de Imágenes de Resonancia Magnéticas (MRI) por tensor de difusión (DT). Se comparan las técnicas clásicas de regresión que existen para este tipo de datos con la nueva metodología de regresión PLS y se observa un mejor desempeño de este último modelo, con lo cual se muestra que la nueva técnica tiene algunas ventajas sobre las ya existentes.

Palabras clave: Variedad Riemanniana, datos variedad-valuados, regresión sobre variedades, imágen de resonancia magnética, imágen por tensor difusión, regresión PLS.

Partial Least Squares Regression (PLS) for Manifold-Valued Data

Raúl Alberto Pérez Agámez

Colegio de Postgraduados, 2012

In this dissertation I develop the methodology of partial least squares regression (PLS) for manifold-valued data. First, makes a review of some modern methods of nonlinear regression, then continues with an exploration of regression methodology for manifold-valued data and finally develops a new methodology of regression for manifold-valued data for which is shown that in certain special situations produces better results than traditional regression techniques that are applicable to this type of data.

Although the methodology developed is applicable to general manifolds, this is illustrated in applications of datasets of certain organic structures that are represented geometrically using the axial medial representation (m-rep) of geometric objects and at symmetric positive definite matrices (PD) obtained from Magnetic Resonance Imaging (MRI) diffusion tensor (DT). We compared different techniques are classic regression for these data with the new methodology and PLS regression showed a better performance This latter model, which shows that the new technique has several advantages over existing ones.

Key words: Riemannian manifold, manifold-valued data, regression on manifolds, Magnetic Resonance Imaging (MRI), Diffusion Tensor Imaging (DTI), PLS regression.

AGRADECIMIENTOS

Este trabajo no habría sido posible sin el apoyo y orientación de mi directora de tesis, la Dra. Graciela González Farías, bajo cuya orientación escogí este tema de investigación sobre el cual he desarrollado esta tesis. Al Dr. José Villaseñor Alva, mi consejero durante mis estudios en el Colegio de Postgraduados (CP), por sus valiosos comentarios con respecto al tema de investigación y por sus consejos y ánimos en momentos difíciles. A la Dra. Elizabeth González Estrada del CP, al Dr. Filemón Ramiréz Pérez de la Universidad Autónoma Chaíngo y al Dr. Barray C. Arnold de la Universidad de California Riverside, quienes también hicieron parte de mi consejo particular durante mis estudios y quienes con sus valiosos aportes y sugerencias con respecto al tema de investigación hicieron posible la culminación con éxito de este trabajo.

Agradezco a la Universidad Nacional de Colombia por el esfuerzo que han realizado, desde mis compañeros de trabajo de la Escuela de Estadística de la sede Medellín hasta distintos entes administrativos, en pro de mi formación académica.

Agradezco al Departamento Administrativo de Ciencia, Tecnología e Innovación de Colombia, COLCIENCIAS, por el apoyo en la financiación de este programa doctoral a través de un crédito-beca condonable, los cuales fueron destinados a gastos de colegiaturas, sostenimientos y otros gastos necesarios.

Agradezco al Dr. Rogelio Ramos Quiroga, profesor del CIMAT de México, por su apoyo en los momentos difíciles durante mi estancia en México.

Agradezco a todo el personal del postgrado de Estadística del Colegio de Postgraduados: secretarias, personal técnico y personal del laboratorio de cómputo por toda la comprensión y el apoyo brindado en los momentos que lo necesitaba, especialmente a la Emma Olivares, por saber ayudarme en los distintos apuros que mantenía involucrado. A los doctores Humberto Vaquera Huerta y Gustavo Ramiréz Valverde, por el apoyo brindado durante mi incorporación y estancia al CP, a los doctores Sergio Pérez Elizalde y Paulino Pérez Pérez por todos sus opiniones y sugerencias con respecto a mi tema de investigación.

A mis amigos y amigas, los que compartieron conmigo durante el periodo en que se desarrolló este trabajo, por haberme brindado esos momentos tan maravillosos de esparcimiento, entre ellos a Luis Fernando Contreras Cruz y Diego Hernandez Jarquín y también a los que por la distancia no pudieron hacerlo, pero me esperaron pacientemente, entre ellos a Antonio Osuna, quién siempre estuvo pendiente de mi estancia en

México, a mi compañero de trabajo y compadre Victor Ignacio López Ríos por su apoyo incondicional en todo lo que estaba a su alcance.

No podría terminar sin agradecer a mi familia, mis padres y hermanos, los cuáles a lo largo de mis años de estudios siempre confiaron en mis resultados, a mi señora Kellys Nallith Salcedo Hurtado por su constante apoyo, comprensión y ánimo en todo momento y a mis hijos Daniela Pérez Salcedo y Luís Manuel Pérez Salcedo, por estar siempre a mi lado.

DEDICATORIA

A mis padres y hermanos. . . ,

A mi esposa *Kellys Nallith* y

A mis hijos *Luis Manuel y Daniela.*

Índice

Índice de tablas

Índice de figuras

Capítulo 1

Introducción

Los recientes avances en física, ingeniería biomédica y ciencias de la computación, han generado el desarrollo de varias técnicas no invasivas, seguras y relativamente absequibles a imágenes médicas. Estas técnicas producen imágenes de alta resolución que le permiten a los médicos explorar en la búsqueda de enfermedades del ser humano. Los análisis cuidadosos de imágenes médicas extienden su utilidad más allá de una simple inspección visual, dando respuesta a preguntas críticas acerca de la anatomía, fisiología y enfermedades humanas.

Entre las áreas de investigación que componen el campo del análisis de imágenes médicas están la caracterización de formas anatómicas, la cual estudia la variabilidad en la forma geométrica de los objetos que describen la imágen.

1.1. Caracterización de Formas Anatómica

El término caracterización de formas puede ser definido como la aplicación de metodologías estadísticas enfocadas a medir y describir la variabilidad sobre formas geométricas de objetos, dicha forma puede caer dentro de un número de categorías conocidas. La caracterización de formas anatómicas estudia objetos biológicos, tales como los órganos del ser humano y calcula la forma geométrica de dichos órganos a partir del análisis de imágenes médicas.

1.1. Caracterización de Formas Anatómica

Una motivación para la caracterización de formas anatómicas recae en el potencial que tienen al servir como una herramienta de diagnóstico. Existen diversas enfermedades de las cuales se sabe que alteran la forma de ciertos órganos. Por ejemplo, se ha reportado que la forma de los hipocampos va cambiando con el progreso de la esquizofrenia. Al caracterizar las diferencias de formas entre hipocampos saludables (sanos) y enfermos y teniendo en cuenta la variabilidad normal en la forma de los hipocampos sobre la población humana, puede ser posible derivar un conjunto de reglas de decisión que pueden clasificar nuevos hipocampos, tanto sanos como enfermos, en distintos instantes de tiempo, dichas reglas pueden luego ser adicionadas a los diagnósticos iniciales de esquizofrenia.

Más allá de proporcionar diagnósticos, la caracterización de formas también puede ofrecer pistas importantes acerca de la naturaleza de ciertas enfermedades, determinando cuándo, donde y cómo es afectada la forma de los distintos órganos involucrados en la enfermedad. Aunque la naturaleza de la esquizofrenia no es totalmente entendida, evidencias recientes sugieren que la esquizofrenia afecta la forma de algunas partes de los hipocampos más que a otras. Si confirmamos tales evidencias, podríamos permitir a los neuro científicos enfocar sus investigaciones sobre las celdas contenidas en esa área particular de los hipocampos.

La caracterización de formas también puede sumarse al campo médico produciendo atlas anatómicos estadísticos. En la actualidad los libros de anatomía muestran algunos instantes típicos de la anatomía humana al igual que algunos instantes de anormalidades anatómicas. Los investigadores están usando técnicas de caracterización de formas para estimar la variabilidad sobre la forma de distintas partes del cuerpo humano, con el objetivo de construir atlas que puedan mostrar muchas variaciones reales de la anatomía. Estos atlas no son solamente hechos para propósitos educacionales, sino que también proporcionan algoritmos de segmentación automática de imágenes, generando un conocimiento inicial sobre la forma de los objetos que están siendo segmentados.

Para llevar a cabo un estudio de caracterización de forma, los investigadores seleccionan un grupo de individuos calificados y adquieren imágenes médicas de las áreas anatómicas de interés. En estudios que involucran una enfermedad, se selecciona un grupo de pacientes que han sufrido la enfermedad a lo largo del tiempo y un grupo de sujetos de control sanos con características similares a los pacientes. Los grupos de pacientes son alguna veces divididos dentro de subgrupos separados por pacientes en las diferentes etapas de la enfermedad o separados por diferentes tipos de tratamientos que se están llevando a cabo. La composición de los grupos en el estudio de caracterización de forma, deberían reflejar la variabilidad intra-población debido a ciertas covariables tales como la edad, el sexo, grupo étnico y otros factores. Frecuentemente es difícil y costoso encontrar a un número grande de individuos calificados, por lo que los métodos de caracterización

de forma están forzados a trabajar con tamaños muestrales pequeños.

La metodología de caracterización de formas establece un vínculo entre los métodos estadísticos de imágenes médicas y los métodos estadísticos estándar. Los métodos estadísticos estándar requieren que la información acerca de la forma sea representada por un conjunto fijo de variables aleatorias, llamadas características. Cada individuo en una muestra debe ser representado por una lista de números, los cuales son realizaciones de las características para ese individuo particular. El desafío de la caracterización de forma es derivar un conjunto apropiado de características que representen la información relevante relacionada a la forma contenida en una colección de imágenes médicas y describir cada imagen como una lista de realizaciones de características.

La transición de las imágenes a las características está basada en áreas del análisis de imágenes. Los algoritmos de segmentación automática pueden ser usados para hallar y extraer objetos de interés sobre imágenes médicas.

Los objetos extraídos necesitan ser representados de manera tal que reflejen sus propiedades geométricas. Por ejemplo, es común representar un objeto como una malla de puntos sobre la frontera. La estructura de la malla debería ser la misma para todos los instantes del objeto en la muestra y los puntos de las mallas correspondientes a diferentes instancias deberían ser colocados sobre las correspondientes ubicaciones en el objeto. Aunque es posible usar todos los parámetros que definen una representación de objeto como características, tal elección nos lleva frecuentemente a desarrollos estadísticos pobres. Usualmente es ventajoso desarrollar procedimientos adicionales y filtrados para derivar características a partir de la representación de objetos.

Los trabajos de caracterización de forma hacen uso de las técnicas estadísticas estándar tales como clasificación y estimación de densidades. La clasificación es usada en aplicaciones que estudian los efectos de enfermedades sobre la forma anatómica y se sitúa en el centro de las aplicaciones de diagnósticos basadas en formas. La estimación de densidades es usada para tareas tales como la construcción de atlas y la segmentación automática, donde es necesario evaluar la validez de un objeto asignándole una puntuación de densidad de probabilidad.

Ambas tareas de clasificación y estimación de densidades consisten de 4 fases: (i) selección del modelo, (ii) el entrenamiento del modelo, (iii) pruebas del modelo y (iv) la aplicación del modelo. La fase de selección del modelo involucra elegir un modelo estadístico apropiado para la tarea a mano. En estimación de densidades; la selección del modelo involucra elegir una distribución de probabilidad que pueda describir razonablemente la variabilidad presente en los datos. En clasificación, la selección del modelo significa elegir

uno entre muchos métodos de clasificación competentes. Los modelos estadísticos son frecuentemente seleccionados empíricamente, usando algunos conocimientos heurísticos acerca de la naturaleza del problema. Durante la fase de entrenamiento, los parámetros del modelo estadístico son entrenados usando una "muestra de entrenamiento". En clasificación, la muestra de entrenamiento consiste de múltiples instancias de cada una de las clases con miembros de cada clase conocida. Para estimación de densidad, el conjunto de entrenamiento simplemente contiene múltiple instancias válidas. Durante la fase de pruebas, la calidad del modelo estadístico entrenado es evaluada, aplicando éste modelo a otra muestra, la cual es similar a la muestra de entrenamiento, pero definitivamente es una muestra nueva. Las pruebas pueden ser usadas no solamente para validar el modelo estadístico, sino también para re-entrenar sus parámetros. Finalmente, la etapa de aplicación, involucra aplicar el modelo estadístico a nuevas instancias de datos para lo cual no se conoce su validez ni a qué clase pertenecen.

Un conjunto bien seleccionado de características es de importancia crítica para la clasificación y estimación de densidad. Por ejemplo, una buena elección de característica puede llevarnos a que la Distribución Gaussiana sea apropiada para describir la variabilidad en una muestra, mientras que una mala elección de características puede llevarnos a datos asimétricos o multimodales. En clasificación, un buen conjunto de características incluye aquellas que reflejan las diferencias entre clases, mientras que excluyen las que capturan variabilidad intra-población y ruido. Un buen conjunto de características lleva a clasificadores que se comportan bien durante las fases de pruebas y aplicaciones.

Capítulo 2

Elementos Básicos de Geometría Diferencial

En este capítulo se hace una revisión sobre las propiedades matemáticas de ciertos entes geométricos que se considerarán en esta tesis, como lo son los modelos de forma mediante representación medial axial, m-rep y los tensores de difusión, los cuales son elementos de ciertas variedades curvadas de alta dimensión, o más exactamente, son elementos de espacios simétricos Riemannianos. En este sentido es útil pensar un punto sobre un espacio simétrico como una transformación a partir de un punto base fijo. Por ejemplo al construir el espacio de tensores de difusión, el punto base es elegido como la matriz identidad y cualquier tensor de difusión se trata como una transformación a partir de la matriz identidad. Los espacios de transformación que se están usando se conocen como Grupos de Lie, los cuales son asimismo variedades suaves. La utilidad de estudiar estos Grupos de Lie de transformaciones de espacios simétricos es debido a que ellos tienden a ser algebraicos en naturaleza y por lo tanto ciertos cálculos sobre espacios simétricos, tales como distancias y trayectoria más cortas entre dos puntos, tienen frecuentemente soluciones cerradas. Estos mismos cálculos pueden requerir de la solución de ecuaciones diferenciales complejas si la variedad considerada no es un espacio simétrico. Debido a que la distancia y trayectoria más corta entre dos puntos son esenciales en la definición de ciertas estadísticas para variedades, los espacios simétricos son particularmente útiles para hacer análisis estadístico.

Muchos objetos geométricos son representables como Grupos de Lie (es decir, como espacios simétricos). Las transformaciones de espacios euclideanos tales como traslaciones, rotaciones, escalamientos y transformaciones afines aparecen como elementos de Gru-

pos de Lie. Las primitivas geométricas tales como vectores unitarios, planos orientados y matrices simétricas positivas definidas (PD), pueden ser vistas como puntos sobre espacios simétricos. Aquí se hace una revisión de la teoría matemática básica de Grupos de Lie y espacios simétricos. El estudio de tales espacios requiere inicialmente de algún conocimiento sobre topología básica y teoría de variedades, lo cuál tratará en las secciones que siguen. Los distintos espacios descritos en este capítulo son todos, de una u otra forma, generalizaciones del espacio euclídeo $\mathbb{R}^n$. El espacio euclídeo es un espacio topológico, como variedad Riemanniana es un Grupo de Lie y un espacio simétrico.

2.1. Conceptos básicos de Topología

El estudio de un espacio topológico surgió de la necesidad de generalizar la noción de continuidad en espacios euclídeos a espacios más generales. La topología es fundamental para construir la teoría de espacios de variedades y funciones. En esta sección se revisan los conceptos básicos necesarios para el estudio de variedades diferenciales. Para ver más al respecto puede ir a cualquier libro básico de topología, como por ejemplo, Munkres 1975.

2.1.1. Elementos Básicos de topología

Recuerde que la continuidad de una función sobre los reales es formulada en términos de intervalos abiertos, es decir, mediante la definición $\epsilon - \delta$ usual. Una topología define cuáles subconjuntos de un conjunto abierto X son abiertos, en la misma forma que un intervalo es abierto. Como se verá al final de esta subsección, los conjuntos abiertos en $\mathbb{R}^n$ son construidos como uniones de bolas abiertas de la forma, $B(x,r) = \{y \in \mathbb{R}^n : \|x - y\| < r\}$. Para un conjunto general X, este concepto de conjuntos abiertos puede ser formalizado mediante el siguiente conjunto de axiomas.

Definición 2.1. Una **Topología** sobre un conjunto abierto X es una colección $\mathcal{T}$ de subconjuntos de X, tales que cumple las siguientes condiciones:

1. El vacío Φ y X están en $\mathcal{T}$.
2. La unión de una colección arbitraria de elementos de $\mathcal{T}$ también está en $\mathcal{T}$.
3. La intersección de una colección finita de elementos de $\mathcal{T}$ también está en $\mathcal{T}$.

A la pareja $(X, \mathcal{T})$ se le llama un **Espacio topológico**. Los elementos de $\mathcal{T}$ se llaman conjuntos abiertos.

Ejemplo 2.1. Dado un conjunto X, se define la **topología trivial** de X, como $\mathcal{T} = \{X, \phi\}$.

Ejemplo 2.2. Dado un conjunto X, se define la **topología discreta** de X, como $\mathcal{T} = \mathcal{P}(X)$, es decir el conjunto de partes de X o colección de todos los subconjuntos de X.

Ejemplo 2.3. Dado el conjunto $X = \mathbb{R}$, se define la **topología usual** $\mathcal{T}$ de $\mathbb{R}$, como la colección de todos los conjuntos que son intervalos abiertos o uniones arbitrarias de ellos. De manera similar se define la topología usual $\mathcal{T}$ de $\mathbb{R}^2$ como la colección de todos los rectángulos abiertos o uniones arbitrarias de ellos.

Definición 2.2. Sea $(X, \mathcal{T})$ un espacio topológico. Un conjunto $C \subseteq X$ es un conjunto cerrado si su complemento es abierto, es decir, si $X - C = \{x \in X \ : \ x \notin C\}$ es abierto.

Pueden haber conjuntos abiertos y cerrados al mismo tiempo y conjuntos que no son ni abiertos ni cerrados. Los conjuntos Φ y X son ambos abiertos y cerrados.

Definición 2.3. Sea $(X, \mathcal{T})$ un espacio topológico. Una vecindad abierta de $x \in X$, es un abierto U tal que $x \in U$.

Definición 2.4. Dado un espacio topológico $(X, \mathcal{T})$. Una **base topológica** es un conjunto $\mathcal{B} \subseteq \mathcal{T}$ tal que todo abierto (no vació) $U \in \mathcal{T}$ se puede expresar como una unión de elementos de $\mathcal{B}$.

2.1.2. Espacio Métrico

La topología sobre $\mathbb{R}^n$ es definida completamente mediante la distancia euclídea entre puntos. Este método para definir una topología puede ser generalizado a cualquier espacio en donde una distancia es definida.

Definición 2.5. Un **Espacio Métrico** es un conjunto X con una función $d \ : \ X \times X \to \mathbb{R}$ que cumple lo siguiente:

1. $d(x, y) \geq 0$ y $d(x, y) = 0$ si y sólo si $x = y$.
2. $d(x, y) = d(y, x)$.

3. $d(x, y) + d(y, z) \geq d(x, z)$.

A la función d anterior se llama una **métrica** o **función distancia** sobre X y a al par (X, d) **espacio métrico**.

Usando la función distancia de un espacio métrico, una base para una topología sobre X puede ser definida como la colección de bolas abiertas $B(x, r) = \{y \in X \ : \ d(x, y) < r\}$ para todo $x \in X, r \in \mathbb{R}$. Una propiedad importante de los espacios métricos en la revisión de la teoría de variedades es la que sigue.

Definición 2.6. Una métrica d sobre un conjunto X se llama **completa** si cualquier sucesión de Cauchy converge en X. Una sucesión de Cauchy es una sucesión $x_1, x_2, \ldots \in X$ tales que para todo $\epsilon > 0$ existe un entero N talque $d(x_i, x_j) < \epsilon$ para todo $i, j > N$.

2.1.3. Continuidad

Como se mencionó inicialmente, la topología se desarrolló con el deseo de generalizar la noción de continuidad de mapeos o aplicaciones de espacios euclídeos. La generalización se hace como sigue.

Definición 2.7. Sean X y Y espacios topológicos. Un mapeo $f \ : \ X \to Y$ es **continuo** si para cada conjunto abierto $U \subset Y$, el conjunto $f^{-1}(U)$ es abierto en X.

Es fácil verificar que para un mapeo $f \ : \ \mathbb{R}^d \to \mathbb{R}^n$, la anterior definición es equivalente a la definición ϵ, δ estándar.

Definición 2.8. Sean X y Y espacios topológicos. Un mapeo $f \ : \ X \to Y$ es un **homeomorfismo** si es biyectivo y tanto f como f^{-1} son continuas. En este caso se dice que X y Y son **homeomórficos**.

Cuando X y Y son homeomórficos, hay una correspondencia biyectiva tanto entre puntos como entre conjuntos abiertos de X y Y. Por lo tanto, como espacios topológicos X y Y son indistinguibles, lo que significa que cualquier propiedad o teorema que sea cierto para el espacio X basado únicamente en la topología de X también es cierto para Y.

2.1.4. Algunas propiedades de Espacios Topológicos

Definición 2.9. Un espacio topológico X se dice que es **Hausdorff**, si para cualesquiera dos puntos distintos $x, y \in X$ existen conjuntos abiertos disjuntos U y V con $x \in U$ y $y \in V$.

Notar que cualquier espacio métrico es un espacio de Hausdorff. Dados dos puntos cualesquiera x, y en un espacio métrico X, se tiene que $d(x, y) > 0$. Luego las dos bolas abiertas $B(x, r)$ y $B(y, r)$, donde $r = \frac{1}{2}d(x, y)$, son conjuntos disjuntos abiertos que contienen a x y a y respectivamente. Sin embargo, no todos los espacios topológicos son de Hausdorff. Por ejemplo, al tomar cualquier conjunto X con más de un punto y dotarlo de la topología trivial, se tiene que Φ y X son los únicos conjuntos abiertos.

Definición 2.10. Sea X un espacio topológico. Una colección de subconjuntos abiertos $\mathcal{O}$ de X se dice que es un **cubrimiento abierto**, si $X = \bigcup_{U \in \mathcal{O}} U$. Un espacio topológico X se dice que es **compacto**, si para cualquier cubrimiento abierto $\mathcal{O}$ de X existe una subcolección finita de conjuntos de $\mathcal{O}$ que cubre a X.

En el teorema de Heine-Borel, (Rudin 1976), se dan criterios intuitivos para que un subconjunto de $\mathbb{R}^n$ sea compacto. Este teorema dice que cualquier subconjunto cerrado y acotado de $\mathbb{R}^n$ es compacto. De donde por ejemplo, una bola cerrada $\bar{B}(x, r)$ es compacta, como lo es la esfera unitaria $S^{n-1} = \{x \in \mathbb{R}^n \ : \ \|x\| = 1\}$. La esfera, como espacio euclídeo, es un ejemplo importante en lo que sigue, ya que es un ejemplo simple de espacio simétrico y además es parte integral de la representación medial de objetos que se usará en esta tesis.

Definición 2.11. Una **separación** de un espacio topológico X es un par de conjuntos disjuntos U, V tales que $X = U \cup V$. Si no existe ninguna separación de X se dice que es **conectado**.

2.2. Variedades Diferenciales

Las variedades diferenciables son espacios que localmente se comportan como espacios euclídeos. En la mayoría de ellas al igual que en los espacios topológicos, es natural hablar de continuidad. Las variedades diferenciales son un entorno natural para el cálculo. Nociones tales como diferenciación, integración, campos vectoriales y ecuaciones diferenciales tienen sentido sobre variedades diferenciables. Ahora se dará una revisión básica

de los resultados que se necesitarán más adelante. Para una visión más general de geometría diferencial pueden ver, Spivak 1999, Milnor 1997, Helgason 1978, Auslander and MacKenzie 1977.

2.2.1. Variedades Topológicas

Una variedad es un espacio topológico que es localmente equivalente a un espacio euclideano.

Definición 2.12. Una **variedad topológica** es un espacio topológico Hausdorff M con una base contable tal que para cada $p \in M$ existe una vecindad U de p que es homeomorfo a $\mathbb{R}^n$ para algún entero n. Es decir, existe un homeomorfismo $x : U \to \Theta \subseteq \mathbb{R}^n$, para un abierto Θ de $\mathbb{R}^n$.

En cada punto $p \in M$ la dimensión n de $\mathbb{R}^n$ en la definición anterior, es única. Si el entero n es el mismo para cualquier punto en M, entonces M se llama una variedad **n-dimensional**. El ejemplo más simple de una variedad es $\mathbb{R}^n$, ya que es trivialmente homeomorfo a sí mismo. De la misma forma, cualquier conjunto abierto de $\mathbb{R}^n$ también es una variedad.

2.2.2. Estructura Diferenciable sobre una Variedad

Lo que sigue en el desarrollo de la teoría de variedades es definir la noción de diferenciación de mapeos en variedades. La diferenciación de mapeos en espacios euclídeos es definida como una propiedad local. Aunque una variedad es localmente homeomórfica a un espacio euclídeo, se requiere de más estructuras para hacer posible la diferenciación. Primero, recordemos que una función sobre un espacio euclídeo $f : \mathbb{R}^n \to \mathbb{R}$ es **suave** o $\mathbf{C}^\infty$ si existen todas sus derivadas parciales. Un mapeo o aplicación de espacios euclídeos $f : \mathbb{R}^m \to \mathbb{R}^n$ se puede pensar como una n-tupla de funciones real-valuadas sobre $\mathbb{R}^m$, es decir, $f = (f^1, f^2, \ldots, f^n)$ y f es suave si cada una de las f^i lo es.

Dado dos vecindades U y V en una variedad M, se dice que dos homeomorfismos $x : U \to \mathbb{R}^n$ y $y : V \to \mathbb{R}^n$ están $\mathbf{C}^\infty$**-relacionados** si el mapa $x \circ y^{-1} : y(U \cap V) \to x(U \cap V)$ es $\mathbf{C}^\infty$.

A la pareja (x, U) se le llama un **entorno coordenado de p** (o **sistema de coorde-**

nadas locales alrededor de p) y se puede considerar como la asignación de un conjunto de coordenadas a los puntos en la vecindad U de p, ver figura 2.1. Es decir, a cualquier punto $p \in U$ le son asignadas las coordenadas $x^1(p), x^2(p), \ldots, x^n(p)$. Como se verá más adelante, los entornos coordenados son importantes para escribir expresiones locales para derivadas, vectores tangentes y métricas Riemannianas sobre una variedad. Una colección de entornos coordenados cuyo dominio cubre a M se le llama un **atlas**, es decir, $\mathcal{A} = \{(x_\alpha, U_\alpha) \ : \ \alpha \in I\}$ es un atlas si $M = \cup_{\alpha \in I} U_\alpha$.

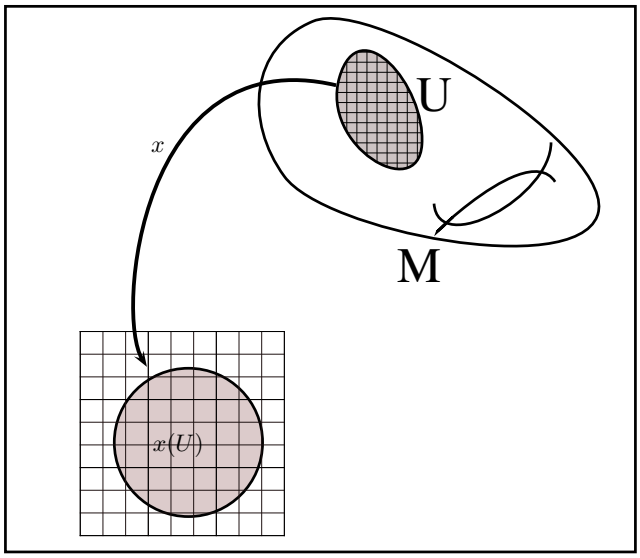


Figura 2.1: Gráfico de coordenadas locales en $\mathbb{R}^2$

Definición 2.13. Un atlas $\mathcal{A}$ sobre una variedad M se dice que es **maximal** si para cualquier otro atlas $\mathcal{A}'$ sobre M, cualquier entorno de coordenadas locales $(x, U) \in \mathcal{A}'$ también es miembro de $\mathcal{A}$, es decir, $\mathcal{A}$ contiene a $\mathcal{A}'$.

Definición 2.14. Una **estructura suave** sobre una variedad M es un atlas maximal $\mathcal{A}$ sobre M.

La variedad M en conjunto con dicho atlas se denomina una **variedad suave**.

Teorema 2.1. Dada una variedad M con un atlas $\mathcal{A}$, existe un único atlas $\mathcal{A}'$ tal que $\mathcal{A} \subset \mathcal{A}'$.

Ejemplo 2.4. Considere la esfera unitaria S^2 como un subconjunto de $\mathbb{R}^3$. El hemisferio superior $U = \{(x, y, z) \in S^2 \ : \ z > 0\}$ es una vecindad abierta de S^2. Ahora, se considera

el homeomorfismo

$$\phi \ : \ S^2 \longrightarrow \mathbb{R}^2$$
$$(x, y, z) \to \phi(x, y, z) = (x, y).$$

Este homeomorfismo origina un entorno de coordenadas locales (ϕ, U).

Entornos de coordenadas similares se pueden producir para el hemisferio inferior y para los hemisferios sobre las dimensiones x y y. Se puede verificar que dichos entornos están $\mathbf{C}^\infty$-relacionados y que cubren a S^2. Por lo tanto, estos atlas forman un atlas sobre S^2 y por el teorema anterior existe un único atlas maximal que contiene a estos atlas y hacen de S^2 una variedad suave.

Ahora, se considera la función $f \ : \ M \to \mathbb{R}$ sobre una variedad suave M. Esta función se dice que es una **función suave** si para cualquier entorno de coordenadas locales (x, U) sobre M, la función $f \circ x^{-1} \ : \ U \to \mathbb{R}$ es suave.

Más generalmente, un mapeo $f \ : \ M \to N$ de variedades suaves se dice que es un **mapeo suave**, si para cada entorno de coordenadas locales (x, U) sobre M y cada entorno de coordenadas locales (y, V) sobre N, el mapeo

$$y \circ f \circ x^{-1} \ : \ x(U) \subseteq \mathbb{R}^n \to y(V) \subseteq \mathbb{R}^{n'}$$

es una mapeo suave. Notar que el mapeo de variedades fue convertido localmente a un mapeo de espacios euclídeos, en donde la diferenciabilidad es fácilmente definida, ver figura 2.2.

Como en el caso de espacios topológicos, existe el deseo de saber cuando dos variedades suaves son equivalentes, lo que quiere decir que ellas son homeomorfas como espacios topológicos y también que tienen estructuras suaves equivalentes.

Teorema 2.2. Dadas dos variedades suaves M y N, una mapeo biyectivo $f \ : \ M \to N$ se llama un **difeomorfísmo** si tanto f como f^{-1} son mapeos suaves. En este caso se dice que M y N son **difeomorfas**.

2.2.3. Espacio Tangente

Dada una variedad $M \subset \mathbb{R}^d$, es posible asociar un subespacio lineal de $\mathbb{R}^d$ a cada punto $p \in M$, llamado el **espacio tangente** en p. El espacio tangente a M en p se denota

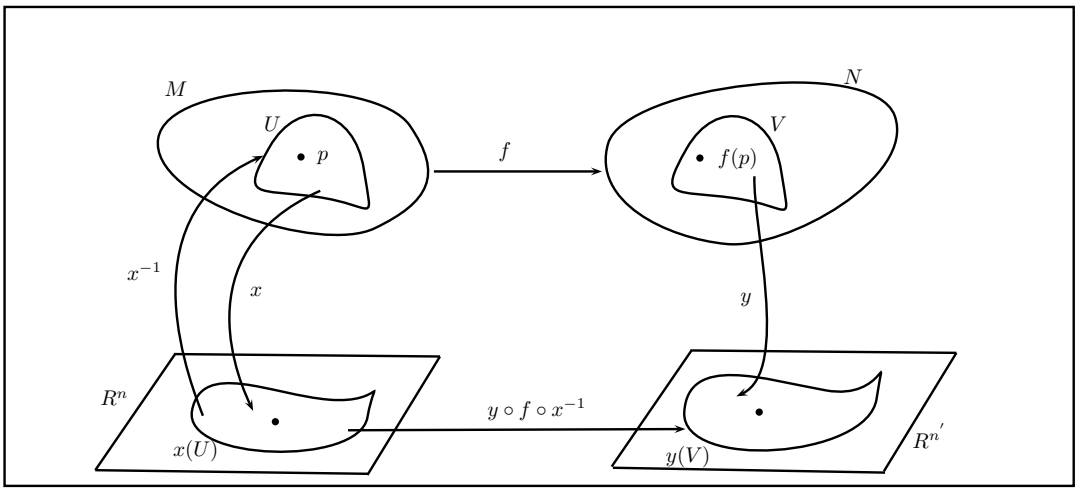


Figura 2.2: Mapeo diferenciable entre variedades

por T_pM y puede ser considerado intuitivamente como el subespacio lineal que mejor se aproxima a M en una vecindad del punto p. Los vectores en el espacio tangente se llaman vectores **vectores tangentes** en p, ver figura 2.3.

Los vectores tangentes se pueden considerar como derivadas direccionales. Considere una curva suave $\gamma \ : \ (-\epsilon, \epsilon) \to M$, con $\gamma(0) = p$. Entonces dada cualquier función suave $f \ : \ M \to \mathbb{R}$, la composición $f \circ \gamma \ : \ (-\epsilon, \epsilon) \to \mathbb{R}$, es una función suave y existe la siguiente derivada:

$$\frac{d}{dt}(f \circ \gamma)(0).$$

Esto conduce a una relación de equivalencia $\sim$ entre las curvas suaves que pasan por p en $t = 0$, es decir, $\mathcal{C}_p = \{\gamma \ : \ (-\epsilon, \epsilon) \to M \ : \ \epsilon_\gamma > 0, \gamma(0) = p, \gamma \text{ es diferenciable}\}$, a saber, si γ_1 y γ_2 son dos curvas suaves que pasan a través del punto p en $t = 0$, entonces

$$\gamma_1 \sim \gamma_2 \ , \ \text{ si para algún entorno de coordenadas } \ (x, U) = ((x^1, x^2, \ldots, x^n), U) \ \text{ de } p$$

se cumple que,

$$\frac{d}{dt}(f \circ \gamma_1)(0) = \frac{d}{dt}(f \circ \gamma_2)(0),$$

es decir, las curvas son equivalentes si los vectores tangentes en $\mathbb{R}^n$ de ambas curvas vistas en coordenadas locales coinciden, para cualquiera función suave $f \ : \ M \to \mathbb{R}$, ver figura 2.4. Notar que $f \circ \gamma_1(0) = f(\gamma_1(0)) = f(\gamma_2(0)) = f \circ \gamma_2(0) = p$. Ahora un vector tangente

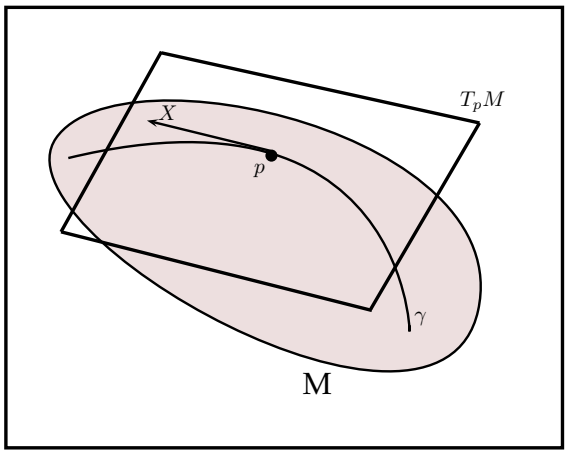


Figura 2.3: Espacio tangente a M en p

se define como una de estas clases de equivalencias de curvas.
Se puede mostrar, ver Auslander and MacKenzie 1977, que estas clases de equivalencia forman un espacio vectorial, a saber, el espacio tangente T_pM, el cual tiene la misma dimensión que M. Dado un sistema de coordenadas locales (x, U) que contiene a p, una base para el espacio tangente T_pM esta dada por los operadores derivadas parciales $\partial/\partial x^i$, las cuales son los vectores tangentes asociados con las curvas coordenadas de x.

Ejemplo 2.5. Nuevamente, se considera la esfera S^2 como un subconjunto de $\mathbb{R}^3$. El espacio tangente en un punto $p \in S^2$ es el conjunto de todos los vectores en $\mathbb{R}^3$ que son perpendiculares a p, es decir, $T_pS^2 = \{v \in \mathbb{R}^3 \ : \ \langle v, p\rangle = 0\}$. Este espacio es en realidad un espacio vectorial bi-dimensional, el cual es el espacio de todos los vectores tangentes en el punto p para curvas suaves a lo largo de la esfera y que pasan a través del punto p.

Un **campo vectorial** sobre una variedad M es una función que asigna de manera suave a cada punto $p \in M$ un vector tangente $X_p \in T_pM$. Este mapeo es suave en el sentido de que las componentes de los vectores pueden ser escritas como funciones suaves en cualquier sistema de coordenadas locales. Es decir, un campo vectorial es una aplicación $X \ : \ M \to TM$, tal que, $\pi \circ X = Id_M$, en donde, $\pi \ : \ TM \to M$, $X_p \mapsto \pi(X_p) = p$, es la proyección canónica y $TM = \cup_{p\in M} T_pM$: es la variedad tangente de M.

Un campo vectorial se puede ver como un operador $X : C^\infty(M) \to C^\infty(M)$, el cual mapea una función suave $f \in C^\infty(M)$ a una función suave $Xf \ : \ M \to M$ tal que,

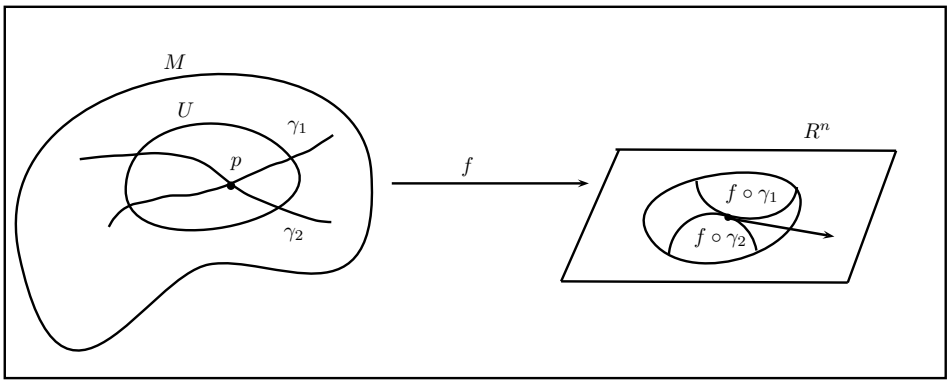


Figura 2.4: Vectores tangentes a M en P

$p \to X_p f$, en otras palabras, la derivada direccional es aplicada en cada punto sobre M, con $C^\infty(M) = \{f \ : \ M \to \mathbb{R} \ : \ f$ es diferenciable o suave$\}$.

Para dos variedades M y N, un mapeo suave $\phi \ : \ M \to N$ induce un mapeo lineal de los espacios tangentes

$$\phi_\star \ : \ T_pM \to T_{\phi(p)}M,$$

dicho mapeo es llamado la **diferencial** de ϕ en p. Esta diferencial esta dada por $\phi_\star(X_p)f = X_p(f \circ \phi)$, para cualquier $X_p \in T_pM$ y para cualquier función suave $f \in C^\infty(M)$. Un mapeo suave de variedades no siempre induce a un mapeo de campos vectoriales (por ejemplo, cuando el mapa no es sobre). Sin embargo, un concepto relacionado esta dado en la definición que sigue.

Definición 2.15. Dado un mapeo de variedades suaves $\phi \ : \ M \to N$, se dice que un campo vectorial X sobre M y un campo vectorial Y sobre N están $\phi-$**relacionados**, si $\phi_\star(X(p)) = Y(q)$ es cierto para cada $q \in N$ y para cada $p \in \phi^{-1}(q)$.

2.3. Geometría Riemanniana

Como se mencionó inicialmente, la idea de distancias sobre una variedad es importante en la definición de estadísticas sobre variedades. La noción de distancia sobre una variedad cae en el ámbito de la geometría Riemanniana, la cual se relaciona con la teoría de variedades suaves. En esta sección se revisan algunos conceptos necesarios para lo que

sigue. Para ver más sobre geometría Riemanniana pueden revisar Boothby 1986, Spivak 1999 y Lee 1997.

Recordemos la definición de longitud de un curva suave sobre un espacio euclideano. Sea $\gamma : [a,b] \to \mathbb{R}^d$ un segmento de curva suave. Entonces en cualquier punto $t_0 \in [a,b]$, la derivada de la curva $\gamma'(t_0)$ da la velocidad de la curva al tiempo t_0. La longitud del segmento de curva γ está dada por la integral de la velocidad de la curva, es decir,

$$L(\gamma) = \int_a^b \|\gamma'(t)\| dt.$$

Esta definición de longitud requiere la habilidad de tomar la norma de los vectores tangentes. En el entorno de variedades, esto es tratado por la definición de una métrica Riemanniana.

2.3.1. Métrica Riemanniana

Definición 2.16. Una **métrica Riemanniana** sobre una variedad M es una función que asigna suavemente a cada punto $p \in M$ un producto interno $\langle . \ , \ . \rangle$ sobre el espacio tangente T_pM. Una **variedad Riemanniana** es una variedad suave equipada con una métrica Riemanniana.

Ahora la norma de un vector tangente $v \in T_pM$ se define como $\|v\| = \langle v \ , \ v \rangle^{\frac{1}{2}}$. Dada las coordenadas locales $x^1, x^2, \ldots, x^n$ sobre una vecindad de p, los vectores coordenados $v^i = \partial/\partial x^i$ en p, forman una base para el espacio tangente T_pM. La métrica Riemanniana se puede expresar en esta base como una matriz $n \times n$ denotada por g, llamada el tensor métrico, cuyas entradas están dadas por

$$g_{ij} = \langle v^i \ , \ v^j \rangle.$$

Las g_{ij} son funciones suaves de las coordenadas $x^1, x^2, \ldots, x^n$.

Dado un segmento de curva suave $\gamma : [a,b] \to M$, la longitud de γ se puede definir de forma similar al caso euclideano como sigue

$$L(\gamma) = \int_a^b \|\gamma'(t)\| dt, \tag{2.1}$$

en donde ahora el vector tangente $\gamma'(t)$ es un vector sobre $T_{\gamma(t)}M$ y la norma esta dada

por la métrica Riemanniana en $\gamma(t)$.

Dada una variedad M y una variedad N con métrica Riemanniana $\langle .\ ,\ . \rangle$, un mapeo $\phi\ :\ M \to N$ induce una métrica $\phi_\star\langle .\ ,\ . \rangle$ sobre M definida por:

$$\phi_\star\langle X_p\ ,\ Y_p\rangle = \langle \phi_\star(X_p)\ ,\ \phi_\star(X_p)\rangle.$$

Esta métrica se llama la métrica **pull-back** inducida por ϕ, ya que ésta mapea la métrica en la dirección opuesta del mapa ϕ.

2.3.2. Geodésica

Sobre espacios euclídeos la trayectoria más corta entre dos puntos es una línea recta y la distancia entre los puntos es medida como la longitud de ese segmento de línea recta. Esta noción de trayectoria más corta puede ser extendida a variedades Riemanniana considerando el problema de hallar el segmento de curva suave más corta entre dos puntos sobre la variedad. Si $\gamma\ :\ [a,b] \to M$ es una curva suave sobre la variedad Riemanniana M con puntos finales $\gamma(a) = x$ y $\gamma(b) = y$, una **variación** de γ que **mantiene los puntos finales fijos** es una familia α de curvas suaves

$$\alpha\ :\ (-\epsilon, \epsilon) \times [a,b] \to M,$$

tal que

1. $\alpha(0,t) = \gamma(t)$,
2. $\tilde{\alpha}(s_0)\ :\ t \mapsto \alpha(s_0,t)$, es un segmento de curva suave para $s_0 \in (-\epsilon,\epsilon)$,
3. $\alpha(s,a) = x$ y $\alpha(s,b) = y$ para todo $s \in (-\epsilon,\epsilon)$.

Ahora la trayectoria suave más corta entre los puntos $x, y \in M$ puede ser vista como hallar un punto crítico para la función longitud de la ecuación 2.1, donde la longitud de $\tilde{\alpha}$ es considerada como una función de s. La trayectoria $\gamma = \tilde{\alpha}(0)$ es una trayectoria crítica para L si

$$\frac{dL(\tilde{\alpha}(s))}{ds} = 0.$$

Resulta más fácil trabajar con la trayectoria crítica del **funcional energía**, el cual esta dado por

$$E(\gamma) = \int_a^b \|\gamma'(t)\|^2 dt.$$

Se puede demostrar, ver Spivak 1999, que una trayectoria crítica para E también es crítica para L. Recíprocamente, una trayectoria crítica para L, una vez parametrizada de forma proporcional a la longitud de arco, es una trayectoria crítica para E. Luego, al asumir curvas que están parametrizadas proporcional a la longitud de arco, no hay diferencia entre curvas con longitud mínima y aquellas con mínima energía. Una trayectoria crítica del funcional E se llama una **geodésica**.

Dado un gráfico (x, U), una curva geodésica $\gamma \subset U$ se puede escribir en coordenadas locales como $\gamma(t) = (\gamma^1(t), \gamma^2(t), \ldots, \gamma^n(t))$. Usando algún sistema de coordenadas locales, γ cumple la siguiente ecuación diferencial, ver Spivak 1999:

$$\frac{d^2\gamma^k}{dt^2} = -\sum_{i,j=1}^{n} \Gamma_{ij}^k(\gamma(t)) \frac{d\gamma^i}{dt} \frac{d\gamma^j}{dt}. \tag{2.2}$$

Los símbolos Γ_{ij}^k se llaman los **símbolos de Christoffel** y se definen como sigue:

$$\Gamma_{ij}^k = \frac{1}{2} \sum_{l=1}^{n} g^{kl} \left(\frac{\partial g_{jl}}{\partial x^i} + \frac{\partial g_{il}}{\partial x^j} - \frac{\partial g_{ij}}{\partial x^l} \right),$$

en donde, g^{ij} denota las entradas de la matriz inversa g^{-1} de la métrica Riemanniana.

Ejemplo 2.6. Sobre el espacio euclídeo $\mathbb{R}^n$, la métrica Riemanniana está dada por la matriz identidad en cada punto $p \in \mathbb{R}^n$. Debido a que la métrica es constante, los símbolos de Christoffel son cero, por lo tanto la ecuación geodésica en 2.2 se reduce a

$$\frac{d^2\gamma^k}{dt^2} = 0.$$

Las únicas soluciones de esta ecuación son líneas rectas, por lo tanto las geodésicas sobre $\mathbb{R}^n$ deben ser líneas rectas.

Dados dos puntos sobre una variedad Riemanniana, no hay garantía de que exista una geodésica entre ellos. También pueden existir múltiples geodésicas uniendo los puntos, es decir, no existe garantía de que la geodésica sea única. Además, una geodésica no tiene que ser un mínimo global de la longitud funcional, es decir, pueden existir geodésicas de diferentes longitudes entre los mismos dos puntos.

Ejemplo 2.7. Considere el plano con el origen removido, $\mathbb{R}^2 - \{0\}$, con la misma métrica como en $\mathbb{R}^2$. La geodésicas aun están dadas por líneas rectas. No existe una geodésica entre los puntos $(1, 0)$ y $(-1, 0)$.

Ejemplo 2.8. Las geodésicas sobre la esfera S^2 están dadas por los círculos mayores, es decir, círculos sobre la esfera con diámetro máximo. Existen un número infinito de geodésicas de igual longitud entre los polos norte y sur, es decir, los meridianos. Además, dados dos puntos cualesquiera sobre S^2 que no son puntos antipodales, existe un único círculo mayor entre ellos. El círculo mayor es separado por dos segmentos geodésicos entre los dos puntos. Uno de estos segmentos geodésicos es más grande que el otro.

La idea de mínimo global de la longitud, nos lleva a la definición de una **distancia métrica** $d : M \times M \to \mathbb{R}$ (no se puede confundir con la métrica Riemanniana). Esta distancia métrica se define como sigue:

$$d(p, q) = \text{Inf}\{L(\gamma) \ : \ \gamma \text{ es una curva suave entre } p \text{ y } q\}.$$

Si existe una geodésica entre los puntos p y q que cumple esta distancia, es decir, si $L(\gamma) = d(p, q)$, entonces a γ se le llama una **geodésica minimal**. Las geodésicas minimales existen bajo ciertas condiciones.

Definición 2.17. Una variedad Riemanniana M se dice que es **completa** si cualquier segmento geodésico $\gamma : [a, b] \to M$ se puede extender a una geodésica desde los reales $\mathbb{R}$ a M.

Teorema 2.3. (**Hopf-Rinow**). Si M es una variedad Riemanniana completa y conectada, entonces la distancia métrica $d(. \ , \ .)$ inducida sobre M es completa. Además, entre cualesquiera dos puntos sobre M existe una geodésica minimal.

Ejemplo 2.9. Tanto el espacio euclideano $\mathbb{R}^n$ y la esfera S^2 son completa. Una línea recta sobre $\mathbb{R}^n$ puede extenderse en ambas direcciones indefinidamente. También, un círculo mayor sobre S^2 se extiende indefinidamente en ambas direcciones (aunque estas se traslapan sobre sí mismo). Como se garantiza por el teorema de Hopf-Ronow, existe una geodésica minimal entre cualesquiera dos puntos sobre $\mathbb{R}^n$, es decir, el único segmento de línea recta entre los puntos. También entre cualesquiera dos puntos sobre la esfera existe una única geodésica minimal, a saber, el más corto de los dos segmentos del círculo mayor entre los dos puntos. De hecho, para puntos antipodales sobre S^2 la geodésica minimal no es única.

Dadas las condiciones iniciales $\gamma(0) = p$ y $\gamma'(0) = v$, la teoría de ecuaciones diferenciales parciales de segundo orden garantiza la existencia de una única solución a la ecuación

de la definición de γ en 2.2, al menos localmente. Así, existe una única geodésica γ con $\gamma(0) = p$ y $\gamma'(0) = v$ definida en algún intervalo $(-\epsilon, \epsilon)$. Cuando la geodésica γ existe en el intervalo $[0, 1]$, el **mapa exponencial Riemanniano** en el punto p, ver figura 2.2, se define como sigue:

$$\begin{aligned} \mathrm{Exp}_p \ : \ & T_pM \to M \\ & v \longrightarrow \mathrm{Exp}_p(v) = \gamma(1). \end{aligned}$$

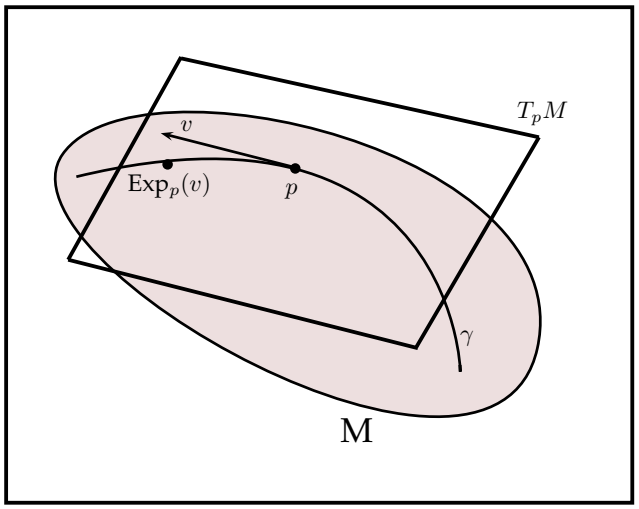


Figura 2.5: Mapa exponencial Riemanniano

Si M es una variedad completa, el mapa exponencial Riemanniano esta definido para todos los vectores $v \in T_pM$.

Teorema 2.4. Dada una variedad Riemanniana M y un punto $p \in M$, el mapa Exp_pM es un difeomorfismo sobre alguna vecindad $U \subseteq T_pM$ que contiene al cero.

Este teorema implica que el Exp_p tiene una inversa definida por lo menos sobre una vecindad $\mathrm{Exp}_p(U)$ de p, donde U es lo mismo que en el teorema 2.4. A esta inversa se le llama el **mapa logarítmico Riemanniano** y es definido por

$$\begin{aligned} \mathrm{Log}_p(U) \ : \ & \mathrm{Exp}_p(U) \subseteq M \to T_pM \\ & X \longrightarrow \mathrm{Log}_p(X) = v. \end{aligned}$$

Definición 2.18. Una **isometría** es un difeomorfismo $\phi : M \to N$ de variedades Riemannianas que preserva la métrica Riemanniana. Es decir, si $\langle . \ , \ . \rangle_M$ y $\langle . \ , \ . \rangle_N$ son las métricas para M y N, respectivamente, entonces $\phi_\star \langle . \ , \ . \rangle_N = \langle . \ , \ . \rangle_M$.

De la definición anterior se sigue que una isometría preserva longitudes de curvas. Es decir que, si c es una curva suave sobre M, entonces la curva $\phi \circ c$ es una curva de la misma longitud sobre N. Además, la imagen de una geodésica bajo una isometría es nuevamente una geodésica.

2.4. Grupos de Lie

El conjunto de todas las posibles traslaciones del espacio euclídeo $\mathbb{R}^n$ es nuevamente el espacio $\mathbb{R}^n$. Un punto $p \in \mathbb{R}^n$ es transformado por el vector $v \in \mathbb{R}^n$, mediante la suma de los vectores $p + v$. Esta transformación tiene una única transformación inversa, llamada traslación por el vector negativo, -v. La operación de traslación es un mapeo suave del espacio $\mathbb{R}^n$. La composición de las dos traslación (es decir, suma en $\mathbb{R}^n$) y una traslación invertida (es decir, el negativo en $\mathbb{R}^n$) también es un mapeo suave. Un conjunto de transformaciones con estas propiedades, es decir, una variedad suave con operaciones de grupo suaves, se conoce como un grupo de Lie. Muchas otras transformaciones de interés de espacios euclídeos también son grupos de Lie, incluyendo las rotaciones, reflexiones y magnificaciones. Sin embargo, los grupos de Lie aparecen más generalmente como transformaciones suaves de variedades. Ahora se hace un breve introducción sobre grupos de Lie. Más acerca del tema se puede encontrar en: Boothby 1986, Duistermaat and Kolk 2000, Hall 2003, Helgason 1978, Kawakubo 1991 y Spivak 1999, entre otros. Se asume que el lector conoce las base de teoría de grupos de Lie, puede ver más a cerca de esto en Herstein 1975.

Definición 2.19. Un **grupo** es un conjunto G con una operación arbitraria, denotada aquí por $\star$, tal que:

1. $(x \star y) \star z = x \star (y \star z)$, para todo $x, y, z \in G$,
2. Existe un elemento **identidad**, $e \in G$, que cumple $x \star e = e \star x = x$ para todo $x \in G$,
3. Cada $x \in G$ tiene un **inverso**, $x^{-1} \in G$, que cumple $x \star x^{-1} = x^{-1} \star x = e$.

Como se mencionó al inicio de esta sección, un grupo de Lie le da la estructura de variedad suave a un grupo.

Definición 2.20. Un **grupo de Lie** G, es una variedad suave que también forma un grupo, donde las dos operaciones de grupo multiplicación e inversa son mapeos suaves de variedades.

Es decir,

$$\begin{aligned} Multiplicacion \ : \ & G \times G \to G \\ & (x, y) \longmapsto x \star y. \end{aligned}$$

y

$$\begin{aligned} Inversa \ : \ & G \to G \\ & x \longmapsto x^{-1}. \end{aligned}$$

son mapeos suaves de variedades.

Ejemplo 2.10. El espacio de todas las matrices no-singulares $n \times n$ forma un grupo de Lie llamado **grupo de Lie general** y se denota por $GL(n)$. La operación de grupo es la multiplicación de matrices y $GL(n)$ puede ser dotado de una estructura de variedad suave como un subconjunto abierto de $\mathbb{R}^{n^2}$. Las ecuaciones para la multiplicación de matrices e inversa son operaciones suaves sobre las entradas de las matrices. Por lo tanto, $GL(n)$ satisface los requerimientos de un grupo de Lie según la definición (2.19). Un **grupo de matrices** es cualquier subgrupo cerrado de $GL(n)$.

Los grupos de matrices heredan la estructura de suavidad del grupo $GL(n)$ como un subconjunto de $\mathbb{R}^{n^2}$ y por lo tanto también son grupos de Lie. Para ver más acerca de la teoría de grupo de matrices, consultar a Curtis 1984 y Hall 2003.

Ejemplo 2.11. Las matrices de rotación $n \times n$, forman un subgrupo cerrado de matrices del grupo $GL(n)$ y por lo tanto forman un grupo de Lie. Este grupo se llama el **grupo ortogonal especial** y se define como

$$SO(n) = \{R \in GL(n) \ : \ R^T R = I, \ \text{y} \ \det(R) = 1\}.$$

Este espacio es un subconjunto acotado y cerrado de $\mathbb{R}^{n^2}$, por lo tanto es compacto por el teorema de Heine-Borel.

Dado un punto y sobre un grupo de Lie G, es posible definir los siguientes dos difeomorfismos:

$$Multiplicacion\ por\ la\ Izquierda\ G \to G\ , \quad x \mapsto yx$$

y

$$Multiplicacion\ por\ la\ Derecha\ G \to G\ , \quad x \mapsto xy.$$

Un campo vectorial X sobre un grupo de Lie G, se llama **invariante izquierda** si dicho campo es invariante bajo la multiplicación por la izquierda, es decir, $L_{y\star}X = X$ para cualquier $y \in G$. Los campos vectoriales **invariante derecha** son definidos de manera similar. Un campo vectorial invariante izquierda (o invariante a derecha) es únicamente definido mediante sus valores sobre el espacio tangente en la identidad, es en T_eG.

Recordemos que los campos vectoriales sobre G pueden ser vistos como operadores sobre el espacio de funciones suaves, $C^\infty(G) = \{f \ : \ G \to \mathbb{R} \ : \ f \text{ es suave o diferenciable}\}$. Por lo tanto dos campos vectoriales X y Y pueden ser compuestos para formar otro operador XY sobre $C^\infty(G)$. Sin embargo, el operador XY no es necesariamente un campo vectorial. Sin embargo, el operador $XY - YX$ si es un campo vectorial sobre G. Esto nos lleva a la definición del **corchete de Lie** de los campos vectoriales X y Y sobre G, definido como

$$[X, Y] = XY - YX. \tag{2.3}$$

Definición 2.21. Una **álgebra de Lie** es un espacio vectorial V equipado con un producto bilineal $[.\ ,\ .] \ : \ V \times V \to V$, llamado un **corchete de Lie**, que cumple

1. $[X, Y] = -[Y, X]$,
2. $[[X, Y], Z] + [[Y, Z], X] + [[Z, X], Y] = 0$, para todo $X, Y, Z \in V$.

El espacio tangente de un grupo de Lie G, que se denota por $\mathfrak{g}$, forma una álgebra de Lie. El corchete de Lie sobre $\mathfrak{g}$ es inducido mediante el corchete de Lie sobre el correspondiente campo vectorial invariante a izquierda. Si X, Y son dos vectores en $\mathfrak{g}$, entonces sean $\tilde{X}, \tilde{Y}$ los únicos campos vectoriales invariante a izquierda correspondientes sobre G, entonces el corchete de Lie sobre $\mathfrak{g}$ está dado por

$$[X, Y] = [\tilde{X}, \tilde{Y}](e).$$

El corchete de Lie proporciona una prueba para saber si el grupo de Lie G es conmutativo. Un grupo de Lie G es conmutativo si y sólo si el corchete de Lie sobre la correspondientes álgebra de Lie $\mathfrak{g}$ es cero, es decir, si $[X, Y] = 0$ para todo $X, Y \in \mathfrak{g}$.

Ejemplo 2.12. El álgebra de Lie para el espacio euclídeo $\mathbb{R}^n$ es nuevamente $\mathbb{R}^n$. El corchete de Lie es cero, es decir, $[X, Y] = 0$ para todo $X, Y \in \mathbb{R}^n$. En realidad el corchete de Lie para el álgebra de Lie de cualquier grupo de Lie conmutativo es siempre cero.

Ejemplo 2.13. El álgebra de Lie para $GL(n)$ es $\mathfrak{gl}(\mathfrak{n})$, el espacio de todas las matrices reales $n \times n$. La operación corchete de Lie para $X, Y \in \mathfrak{gl}(n)$, está dada por:

$$[X, Y] = XY - YX.$$

Aquí el producto XY denota la multiplicación real de matrices, el cual resulta ser lo mismo que la composición de los operadores de campos vectoriales, (comparar con (2.3)). Todas las álgebras de Lie correspondiente a grupos de matrices son sub-álgebras de $\mathfrak{gl}(n)$.

Ejemplo 2.14. El álgebra de Lie para el grupo de rotaciones $SO(n)$ es $\mathfrak{so}(n)$, el espacio de matrices cuasi-simétricas. Una matriz es cuasi-simétrica si $A = -A^T$.

Teorema 2.5. Un producto directo $G_1 \times G_2 \times \ldots \times G_n$ de grupos de Lie también es grupo de Lie.

2.4.1. Mapa Exponencial y Logarítmico de Grupos de Lie

Definición 2.22. Un mapeo de grupos de Lie $\phi : G_1 \to G_2$ se llama un homeomorfismo de grupos de Lie, si es un mapeo suave y un homeomorfismo de grupos, es decir, $\phi(e_1) = e_2$ cuando e_1, e_2 son los respectivos elementos identidad de G_1 y G_2 y $\phi(gh) = \phi(g)\phi(h)$, para todo $g, h \in G_1$.

La imagen de un homomorfísmo de grupos de Lie $h : \mathbb{R} \to G$, se llama un **subgrupo uni-paramétrico**. Un subgrupo uni-paramétrico es al mismo tiempo una curva suave y un subgrupo de G. Esto no significa, sin embargo, que cualquier subgrupo uni-paramétrico es un subgrupo de Lie de G (puede fallar el hecho de ser una sub-variedad inmersa de G, lo cual es requisito para ser un subgrupo de Lie de G). Existe una correspondencia biyectiva entre el álgebra de Lie y los subgrupos uni-paramétricos.

Teorema 2.6. Sea $\mathfrak{g}$ el álgebra de Lie de un grupo de Lie G. Dado cualquier vector $X \in \mathfrak{g}$, existe un único homeomorfismo de grupos de Lie $h_X : \mathbb{R} \to G$, tal que $h'_X(0) = X$.

El **mapa exponencial de grupos de Lie**, $\exp : \mathfrak{g} \to G$ (no confundir con el mapa exponencial Riemanniano) se define como sigue

$$\exp(X) = h_X(1).$$

Ejemplo 2.15. Para el grupo de Lie $\mathbb{R}^n$, el único homeomorfismo de grupos de Lie, $h_X : \mathbb{R} \to \mathbb{R}^n$, de acuerdo al teorema (6), está dado por: $h_X(t) = tX$. Por lo tanto los subgrupos uni-paramétricos están dados por líneas rectas a través del origen. El mapa exponencial de grupos de Lie es la identidad. En este caso el mapa exponencial grupo de Lie es el mismo que el mapa exponencial Riemanniano en el origen. Pero esto, no siempre es el caso.

Para el grupo de matrices, el mapa exponencial de grupo de Lie de una matriz $X \in \mathfrak{gl}(n)$ se calcula mediante la fórmula

$$\exp(X) = \sum_{k=0}^{\infty} \frac{1}{k!} X^k. \tag{2.4}$$

Dicha serie converge absolutamente para todo $X \in \mathfrak{gl}(n)$.

Ejemplo 2.16. Para el grupo de Lie de rotaciones 3D, $SO(3)$, el mapa matriz exponencial toma una forma simple. Para una matriz $X \in \mathfrak{so}(3)$, la siguiente identidad es cierta:

$$X^3 = -\theta X \ , \quad \text{donde} \quad \theta = \sqrt{\frac{1}{2}\text{tr}(X^T X)}.$$

Sustituyendo esta identidad dentro de la serie infinita (2.4), el mapa exponencial para $\mathfrak{so}(3)$ se puede reducir a

$$\exp(X) = \begin{cases} I & \text{si} \quad \theta = 0 \\ I + \frac{\sin\theta}{\theta} X + \frac{1-\cos\theta}{\theta^2} X^2 & \text{si} \quad \theta \in (0, \pi). \end{cases}$$

El mapa logarítmico de grupos de Lie para una matriz rotación $R \in SO(3)$, está dado por

$$\log(R) = \begin{cases} I & \text{si} \quad \theta = 0 \\ \frac{\theta}{2\sin\theta}(R - R^T) & \text{si} \quad |\theta| \in (0, \pi), \end{cases}$$

donde, $\text{tr}(R) = 2\cos\theta + 1$.

El mapa exponencial para rotaciones 3D tiene un significado intuitivo. Cualquier vector $X \in \mathfrak{so}(3)$, es decir, una matriz cuasi-simétrica, puede ser escrita en la forma siguiente

$$X = \begin{pmatrix} 0 & -z & y \\ z & 0 & -x \\ -y & x & 0 \end{pmatrix}.$$

Si $v = (x, y, z) \in \mathbb{R}^3$, entonces la matriz rotación dada por el mapa exponencial $\exp(X)$ es una rotación 3D mediante un ángulo de $\theta = \|v\|$ alrededor del eje unitario $v/\|v\|$.

2.4.2. Métricas Bi-Invariantes

Definición 2.23. Una métrica Riemanniana $\langle . \ , \ . \rangle$ sobre un grupo de Lie G, se dice que es una **métrica bi-invariante** si es invariante tanto bajo multiplicación a izquierda como a derecha, es decir, $R_g^\star \langle . \ , \ . \rangle = L_g^\star \langle . \ , \ . \rangle = \langle . \ , \ . \rangle$ para todo $g \in G$.

Teorema 2.7. Para un grupo de Lie G con una métrica bi-invariante el mapa exponencial de grupo de Lie coincide con el mapa exponencial Riemanniano en la identidad, es decir, para cualquier vector tangente $X \in \mathfrak{g}$

$$\exp(X) = \mathrm{Exp}_e(X).$$

Usando la invarianza a izquierda de la métrica Riemanniana, cualquier geodésica en un punto $g \in G$ se puede escribir como la multiplicación a izquierda de una geodésica en la identidad. Es decir, la geodésica γ con condición inicial $\gamma(0) = g$ y $\gamma^{'}(0) = L_{g\star}(X)$ esta dada por

$$\gamma(t) = g\exp(tX).$$

Teorema 2.8. Un grupo de Lie compacto G tiene una única métrica bi-invariante.

2.5. Espacios Simétricos

Un espacio simétrico Riemanniano es una variedad conectada M tal que en cada punto el mapeo que regresa geodésicas a través de ese punto es una isometría. Para un tratamiento más detallado de espacios simétricos, ver Helgason 1978, Boothby 1986. Algunos ejemplos comunes de espacios simétricos son los espacios euclideanos $\mathbb{R}^n$, esferas S^n y espacios hiperbólicos H^n. Los espacios simétricos y los métodos para calcular geodésicas y distancias sobre ellos, aparecen de forma natural a partir de ciertas acciones de grupos de Lie sobre variedades.

Antes de definir lo que son los espacios simétricos, es necesario dar algunas definiciones preliminares acerca de mapeos de conjuntos. Sea X y ϕ cualquier mapeo de X en sí mismo. Un punto $x \in X$ se llama un **punto fijo** de ϕ si $\phi(x) = x$. El mapeo ϕ se llama **involutivo** si ϕ no es el mapeo identidad pero su cuadrado si lo es, es decir, $\phi \circ \phi =$id.

Definición 2.24. Un **espacio simétrico** es una variedad Riemanniana conectada M tal que en cada punto $p \in M$ existe una isometría involutiva $\phi_p \ : \ M \to M$ que tiene a p como un punto fijo aislado.

El término **aislado** significa que existe una vecindad U de p tal que p es el único punto en U que es un punto fijo de ϕ_p. Esta definición es algo alusiva a lo difícil que es obtener un sentido intuitivo para las clases de variedades que son espacios simétricos. Afortunadamente, esta definición es suficiente para implicar algunas propiedades muy buenas de espacios simétricos.

El siguiente teorema, ver Boothby 1986, muestra que la isometría involutiva ϕ_p de la definición 2.21, es más fácilmente vista como el mapeo que regresa geodésicas a través del punto p.

Teorema 2.9. Un espacio simétrico Riemanniano es completo, y si ϕ_p es una isometría involutiva de M, entonces ϕ_p es una reflexión del espacio tangente T_pM, es decir, $\phi_p(X) = -X$, y ϕ_p regresa geodésicas a través de p, es decir, $\phi_p(\mathrm{Exp}(X)) = \mathrm{Exp}_p(-X)$ para todo $X \in T_pM$ tal que dicha geodésica exista.

Debido a que los espacios simétricos aparecen naturalmente a partir de ciertos grupos de Lie de transformaciones de una variedad M. Ahora se darán algunas definiciones básicos acerca de acciones de grupos de Lie.

2.5.1. Acciones de grupos de Lie

Definición 2.25. Dada una variedad suave M y un grupo de Lie G, una **acción de grupo suave** de G sobre M, es una mapeo suave $G \times M \to M$, definido como $(g, p) \mapsto g.p$, tales que para todo $g, h \in G$ y todo $p \in M$ se cumple que:

1. $e.p = p$,
2. $(gh).p = (g.(h.p))$.

La acción de grupo se podría pensar como una transformación de la variedad M, de la misma forma que las matrices son transformaciones del espacio Euclideano.

La **órbita** de una punto $p \in M$ se define como: $G(p) = \{g.p \ : \ g \in G\}$. En el caso de que M tenga una sola órbita, entonces a M se le llama un **espacio homogéneo** y en este caso se dice que la acción de grupo es **transitiva**. El **sub-grupo de isotropía** de p se define como: $G_p = \{g \in G \ : \ g.p = p\}$, es decir, G_p-es el subgrupo de G que deja fijo al punto p.

Sea H un subgrupo de Lie cerrado del grupo de Lie G. La **cerradura izquierda** de un elemento $g \in G$ se define como $gH = \{gh \ : \ h \in H\}$. El espacio de todas de tales cerraduras se denota por G/H y es una variedad suave. Existe una biyección natural $G(p) \cong G/G_p$, dada por el mapeo $g.p \mapsto gG_p$, es decir,

$$\begin{aligned} G(p) &\to G/G_p \\ g.p &\longmapsto gG_p \end{aligned}$$

Ahora, sea M un espacio simétrico y se elige un punto base arbitrario $p \in M$. Siempre se puede escribir a M como un espacio homogéneo $M = G/G_p$, en donde G es un grupo conectado de isometrías de M y el subgrupo de isotropía G_p es compacto. El hecho de que G es un grupo de isometrías significa que, $d(p,q) = d(g.p, g.q)$, para todo $p, q \in M$ y $g \in G$.

Un elemento $g \in G$ induce una mapeo suave $\phi_g \ : \ M \to M$ vía la acción de grupo, definido como, $\phi_g(p) = g.p$. También, este mapeo tiene una inversa suave, a saber $\phi_{g^{-1}}$. Por lo tanto, ϕ_g-es un difeomorfismo.

Definición 2.26. Dada una acción del grupo de Lie G sobre una variedad M, una métrica Riemanniana **G-invariante** $\langle . \ , \ . \rangle$ sobre M es una métrica tal que el mapeo ϕ_g es una isometría para toda $g \in G$, es decir que, $\phi_g^\star \langle . \ , \ . \rangle$.

Ejemplo 2.17. La métrica euclídea estándar sobre $\mathbb{R}^n$ es invariante bajo la acción de grupo $SO(n)$. En otras palabras, una rotación del espacio euclídeo es una isometría. La acción de $\mathbb{R}^n$ sobre sí mismo mediante traslaciones, es otro ejemplo de un grupo de isometrías. Estos dos grupos pueden ser combinados para formar el **grupo especial euclideano**, denotado por $SE(n) = SO(n) \ltimes \mathbb{R}^n$. El producto semi-directo $\ltimes$ significa que $SE(n)$ como conjunto es el producto directo de $SO(n)$ y $\mathbb{R}^n$, pero la multiplicación está dada por la fórmula

$$(R_1, v_1) \star (R_2, v_2) = (R_1 R_2, R_1.v_2 + v_1).$$

2.5.2. Espacios simétricos como grupos de Lie cocientes

El siguiente teorema, ver Boothby 1986, da un criterio para que una variedad posea un métrica G-invariante.

Teorema 2.10. Considere un grupo de Lie G que actúa transitivamente sobre una variedad M. Si para algún punto $p \in M$ el subgrupo de isotropía G_p es un subgrupo de Lie compacto conectado de G, entonces M tiene una métrica G-invariante.

2.5. Espacios Simétricos

Los espacios simétricos aparecen naturalmente a partir de espacios homogéneos con métricas G-invariantes, como lo muestra el siguiente teorema, ver Boothby 1986.

Teorema 2.11. Suponga que G, M y p cumplen las condiciones del teorema 2.10. Si $\alpha : G \to G$ es un automorfismo involutivo (es decir, un isomorfismo de G en sí mismo) con un conjunto fijado G_p, entonces M es un espacio simétrico.

El recíproco del teorema anterior también es cierto, como lo dice el siguiente teorema, ver Helgason 1978.

Teorema 2.12. Si M es un espacio simétrico y p es cualquier punto en M, entonces M es difeomorfo al grupo de Lie cociente G/G_p, en donde, $G = I_0(M)$ es el componente conectado del grupo de Lie de isometrías de M y G_p es el subgrupo de Lie compacto de G que deja al punto p fijo. Además, existe un automorfismo involutivo $\alpha : G \to G$ que deja fijo a G_p.

Teorema 2.13. Un grupo de Lie conectado G con métrica bi-invariante es un espacio simétrico.

Ejemplo 2.18. El espacio euclídeo $\mathbb{R}^n$ es un espacio simétrico, como puede verse por teorema 2.13. La isometría involutiva ϕ_p esta dada por la reflexión alrededor de p, es decir, ϕ_p regresa líneas a través de p mediante la ecuación

$$\phi_p(q) = 2p - q.$$

Las geodésicas sobre un espacio simétrico $M = G/G_p$, son calculadas a través de la acción de grupo. Debido a que G es un grupo de isometrías que actúa transitivamente sobre M, es suficiente considerar únicamente geodésicas iniciando en el punto base p. Para un punto arbitrario $q \in M$, las geodésicas que inician en q son de la forma $g.\gamma$, donde $g = g.p$ y γ es una geodésica con $\gamma(0) = p$. La geodésicas son la imagen de la acción de un subgrupo uniparamétrico de G que actúa sobre el punto base p, como se enuncia en el siguiente teorema.

Teorema 2.14. Si M es un espacio simétrico con métrica G-invariante, como en el teorema 2.11, entonces una geodésica γ que inicia en el punto $p \in M$, es de la forma

$$\gamma(t) = \exp(tX).p,$$

en donde, X es un vector sobre el álgebra de Lie $\mathfrak{g}$.

Ejemplo 2.19. La esfera S^2 es un espacio simétrico. El grupo de rotación $SO(3)$ actúa transitivamente sobre S^2, es decir, para cualesquiera dos vectores unitarios x, y, existe una

rotación R tal que $Rx = y$. El polo norte $p = (0, 0, 1)$ se mantiene fijo mediante cualquier rotación del plano xy. Por lo tanto, el subgrupo de isotropía para p es equivalente a $SO(2)$. Por lo tanto, la esfera puede ser escrita como el espacio homogéneo $S^2 = SO(3)/SO(2)$. La isometría involutiva ϕ_p esta dada por la reflexión alrededor de p, es decir, mediante una rotación de la esfera al rededor del eje p por un ángulo de π.

Las geodésicas en el punto base $p = (0, 0, 1)$ son los círculos mayores a través de p, es decir, los meridianos. Las geodésicas en un punto arbitrario sobre la esfera S^2, también son los círculos mayores, es decir versiones rotadas de los meridianos. Como se muestra en el teorema 2.14, estas geodésicas son realizadas mediante la acción de grupo de un subgrupo uniparamétrico de $SO(3)$. Tal subgrupo consiste de todas las rotaciones alrededor de un eje fijo en $\mathbb{R}^3$ perpendicular a p. Se considera un vector tangente sobre T_pS^2 como el vector $v = (v_1, v_2, 0)$ sobre el plano xy. Entonces, el mapa exponencial está dado por

$$\mathrm{Exp}_p(v) = \left(v_1 . \frac{\sin \|v\|}{\|v\|}, v_2 . \frac{\sin \|v\|}{\|v\|}, \cos \|v\| \right), \tag{2.5}$$

en donde, $\|v\| = \sqrt{v_1^2 + v_2^2}$. Esta ecuación se puede derivar como una consecuencia de dos rotaciones que rotan el punto base $p = (0, 0, 1)$ al punto $\mathrm{Exp}_p(v)$. Primero, es una rotación alrededor del eje y por un ángulo de $\phi_y = \|v\|$. Luego lo segundo, es un alineamiento de la geodésica con el vector tangente v, esto es una rotación alrededor del eje z por un ángulo de ϕ_z, donde, $\cos(\phi_z) = v_1/\|v\|$ y $\sin(\phi_z) = v_2/\|v\|$.

El mapa logarítmico correspondiente para un punto $x = (x_1, x_2, x_3) \in S^2$, está dado por

$$\mathrm{Log}_p(x) = \left(x_1 . \frac{\theta}{\sin \theta}, x_2 . \frac{\theta}{\sin \theta} \right), \tag{2.6}$$

en donde, $\theta = \cos^{-1}(x_3)$, es la distancia esférica desde el punto base p al punto x. Notar que el punto antipodal, $-p$, no esta en el dominio del mapa logarítmico.

Capítulo 3

Modelos de Regresión y PLS

Los métodos clásicos de regresión están basados principalmente en el supuesto de observaciones independientes e idénticamente distribuidas normal, lo cuál en muchas aplicaciones es complicado de satisfacer, lo que ha llevado a la creación de diferentes métodos de regresión sobre los cuáles se tiene un número mínimo de supuestos que deben de cumplir las observaciones disponibles. Muchos de estos métodos caen en la parte de la teoría estadística no paramétrica, en la cual se desarrollan procesos de inferencia que no necesitan supuestos explícitos con respecto a la forma funcional de la distribución de las observaciones disponibles.

Dentro de la teoría estadística no paramétrica se encuentra la regresión no paramétrica, que consiste de técnicas para el ajuste de funciones de regresión cuando se tiene muy poco conocimiento apriori acerca de su forma funcional. Este método origina funciones suavizadas de la relación considerada. Los primeros métodos de suavizado fueron los promedios móviles, luego aparecieron la estimación vía kernel y la regresión local ponderada. En lo que sigue, se discuten y comparan varios modelos no-paramétricos de regresión no-lineal. En lugar de forzar una forma analítica predefinida sobre los datos, estos métodos aproximan la función no-lineal subyacente usando funciones suavizadas o splines sobre el conjunto de datos de entrenamiento. Por último se hace una breve revisión acerca de la teoría de modelos PLS.

3.1. Modelos de Regresión

Los modelos de regresión describen la relación entre dos clases de mediciones: las mediciones independientes o predictoras, denotadas por X, y las mediciones dependientes o respuestas, denotadas por Y.

La forma general de un modelo de regresión es:

$$y = f(x) + e$$

La respuesta Y, esta compuesta de dos partes:

- La parte sistemática $f(x)$, que depende de X.
- La parte aleatoria e, que es independiente de los predictores.

El primer paso en un análisis de regresión, es seleccionar la forma estructural de f y un método de estimación para los parámetros.

El modelo de regresión más común es el modelo de regresión lineal múltiple, en donde la función f es lineal. Uno de los métodos de estimación de parámetros más comunes es el de mínimos cuadrados ordinario (OLS).

Se requiere alguna precaución con el término "Modelo Lineal", debido a que este puede tener diferentes significados, como lo son:

- **Modelo lineal en los paŕametros.** Un modelo es lineal en los parámetros cuando la respuesta estimada $\hat{y}$ es una función lineal de los parámetros.
- **Modelo de Estructura Lineal.** Un modelo es de estructura lineal si $\hat{y}$ es función lineal de los predictores, X_i's.
- **Modelo de Estimador Lineal.** Un Modelo es de estimador lineal si cada $\hat{y}_i$ es función lineal de las respuestas y_i's.

El modelo de regresión ajustado mediante el método de OLS satisface las tres clases

de linealidades y en la mayoría de los casos, el ajuste da una primera aproximación satisfactoria de la función subyacente.

El estimador de mínimos cuadrados es el mejor estimador lineal insesgado (en el sentido de mínima varianza, es decir es el BLUE (por sus siglas en inglés)) cuando la relación entre X e Y es lineal y se satisfacen algunas suposiciones sobre la distribución de los errores. Sin embargo, el método de ajuste mediante OLS también puede ser aplicado a otros tipos de modelos de regresión.

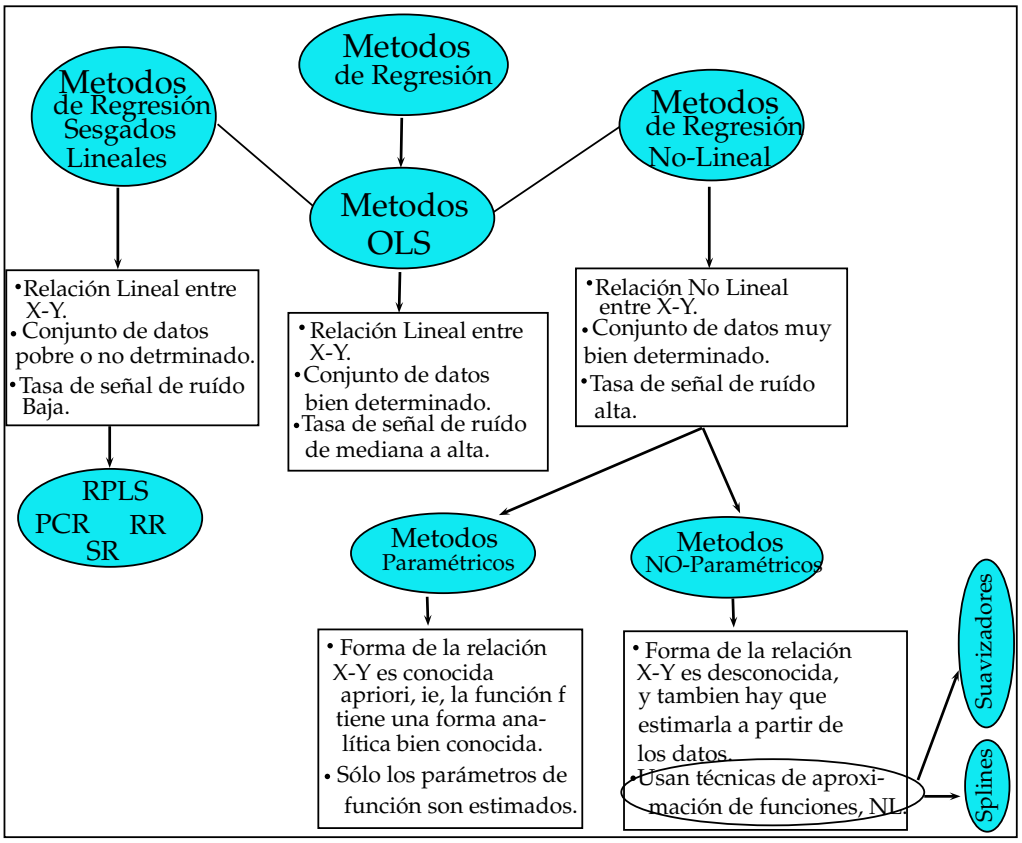


Figura 3.1: Representación Gráfica de Algunos Métodos de Regresión

3.1.1. Métodos de regresión lineal sesgados.

Los métodos de regresión lineal sesgados son frecuentemente usados en análisis de datos químicos y se caracterizan por: una relación lineal entre X e Y, un conjunto pobre de datos (ie, $\frac{\text{número de observaciones}}{\text{número de variables}} < 1$) y por una baja tasa de señal de ruido. En este grupo de métodos de regresión estan: La regresión mediante mínimos cuadrados parciales (PLSR), la regresión por componentes principales (PCR), la regresión de ridge (RR) y regresión paso a paso (SR, stepwise regression), ver figura 3.1 sobre una clasificación de métodos de regresión.

3.1.2. Métodos de regresión No-lineal.

Los métodos de regresión no-lineales se caracterizan por: una relación no-lineal entre X e Y, un conjunto bien determinado de datos (ie, $\frac{\text{número de observaciones}}{\text{número de variables}} >> 1$) y por una alta taza de señal de ruido. Éstos métodos de regresión no-lineal se pueden dividir a su vez en dos grupos:

Métodos de regresión no-lineal paramétricos.

Este grupo contiene métodos de regresión donde la forma de la relación entre X e Y es conocida a priori, es decir, la función f tiene una forma analítica bien definida, y sólamente los parámetros de la función son estimados a partir de los datos.

Métodos de regresión no-lineal no-paramétricos.

Este grupo contiene modelos no-lineales más flexibles, donde también la forma de la no-linealidad es estimada a partir de los datos. En estos métodos de regresión, en lugar de tratar con la forma analítica de la relación y con los parámetros de la función no-lineal, se usan técnicas de aproximación de funciones no lineales.

En la siguiente sección se hace una breve descripción de métodos de regresión no-lineales no-paramétricos que aproximan funciones no-lineales usando funciones suavizadas y splines.

3.2. Aproximación de Funciones

Cuando la forma analítica de la relación entre X e Y no es conocida, se debe aproximar la función usando un conjunto de datos de entrenamiento. Una técnica bien conocida y frecuentemente usada para lograr esto es la aproximación polinomial. Sin embargo, los polinomios en la práctica frecuentemente no trabajan bien y su proceso de ajuste es global, lo cual significa que una perturbación local afecta a toda la función estimada, debido a esto, es mas exitoso usar el acercamiento de aproximación de funciones por tramos. En este acercamiento, la idea es dividir el rango de X en tramos o intervalos, para tratar de estimar la función en cada intervalo de manera independiente a la estimación sobre los otros intervalos individuales. Las dos técnicas más usadas en este caso son: el uso de suavizados y la regresión spline. La principal diferencia entre estas dos técnicas es la forma en la que se realiza la partición del rango de X en intervalos. Ambas técnicas tienen un número más pequeño de intervalos que el número de datos iniciales y usan criterios de mínimos cuadrados para ajustar un polinomio local sobre cada tramo. Aunque existen suavizadores y splines multivariados, a continuación se describe al caso bivariado, ie, cuando solamente hay un predictor y una variable respuesta.

Tanto la técnica de suavizados como la regresión spline son técnicas aplicadas a datos bivariados X e Y, que producen una descomposición de la forma

$$y_i = f(x_i) + e_i \quad , \quad i = 1, \cdots, n$$

donde f es una función suave (llamada función suavizada o spline). La función f es usualmente ajustada basándose en el criterio de mínimos cuadrados.

Las técnicas de suavizados y splines se usan para aproximar funciones con el objetivo de describir la asociación entre el predictor X y la respuesta Y y luego predecir Y a partir de X.

3.2.1. Suavizados

Las funciones suavizadas pueden ser consideradas como estimadores de la esperanza condicional

$$f(x) = E[y \mid x].$$

Existen dos clases básicas de suavizados: El suavizado kernel y el suavizado por ventanas.

El suavizador kernel estima la esperanza condicional anterior en x_i, mediante la asignación de pesos a los puntos, ajustando un polinomio ponderado a todos los puntos y tomando el valor de la respuesta ajustada en x_i. El peso más grande es puesto en x_i y el resto de los pesos son simétricamente decrecientes a medida que los puntos están más alejados de x_i.

El suavizado por ventanas puede ser considerado como un caso especial del suavizado kernel, donde todos los puntos dentro de un cierto intervalo N_i (o ventana) alrededor de x_i tienen peso 1 y todos los puntos fuera del intervalo tienen peso 0. De acuerdo al grado del polinomio, el suavizado puede ser un promedio local (grado cero), un ajuste local lineal (grado uno), un ajuste local cuadrático (grado dos), etc.

En los suavizados por ventana los intervalos se están moviendo de punto a punto, a diferencia de los intervalos splines que son fijos. Esta es una de las principales diferencias entre suavizados y splines. Debido a que existen tantos intervalos (ventanas) como puntos datos, no es factible almacenar un suavizador en término de los coeficientes de los polinomios ajustados. En lugar de eso, es costumbre tratar el suavizado en su forma digital, es decir, como un conjunto de pares de puntos.

La regresión local promedio calcula el valor suavizado $\hat{y}_i$ como el promedio de aquellos y_i's con valores x_i's que están en un intervalo N_i alrededor de x_i:

$$\hat{y}_i = f(x_i) = E[y_j \mid x_j \in N_i].$$

La regresión local promedio, aunque es una técnica comúnmente usada, tiene algunas serias deficiencias. Esta no reproduce una línea recta si los valores de X no están equiespaciados y tiene un mal comportamiento en las fronteras.

El ajuste lineal local alivia ambos problemas. Éste calcula el valor suavizado $\hat{y}_i$ ajustando una línea recta (usualmente por mínimos cuadrados) a los puntos $x_j,\ y_j$ en el intervalo N_i y toma el valor de la respuesta ajustada en x_i. Polinomios de grados superiores se pueden ajustar en una forma similar.

Por ventajas computacionales, el intervalo es tomado tal que sea simétrico, ie, x_i tiene igual número de vecinos a la derecha y a la izquierda. Cuando se mueve al siguiente valor x_i, un punto queda en el intervalo a la izquierda y un punto entra al intervalo sobre la derecha. Tanto la regresión local promedio como el ajuste lineal local pueden ser calculados a través de los valores previos usando formulas de actualización.

Aunque la función suavizada final puede parecer continua, no hay ninguna restricción

de continuidad sobre el ajuste de los polinomios locales. Esta es otra diferencia entre suavizadores y splines.

En ambos suavizados, el punto clave es: cómo seleccionar el tamaño del intervalo, también llamado parámetro generador. Este parámetro controla la cantidad de sesgo y varianza del suavizado. El sesgo al cuadrado (B^2) y la varianza V, son dos componentes del error cuadrático medio (MSE) de un estimador.

$$\begin{aligned} MSE(\hat{y}) &= B^2(\hat{y}) + V(\hat{y}) \\ &= (E(\hat{y}) - y)^2 + E[\hat{y} - E(\hat{y})]^2. \end{aligned}$$

El incremento de la complejidad de un modelo (disminuir el tamaño del intervalo) resulta en el decrecimiento del sesgo y en un crecimiento de la varianza del estimador; lo que se llama equilibrio sesgo-varianza. Uno debe hallar la complejidad óptima para obtener el mínimo MSE. Al contrario, al incrementar el tamaño del intervalo, ie, decrecer la complejidad del modelo, incrementa el sesgo y decrece la varianza del suavizado.

Un valor grande del generador hace menos ondulado el suavizado. En los casos extremos, cuando el intervalo contiene todos los puntos, la regresión local promedio produce una línea recta con pendiente cero, mientras que el ajuste lineal, produce una línea recta con una pendiente distinta de cero. Tal suavizado tiene menos varianza, pero sesgo alto. En el otro caso, un suavizado con intervalos que contienen solamente un punto y_i dado como el valor suavizado, tiene varianza muy alta pero no tiene sesgo.

Idealmente el valor del generador óptimo del suavizado, debería ser estimado vía validación cruzada. Cada punto i es borrado (indicado por $-i$) y el valor del suavizado en x_i con valor generador J es calculado desde los otros $(n-1)$-puntos. La bondad de ajuste es calculada como:

$$e^2_{CV}(J) = \frac{1}{N}\sum_{i=1}^{N}[y_i \ - f_{-i}(x_i \mid J)]^2.$$

El parámetro generador estimado J mediante validación cruzada, es el valor que minimice el criterio anterior. Este proceso resulta en un valor constante del generador sobre el rango completo del predictor.

Esta no es una solución óptima si la varianza del error o la segunda derivada de la función f cambia sobre el rango del predictor. Una varianza de error pequeña y una alta curvatura

en f exigen un valor pequeño del generador, mientras que en el caso de una varianza de error grande y baja curvatura, exigen un valor grande del generador. La solución es una variable generadora de suavizados.

3.2.2. Splines

Los splines son funciones estimadas obtenidas mediante el ajuste de polinomios por trazos. El rango de X es dividido dentro de intervalos fijos, lo cual es una de las diferencias entre funciones suavizadas y splines. Los intervalos son separados por los llamados nodos de localización. En cada intervalo un polinomio es ajustado con las restricciones de que en los nodos de localización la función sea continua, siendo esta otra de las diferencias entre funciones suavizadas y splines. La integral y derivada de un spline es también un spline de un grado mayor o menor, frecuentemente también con una restricción de continuidad. El grado de un spline puede variar de cero a un grado muy alto, sin embargo, los splines de primer, segundo y tercer grado son los de mayor uso.

Un spline esta definido por su grado, el número de nodos de localización, la posición de los nodos y por los coeficientes del polinomio ajustado en cada intervalo. Un spline de grado m con N nodos de localización $(t_k, \quad k = 1, \cdots, N)$ puede ser escrito en forma general como sigue:

$$y_i = \sum_{j=0}^{m} b_{0j} x_i^j + \sum_{k=1}^{N} \sum_{j=0}^{m} b_{kj} (x_i - t_k)_+^j \quad + e_i \ ,$$

donde, $x_i^j \quad$ y $\quad (x_i - t_k)_+^j$, son llamadas funciones bases; y la notación $(...)_+$ significa parte positiva.

Esta notación pone al spline como una ecuación de regresión ordinaria. Los coeficientes b_{0j} y b_{kj} son estimados minimizando un criterio de mínimos cuadrados.

Dependiendo sobre los requerimientos de continuidad en varios nodos de localización, no todas las funciones bases anteriores están presentes en un spline, ie, algunos de los coeficientes b_{0j} son ceros. Un spline frecuentemente usado de grado m, con N nodos, y con restricciones de continuidad sobre la función y sobre sus derivadas hasta de grado

$m-1$ tiene la forma siguiente:

$$y_i = \sum_{j=0}^{m} b_{0j} x_i^j + \sum_{k=1}^{N} b_k (x_i - t_k)_+^m \quad + e_i \ ,$$

el número de coeficientes en tal spline es $m + N + 1$.

Existen varias funciones bases equivalentes de representaciones del mismo spline. Otra forma del anterior spline, que será usada posteriormente en uno de los métodos de regresión, es:

$$y_i = b_0 + \sum_{k=1}^{N} b_k [s_k (x_i - t_k)]_+^m \quad + e_i \ ,$$

donde, s_k es: +1 ó -1.

Al ajustar un spline uno debe seleccionar el grado óptimo m, el número óptimo de nodos N, y la localización óptima de los nodos. El grado del spline es algunas veces fijado a priori. El número y la localización de los nodos, o es fijo o es variable. Los splines con localizaciones de nodos variables son llamados splines adaptativos.

Las siguientes reglas deben de tenerse en cuenta cuando seleccionamos localizaciones de nodos fijos:

- Las localizaciones de nodos deberían ser colocadas en puntos correspondientes a datos.
- Un mínimo de 5 puntos deberían de estar entre las localizaciones de nodos.
- Los puntos extremos deberían de estar centrados sobre los intervalos.
- Los puntos de inflexión deberían de estar cerrando las localizaciones de nodos.

Los splines adaptativos ofrecen funciones de aproximación más flexibles que los splines con nodos de localización fijos. Lo ideal debería ser estimar el grado óptimo m, el número óptimo de nodos N, y las localizaciones de nodos óptimas mediante validación cruzada para obtener el mejor spline predictivo. En una función suavizada, el parámetro generador controla el equilibrio entre sesgo y varianza. En un spline, tal equilibrio es controlado por el grado del polinomio ajustado y por el número de nodos. Al incrementar el grado y el número de nodos, ie, incrementar la complejidad del modelo, incrementa la varianza y

decrece el sesgo del spline. El caso extremo de $m = 1$ y $N = 0$ es la solución de mínimos cuadrados ordinaria.

3.3. Métodos Basados en Suavizados

3.3.1. Esperanza Condicional Alternante (ACE).

El modelo de regresión lineal tiene la siguiente forma:

$$y = b_0 + \sum_j b_j x_j + e \tag{3.1}$$

donde, $\{b_j \ : \ j = 0, 1, \ldots, p\}$ son los coeficientes de regresión. En el modelo ACE la respuesta o una función de la respuesta, es la suma de funciones suaves de los predictores, ie:

$$y = \sum_j f_j(x_j) + e, \quad o \quad g(y) = \sum_j f_j(x_j) + e. \tag{3.2}$$

Si sólo hay un predictor (p=1) y ninguna función de la respuesta es usada, el modelo ACE es simplemente una función suavizada. Las funciones f_j y g, llamadas funciones de transformación, no requieren tener ninguna forma de parametrización particular. Estas son estimadas mediante procesos de mínimos cuadrados, es decir, mediante la minimización de la siguiente función de errores

$$e^2(g, f_1, \ldots, f_p) = E\left[g(y) - \sum_j f_j(x_j)\right]^2. \tag{3.3}$$

Las funciones transformación son normalizadas a tener media cero y varianza unitaria, ie:

$$E[g(y)] = E[f_j(x_j)] = 0, \quad y \quad Var[g(y)] = 1. \tag{3.4}$$

Las funciones transformación g y f_j, van mejorando iterativamente en forma alternante.

Las funciones son inicializadas en un conjunto de valores tales como:

$$g(y) = y/\|y\|, \quad y \quad f_j(x_j) = b_j x_j, \quad j = 1, 2, \ldots, p \tag{3.5}$$

donde, $\| \star \| = \sqrt{E(\star)^2}$ y los b_j son los coeficientes estimados mediante OLS.

3.3.2. Técnica de Regresión Aditiva Múltiple Suave (SMART).

Mientras que los modelos ACE modelan la respuesta como la suma de funciones de transformación de los predictores, SMART pone funciones sobre varias combinaciones lineales de las predictoras, como sigue

$$y = \bar{y} + \sum_m a_m f_m \left(\sum_j b_{jm} x_j \right) + e. \tag{3.6}$$

La respuesta es modelada como una suma de M funciones no-paramétricas de (usualmente) combinaciones lineales de los predictores. Las funciones f_m son requeridas a ser suavizadas o arbitrarias en otros casos. Este método también es llamado regresión de proyección pursuit. Aun para M pequeña, muchas clases de funciones pueden ser estrechamente ajustadas mediante aproximaciones de esta forma. Estas funciones junto con los coeficientes de las combinacines lineales son conjuntamente optimizada, basado sobre el criterio de mínimos cuadrados:

$$e^2(f_1, \ldots, f_M, b_{11}, \ldots, b_{pM}, a_1, \ldots, a_M) = E[y - \hat{y}]^2. \tag{3.7}$$

La solución es invariante a rotación y escalamiento de las variables predictoras. Similar al ACE, las funciones son normalizadas:

$$E[f_m] = 0, \quad y \quad Var[f_m] = 1, \quad m = 1, 2, \ldots, M. \tag{3.8}$$

Los coeficientes en las combinaciones lineales son normalizados a:

$$\sum_j b_{jm}^2 = 1, \quad m = 1, 2, \ldots, M. \tag{3.9}$$

SMART, tambien puede modelar respuestas multiples:

$$y_k = \bar{y}_k + \sum_m a_{km} f_m \left(\sum_j b_{jm} x_j \right) + e_k, \quad k = 1, 2, \ldots, r. \tag{3.10}$$

El criterio de mínimos cuadrados a ser minimizado es modificado como:

$$e^2(f_1, \ldots, f_M, b_{11}, \ldots, b_{pM}, a_{11}, \ldots, a_{rM}) = \sum_k w_k E[y_k - \hat{y}_k]^2. \tag{3.11}$$

Los pesos preespecificados de la respuesta w_k dan alguna flexibilidad al balancear la importancia de las diferentes respuestas. La influencia de cada respuesta es proporcinal a su varianza. Para tener una importancia igual, los pesos de las respuestas deberian ser:

$$w_k = \frac{1}{Var[y_k]}. \tag{3.12}$$

Los coeficientes a_{km} y b_{jm}, y las funciones f_m son optimizadas conjuntamente en un algoritmo iterativo.

3.3.3. Mínimos Cuadrados Parciales No-Lineal (NLPLS).

NPLS es una extensión no-paramétrica no-lineal de PLS aplicando suavizados al modelo de relación no-lineal interno. PLS es un modelo lineal que regresa las variables respuestas sobre el conjunto de variables latentes,

$$y_k = \sum_m a_{km} c_m t_m + e_k, \quad k = 1, 2, \ldots, r, \tag{3.13}$$

dichas variables latentes son combinaciones lineales de las predictoras originales

$$t_m = \sum_m b_{jm} x_j, \quad m = 1, 2, \ldots, M, \tag{3.14}$$

c_m es el coeficiente de relación interna mientras que b_{jm} y a_{km} son los coeficientes de las variables latentes x e y, respectivamente.

PLS, puede ser descompuesto dentro de tres modelos lineales, como sigue, (en forma matricial)

1. $\mathbf{X} = \mathbf{TB} + \mathbf{E_X}$, x-variables latentes (multivariadas).
2. $\mathbf{Y} = \mathbf{UA} + \mathbf{E_Y}$, y-variables latentes (multivariadas).
3. $\mathbf{U} = \mathbf{TC} + \mathbf{E_U}$, relación interna (univariada).

Extender PLS para describir una relación no-lineal, podría involucrar la no-linealización de ninguna de las tres ecuaciones anteriores. Debido a su simplicidad, la elección más sensible es la ecuación de relación interna

$$u = f(t) + e_u. \tag{3.15}$$

Similar a ACE y a SMART, NLPLS también utiliza funciones suavizadas para aproximar la función no-lineal. Sin embargo, hay un diferencia importante entre NLPLS y los dos métodos de suavizados discutidos anteriormente basados sobre métodos de regresión. Mientras que ACE y SMART estiman funciones y coeficientes mediante técnicas de mínimos cuadrados y por lo tanto son métodos insesgados, NLPLS, utiliza un estimador sesgado que no es de mínimos cuadrados para los coeficientes. ACE y SMART, deberían ser usados solamente con datos de una tasa observaciones/variables alta, es decir, conjuntos de datos bien determinados. NLPLS, aunque puede usar mas grados de libertad que el modelo PLS lineal, también puede ser usado para conjuntos de datos no determinados.

PLS lineal calcula una combinación lineal de las variables latentes. En el espíritu del modelo; SMART, NLPLS calculan una combinación lineal de funciones no-paramétricas de las variables predictores. En otras palabras, la relación interna es calculada mediante suavizados en lugar de regresión lineal.

$$y_k = \bar{y}_k + \sum_m a_{km} f_m \left(\sum_j b_{jm} x_j \right) + e_k, \quad k = 1, 2, \ldots, r. \tag{3.16}$$

Aunque NLPLS y SMART tienen formas estructurales iguales, usan diferentes técnicas para estimar los parámetros y las funciones.

3.4. Métodos Basados en Splines

3.4.1. Clasificación y Árboles de Regresión (CART).

La CART, también llamado regresión particionada recursiva, usa aproximaciones constantes por pedazos, ie, splines de grado cero para estimar la función no-lineal. Si sola-

mente hay un predictor, la ecuación para el spline con $m = 0$ es:

$$y_i = \sum_{j=0}^{0} b_0 x_i^0 + \sum_{k=1}^{N} b_k (x_i - t_k)_+^0 + e_i. \tag{3.17}$$

Centrando a y se elimina el primer término constante. En el segundo término la función base tiene valor 1 si $x_i > t_k$ y cero en otro caso, de donde el término completo tiene valor b_k ó 0. La extensión multivariada del modelo es:

$$y_i = \sum_{k=1}^{N} b_k B_k^0 + e_i = \sum_{k=1}^{N} b_k I(x_i \in R_k) + e_i, \tag{3.18}$$

donde, $R = \cup R_k$ y $R_k \cap R_{n \neq k} = 0$ y I es la función indicadora.

El espacio predictor R es particionado dentro de $\{R_k \ , \ k = 1, 2, \ldots, N\}$ subregiones. Una b_k constante es asignada a cada subregión R_k. En el modelo anterior se dice que la respuesta estimada es b_k si el vector predictor $\mathbf{x}_i$ esta en la subregión R_k. Debido a que las subregiones son excluyentes, $\mathbf{x}_i$ puede estar solamente en una de las subregiones, así solamente una función base B_k^0 toma un valor de 1 y el resto tienen valor cero.

La partición óptima del espacio predictor y los valores constantes b_k son estimados a partir del conjunto de datos de entrenamiento, basado sobre el siguiente criterio de mínimos cuadrados:

$$e^2 = E\left[y - \sum_{k=1}^{N} b_k B_k^0\right]^2. \tag{3.19}$$

Las constantes son calculadas mediante promedios locales:

$$b_k = \frac{\sum_i y_i I(\mathbf{x}_i \in R_k)}{\sum_i I(\mathbf{x}_i \in R_k)}. \tag{3.20}$$

El modelo final se da en forma de un árbol.

3.4.2. Regresión Spline Adaptativa Múltivariada (MARS).

La MARS, es una extensión multivariada de spline adaptativo con grado 1 o 3. Un spline bivariado con $m = 1$ puede ser escrito como:

$$y_i = b^0 + \sum_{k=1}^{N} b_k s_k B_k^1 + e_i \tag{3.21}$$

$$= b^0 + \sum_{k=1}^{N} b_k [s_k(x_i - t_k)]_+^1 + e_i \tag{3.22}$$

donde s_k es 1 ó -1. Cada función base involucra solamente un predictor y un nodo de localización. La MARS es una extensión multivariada del anterior modelo bi-variado :

$$y_i = b^0 + \sum_{k=1}^{N} b_k \prod_{j=1}^{J} [s_{kj}(x_{ij} - t_{kj})]_+^1 + e_i. \tag{3.23}$$

Cada función base multivariada es un producto de J funciones base univariadas. El parámetro J puede ser diferente para cada función base multivariada. Similarmente una función base multivariada de tercer grado es:

$$B_k^3 = \prod_{j=1}^{J} [s_{kj}(x_{ij} - t_{kj})]_+^3. \tag{3.24}$$

Una vez las funciones base son calculadas los coeficientes de regresión b_k son estimados mediante el proceso de mínimos cuadrados.

Los splines multivariados anteriores son splines adaptativos en la MARS, lo cual significa que los nodos de localización no son fijos, sino que son optimizados basados sobre el conjunto de entrenamiento.

Como en otros métodos no-lineales, un punto crucial es determinar la complejidad óptima del modelo, ie, el equilibrio óptimo del sesgo-varianza. En la MARS la complejidad esta determinada por: el grado del polinomio ajustado, el número de términos N y por el orden J de las funciones base multivariadas. Estos parámetros deberían ser optimizados mediante validación-cruzada para obtener el mejor modelo predictivo. En la MARS la validación cruzada generalizada (GCV) es aplicada para estimar la complejidad óptima

del modelo. El criterio a minimizar es:

$$GCV(N) = \frac{(1/n)\sum[y_i - \hat{f}_N(x_i)]^2}{[1 - C(N)/n]^2}, \tag{3.25}$$

de donde al incrementar N-el número de términos en el modelo, aumenta la varianza de la función f_N. El anterior criterio de GCV tiene el criterio de mínimos cuadrados en el numerador y un término de penalidad de la complejidad, $C(N)$ en el denominador.

3.5. La Metodología PLS

La regresión por mínimos cuadrados parciales (PLS) es una técnica de relación de variables que fue introducida por Wold 1975a, Wold 1982, Wold 1985 y extendida al campo de la quimiometría por su hermano Wold et al. 1984, Wold 2001, Wold et al. 2001, Wold, Sjöström, and Eriksson 2001, Martens and Naes 1989, Martin et al. 2001, Martens 2001, entre otros autores. El contenido que sigue de este capítulo está principalmente basado en Höskuldsson 1988.

La regresión PLS es una técnica reciente que combina y generaliza características del análisis de componentes principales (PCA) y de la regresión lineal múltiple (MRL). Su objetivo es predecir un conjunto de variables dependientes a partir de un conjunto de variables independientes o predictoras. La predicción es llevada a cabo extrayendo a partir de un conjunto de predictores un conjunto de factores ortogonales llamados variables o componentes latentes, las cuales tienen mejor potencia predictiva que las variables predictoras originales. A partir de estas variables latentes se pueden crear visualizaciones gráficas semejantes a las realizadas en PCA. La calidad de predicción obtenida a partir de un modelo de regresión PLS se evalúa mediante técnicas de validación cruzada tales como el bootstrap y el jackknife.

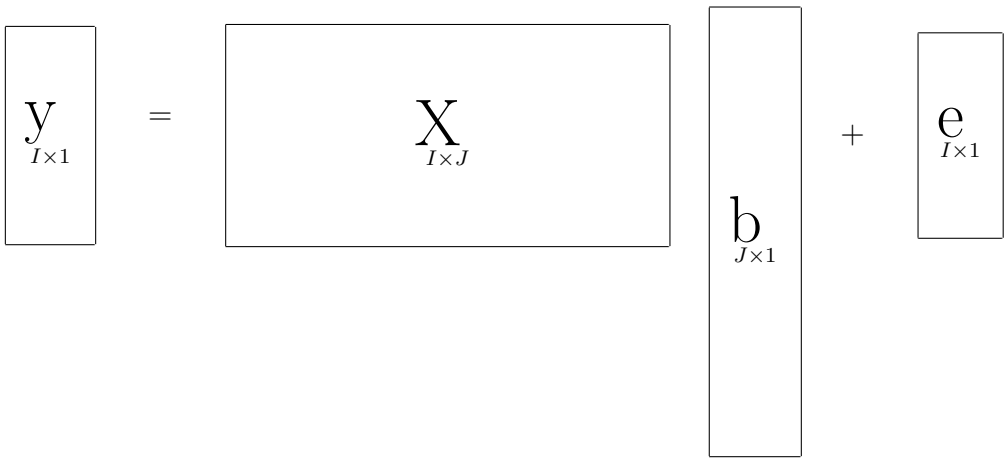
La regresión PLS es particularmente útil cuando se necesita predecir un conjunto de variables dependientes a partir de un conjunto muy grande de variables independientes o predictores. La regresión PLS se originó en las ciencias sociales, específicamente Economía, pero fue más popular inicialmente en Quimiometría, es decir sobre el cálculo estadístico en química, debido en parte tanto a Herman como Svante, Wold, Sjöström, and Eriksson 2001 y en evaluación sensorial debido a, Martens and Naes 1989. La regresión PLS también se ha convertido en una herramienta de elección en las ciencias sociales como una técnica multivariada tanto para datos experimentales como no experimentales, Worsley 1997, McIntosh and N.J. 2004. La regresión PLS fue presentada inicialmente como un

algoritmo similar al método de la potencia, usado en el cálculo de eigenvectores, pero fue rápidamente desarrollado en el campo de la estadística, Burnham, Viveros, and MacGregor 1996, Höskuldsson 2001, Tenenhaus 1998.

3.5.1. Descripción de la Regresión PLS: Una variable dependiente

En esta sección se utilizará la siguiente notación: I representará el número de observaciones sobre 1-variable dependiente que serán almacenadas en un vector $I \times 1$ denotada por $\mathbf{y}$, mientras que los valores de J- predictores medidos sobre estos I-observaciones serán almacenadas en una matriz $I \times J$ denotada por $\mathbf{X}$. En este caso se habla de regresión PLS1.

El objetivo es predecir $\mathbf{y}$ a partir de $\mathbf{X}$ y describir su estructura común. Cuando $\mathbf{X}$ es de rango completo este objetivo se podría lograr usando regresión múltiple ordinaria, pero cuando el número de predictoras es grande comparado al número de observaciones, la matriz $\mathbf{X}$ probablemente sea singular y el acercamiento de regresión clásica ya no es posible, debido a la multicolinealidad. A esta configuración de datos se le llama frecuentemente "el problema de N pequeño P grande " y aparece en muchas áreas de investigación, dicho problema se ilustra gráficamente como sigue:



Aunque la metodología de regresión PLS tiene sus principales aplicaciones en situaciones donde se encuentra presente el problema de "N pequeño P grande " , también se ha mostrado para el caso de PLS1, que dicha técnica origina en algunos casos mejores re-

sultados de predicción aún en situaciones estándares en donde no necesariamente se encuentra presente en problema un N-pequeño, comparado con los modelos de regresión estándar que generalmente se usan en estas situaciones, como lo son regresión OLS, selección de subconjunto de variables, PCR y métodos shrinkage de Stein, ver Garthwaite 1994.

El desempeño del método de regresión PLS es similar al del método de regresión por componentes principales (PCR), pero comparado con otras técnicas de regresión utilizadas para analizar datos con problemas de multicolinealidad, una ventaja de la regresión PLS es que la información en la variable $\mathbf{y}$ es usada en el proceso de construcción de las componentes o variables latentes. Otra ventaja es que el método de regresión PLS frecuentemente requiere un número menor de componentes que la PCR para dar buenas predicciones, además siempre es claro sobre cuáles componentes deberían incluirse en la regresión una vez que el número a de componentes ha sido determinado. Otra ventaja es la facilidad con la que se realizan los cálculos, lo que lo ha llevado a convertirse en un método muy popular, inicialmente en el área de la quimiometría y posteriormente en aplicaciones de muchas otras áreas de la ciencia como la economía, ciencias sociales, ciencias médicas.

Existen muchas variantes del algoritmo con el que fue descrito inicialmente la regresión PLS, pero todas estas son equivalentes. Ahora se pasa a formular el algoritmo original, el cual es uno de los más comúnmente usado hoy en día. Como se mencionó inicialmente en esta sección, se inicia con un conjunto de datos $(\mathbf{X}, \mathbf{y})$ de dimensión $I \times (J+1)$, se asume que tanto la matriz $\mathbf{X}$ como el vector $\mathbf{y}$ han sido centrados. Las etapas del algoritmo son:

Se hace $\mathbf{E}_0 = \mathbf{X}$ y $\mathbf{f}_0 = \mathbf{y}$. Para $k = 1, 2, \ldots, a$, se calculan:

Paso 1. $\mathbf{w}_k = \mathbf{E}_{k-1}^T \mathbf{f}_{k-1}$, $\mathbf{t}_k = \mathbf{E}_{k-1}\mathbf{w}_k$, (estimación de pesos y scores factores de $\mathbf{X}$).

Paso 2. $\mathbf{p}_k = \frac{\mathbf{E}_{k-1}^T \mathbf{t}_k}{\mathbf{t}_k^{'}\mathbf{t}_k} = \frac{\mathbf{X}^T \mathbf{t}_k}{\mathbf{t}_k^{'}\mathbf{t}_k}$, (estimación de cargas (o loadings) de $\mathbf{X}$).

Paso 3. $q_k = \frac{\mathbf{f}_{k-1}^{'}\mathbf{t}_k}{\mathbf{t}_k^{'}\mathbf{t}_k} = \frac{\mathbf{y}^{'}\mathbf{t}_k}{\mathbf{t}_k^{'}\mathbf{t}_k}$, (estimación de cargas (o loadings) de $\mathbf{y}$).

Paso 4. $\mathbf{E}_k = \mathbf{X} - \mathbf{t}_1\mathbf{p}_1^{'} - \mathbf{t}_k\mathbf{p}_k^{'}$, (deflaxión de la matriz $\mathbf{X}$).

Paso 5. $\mathbf{y}_k = \mathbf{y} - \mathbf{t}_1\mathbf{q}_1 - \mathbf{t}_k\mathbf{q}_k$. (deflaxión del vector $\mathbf{y}$).

La interpretación de las diferentes etapas del algoritmo anterior son como sigue: En el paso (1) se construyen las variables latentes, las cuáles son combinaciones lineales de

las variables originales $\mathbf{x}'$s de la etapa anterior. En los pasos (2) y (3), se construyen los loadings tanto para $\mathbf{X}$ como para $\mathbf{y}$, mediante ajustes de regresiones de mínimos cuadrados. Finalmente en los pasos (4) y (5) se obtienen nuevas variables $\mathbf{X}$ y $\mathbf{y}$, las cuales son calculadas como residuales.

Ahora, si $\mathbf{x}_0 = (x_{01}, x_{02}, \ldots, x_{oJ})$ es un conjunto de mediciones sobre un nuevo individuo, se definen $\mathbf{e}_0 = \mathbf{x}_0 - \bar{\mathbf{x}}$ con $\bar{\mathbf{x}} = (\bar{x_1}, \bar{x}_2, \ldots, \bar{x}_J)$ y entonces los nuevos scores y residuales están dados respectivamente por:

$$t_{k0} = \mathbf{e}_{k-1}\mathbf{w}_k \ , \quad \text{y} \quad \mathbf{e}_k = \mathbf{e}_{k-1} - t_{k0}\mathbf{p}_k.$$

El valor correspondiente a $\mathbf{y}_0$ es predicho en la etapa a mediante:

$$\hat{y}_{a0} = \bar{y} + \sum_{k=1}^{a} t_{k0}q_k = \bar{y} + \sum_{k=1}^{a} t_{k0}(\mathbf{t}_k^{'}\mathbf{t}_k)^{-1}\mathbf{t}_k^{'}\mathbf{y}.$$

El número de componentes a es usualmente determinado mediante validación cruzada u otros métodos que han sido propuestos en la literatura.

Algunas propiedades de las variables latentes del algoritmo anterior son las siguientes: De los pasos (1) (2) y (4), se observa que los scores $\mathbf{t}_k$ son ortogonales y los pesos $\mathbf{w}_k$ también son ortogonales y satisfacen la siguiente relación recursiva:

$$\mathbf{w}_{k+1} = \mathbf{s} - \mathbf{S}\mathbf{W}_k(\mathbf{W}_k^{'}\mathbf{S}\mathbf{W}_k)^{-1}\mathbf{W}_k^{'}\mathbf{s} \ ,$$

con $\mathbf{W}_k = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k)$, $\mathbf{S} = \mathbf{X}'\mathbf{X}$ y $\mathbf{s} = \mathbf{X}'\mathbf{y}$. Además, el vector de regresión PLSR con a-componentes puede ser escrito como

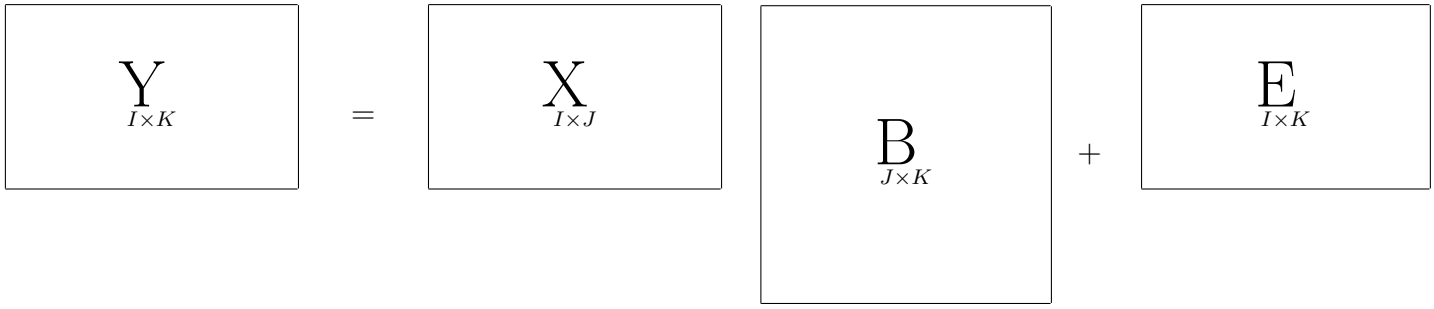
$$\mathbf{b}_a = \mathbf{W}_a(\mathbf{W}_a^{'}\mathbf{S}\mathbf{W}_a)^{-1}\mathbf{W}_a^{'}\mathbf{s}.$$

Un resultado asociado a la regresión, es que el R^2 es al menos tan grande como el R^2 de PCR con el mismo número de componentes y que el vector de regresión tiene propiedades de encogimiento. La regresión PLS1, se contrae en el sentido fuerte, es decir que la norma del vector de coeficientes de regresión no crece cuando el número de componentes aumenta.

3.5.2. Descripción de la Regresión PLS: más de una variable dependiente

En esta sección se utilizará la siguiente notación: I representará el número de observaciones sobre K variables dependientes que serán almacenadas en una matriz $I \times K$ denotada por $\mathbf{Y}$, mientras que los valores de J- predictores medidos sobre estos I-observaciones serán almacenadas en una matriz $I \times J$ denotada por $\mathbf{X}$. En este caso se habla de regresión PLS2.

El objetivo es el mismo que en en PLS1, es decir, predecir a $\mathbf{Y}$ a partir de $\mathbf{X}$ y describir su estructura común. Cuando $\mathbf{Y}$ es un vector y $\mathbf{X}$ es de rango completo, este objetivo se podría lograr usando regresión múltiple ordinaria, pero cuando el número de predictoras es grande comparado al número de observaciones, la matriz $\mathbf{X}$ probablemente sea singular y el acercamiento de regresión ya no es posible, debido a la multicolinealidad. Bajo este esquema también se repite el problema de "N pequeño P grande" y aparece en muchas áreas de investigación y que se ilustra gráficamente como sigue:



Varios acercamientos han sido desarrollados para hacer frente al problema de la multicolinealidad. Un método cercanamente relacionado a la regresión PLS es la regresión por componentes principales (PCR), el cual realiza un PCA a la matriz $\mathbf{X}$ y luego usa las componentes principales de $\mathbf{X}$ como las variables independientes de un modelo de regresión múltiple para predecir a $\mathbf{Y}$. En PCA, $\mathbf{X}$ es descompuesto usando su descomposición espectral en valores y vectores propios o su descomposición en valores y vectores singulares (SVD) de la forma

$$\mathbf{X} = \mathbf{R}\boldsymbol{\Delta}\mathbf{V}^T \ , \quad \text{con} \quad \mathbf{R}^T\mathbf{R} = \mathbf{V}^T\mathbf{V} = \mathbf{I} \ ,$$

donde, $\mathbf{R}$ y $\mathbf{V}$ son matrices de vectores singulares a izquierda y derecha, y $\boldsymbol{\Delta}$ es una

matriz diagonal con los valores singulares como elementos en su diagonal. Los vectores singulares son ordenados de acuerdo a sus correspondientes valores singulares los cuales son la raíz cuadrada de la varianza (es decir, los eigenvalores) de $\mathbf{X}$ explicada por sus vectores singulares. A las columnas de $\mathbf{V}$ se le llaman cargas (loadings) y a las columnas de $\mathbf{G} = \mathbf{R}\boldsymbol{\Delta}$ se le llaman los factores scores o componentes principales de $\mathbf{X}$, o simplemente scores o componentes. La matriz $\mathbf{R}$ de vectores singulares a izquierda de $\mathbf{X}$ (o la matriz $\mathbf{G}$ de componentes principales) son luego usadas para predecir a $\mathbf{Y}$ usando un modelo de regresión lineal múltiple estándar. Este acercamiento trabaja bien debido a que la ortogonalidad de los vectores singulares elimina el problema de multicolinealidad, pero el problema de elegir un subconjunto óptimo de predictores aún permanece. Una posible estrategia, consiste en mantener solamente unas pocas componentes principales, pero estas componentes principales fueron originalmente elegidas para explicar a $\mathbf{X}$ en lugar de $\mathbf{Y}$, y así no se tiene garantía de que las componentes principales, las cuáles explican óptimamente a $\mathbf{X}$, sean relevantes para la predicción de $\mathbf{Y}$.

Al contrario de la regresión por componentes principales, la regresión PLS encuentra las componentes a partir de $\mathbf{X}$ que mejor predicen a $\mathbf{Y}$. La regresión PLS, busca un conjunto de componentes (llamadas vectores latentes) que desarrollan una descomposición simultánea de $\mathbf{X}$ y $\mathbf{Y}$, con la restricción de que estas componentes explican tanto como sea posible de la *covarianza* entre $\mathbf{X}$ y $\mathbf{Y}$. Este paso generaliza a PCA. Luego se sigue mediante una etapa de regresión donde los vectores latentes obtenidos a partir de $\mathbf{X}$ son usados para predecir a $\mathbf{Y}$. La regresión PLS descompone tanto a $\mathbf{X}$ como a $\mathbf{Y}$ como un producto de un conjunto común de factores ortogonales y un conjunto de cargas específicas. Es decir, las variables son descompuestas como

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T \ , \quad \text{con} \quad \mathbf{T}^T\mathbf{T} = \mathbf{I} \ ,$$

con $\mathbf{I}$ la matriz identidad. Por analogía con PCA, a la matriz $\mathbf{T}$ se le llama la *matriz scores* y a la matriz $\mathbf{P}$ se le llama la *matriz de cargas*, (en regresión PLS las cargas no son ortogonales). Del mismo modo, $\mathbf{Y}$ es estimado mediante:

$$\hat{\mathbf{Y}} = \mathbf{T}\mathbf{B}\mathbf{C}^T \ ,$$

donde, $\mathbf{B}$ es una matriz diagonal con los *pesos de la regresión* como elementos en su diagonal y $\mathbf{C}$ es la *matriz de pesos* de las variables dependientes. Las columnas de $\mathbf{T}$ son los *vectores latentes*. Cuando el número de vectores latentes es igual al rango de $\mathbf{X}$, los vectores latentes desarrollan una descomposición exacta de $\mathbf{X}$.

Los vectores latentes podrían ser elegidos en distintas formas. En la formulación anterior cualesquier conjunto de vectores ortogonales que genere las columnas del espacio de $\mathbf{X}$

podría ser usado para jugar el papel de $\mathbf{T}$. Con el fin de especificar a $\mathbf{T}$, se requieren algunas condiciones adicionales. Para la regresión PLS, estas condiciones equivalen a encontrar dos conjuntos de pesos denotados por $\mathbf{w}$ y $\mathbf{c}$ con el fin de crear una combinación lineal de las columnas de $\mathbf{X}$ y $\mathbf{Y}$ (respectivamente), tales que, estas dos combinaciones lineales tienen máxima covarianza. Es decir, el objetivo es obtener un primer par de vectores

$$\mathbf{t} = \mathbf{X}\mathbf{w} \quad \text{y} \quad \mathbf{u} = \mathbf{Y}\mathbf{c} \ ,$$

con la restricción de que $\mathbf{w}^T\mathbf{w} = 1$ y $\mathbf{t}^T\mathbf{t} = 1$ y $\mathbf{t}^T\mathbf{u}$ sea máximo. Cuando los primeros vectores latentes son hallados, estos son sustraídos tanto de $\mathbf{X}$ como de $\mathbf{Y}$ y el proceso es reiterado hasta que $\mathbf{X}$ se convierta en una matriz nula.

Como se mencionó antes, el criterio de optimización de PLS maximiza la covarianza entre una combinación de los predictores X y la respuesta Y. Sin embargo, las medidas usuales de covarianza o correlación están altamente influenciadas por outliers, de donde se tiene que el método de PLS puede ser sensible a outliers. Debido a lo anterior, se han propuesto en algunos casos reemplazar la correlación de Pearson usual por la correlación de Spearman mediante rangos, debido a que ésta última es no sensible a valores outliers tanto en X como en Y, con lo cual se logra obtener robustes de la metodología PLS con respecto a la presencia de outliers, ver Rojo and Nguyen 2009. Otro trabajo en donde se ilustra una técnica PLS robusta es González, Peña, and Romera 2009.

De manera similar a las variantes Bayesinas que existen para alguna técnicas de regresión utilizando métodos de regularización como lo son, regresión de Ridge, regresión LASSO, entre otras, deben existir extensiones alternativas desde el punto de vista Bayesiano para modelos de regresión PLS.

3.5.3. Un Algoritmo para PLS

Las propiedades de la regresión PLS se pueden analizar a partir de un bosquejo del algoritmo original, llamado NIPALS. El primer paso es crear dos matrices $\mathbf{E} = \mathbf{X}$ y $\mathbf{F} = \mathbf{Y}$, dichas matrices son centradas y normalizadas por columna, (es decir, transformadas en Z-scores). La suma de cuadrados de estas matrices se denotan por SS_X y SS_Y. Antes de iniciar el proceso de iteración, el vector $\mathbf{u}$ es inicializado con valores aleatorios. El algoritmo NIPALS desarrolla las siguientes etapas:

Paso 1. $\mathbf{w} \propto \mathbf{E}^T\mathbf{u}$, (estimación de pesos de $\mathbf{X}$), se normaliza a $\mathbf{w}$

Paso 2. $\mathbf{t} \propto \mathbf{Ew}$, (estimación de scores factores de $\mathbf{X}$), se normaliza a $\mathbf{t}$

Paso 3. $\mathbf{c} \propto \mathbf{F}^T\mathbf{t}$, (estimación de pesos de $\mathbf{Y}$), se normaliza a $\mathbf{c}$

Paso 4. $\mathbf{u} \propto \mathbf{Fc}$, (estimación de scores factores de $\mathbf{Y}$)

Si $\mathbf{t}$ no converge, entonces voy al paso 1. y si $\mathbf{t}$-converge, entonces calculo el valor de b el cual es usado para predecir a $\mathbf{Y}$ desde $\mathbf{t}$ mediante, $b = \mathbf{t}^T\mathbf{u}$, y se calcula en factor de carga para $\mathbf{X}$ como $\mathbf{p} = \mathbf{E}^T\mathbf{t}$. Ahora se resta (parcialmente) el efecto de $\mathbf{t}$ tanto sobre $\mathbf{E}$ como sobre $\mathbf{F}$ como sigue: $\mathbf{E} = \mathbf{E} - \mathbf{tp}^T$ y $\mathbf{F} = \mathbf{F} - b\mathbf{tc}^T$. Esta resta es llamada deflaxión de las matrices $\mathbf{E}$ y $\mathbf{F}$. Para el caso $K = 1$ este algoritmo da los mismos resultados que la regresión PLS1. Los vectores $\mathbf{t}, \mathbf{u}, \mathbf{w}, \mathbf{c}$ y $\mathbf{p}$ son almacenados en las matrices correspondientes, el escalar b es almacenado en los elementos de la diagonal de la matriz $\mathbf{B}$. La suma de cuadrados de $\mathbf{X}$ (respectivamente $\mathbf{Y}$) explicada por el vector latente, se calcula como $\mathbf{p}^T\mathbf{p}$ (respectivamente b^2), y la proporción de varianza explicada se obtiene dividiendo la suma de cuadrados explicada por la correspondiente suma total de cuadrados (es decir, por SS_X y por SS_Y). Si $\mathbf{E}$ es una matriz nula, entonces el conjunto completo de vectores latentes es hallado, en otros casos, el proceso puede ser iterado desde desde el paso 1.

La regresión PLS está cercanamente relacionada a la descomposición en valores y vectores propios y a la descomposición en valores y vectores singulares. Si iniciamos el paso 1. del algoritmo, calculando a $\mathbf{w} \propto \mathbf{E}^T\mathbf{u}$, y se sustituyen los términos que están más a la derecha iterativamente, se obtiene la siguiente serie de ecuaciones

$$\mathbf{w} \propto \mathbf{E}^T\mathbf{u} \propto \mathbf{E}^T\mathbf{Fc} \propto \mathbf{E}^T\mathbf{FF}^T\mathbf{t} \propto \mathbf{E}^T\mathbf{FF}^T\mathbf{Ew}.$$

Lo anterior muestra que el vector de pesos $\mathbf{w}$ es el primer vector singular a derecha de la matriz $\mathbf{S} = \mathbf{E}^T\mathbf{F}$. Similarmente, el primer vector de pesos $\mathbf{c}$ es el vector singular a izquierda de $\mathbf{S}$. Con argumentos similares se muestra que los primeros vectores $\mathbf{t}$ y $\mathbf{u}$ son los primeros eigenvectores de $\mathbf{EE}^T\mathbf{FF}^T$ y $\mathbf{FF}^T\mathbf{EE}^T$ respectivamente. Esta última observación, muestra que los vectores de pesos también pueden ser obtenidos a partir de matrices de tamaños $I \times I$, lo cual es útil cuando el número de variables es mucho mayor que el número de observaciones, es decir en problemas de N pequeño y P grande.

3.5.4. Predicción de las Variables Dependientes

Las variables dependientes son predichas usando la formula de regresión multivariada como sigue

$$\widehat{\mathbf{Y}} = \mathbf{TBC}^T = \mathbf{XB}_{\mathrm{PLS}} \ , \quad \text{con} \quad \mathbf{B}_{\mathrm{PLS}} = (\mathbf{P}^{\mathrm{T}+})\mathbf{BC}^T \ ,$$

donde, $\mathbf{P}^{\mathrm{T}+}$ es la pseudo inversa de Moore-Penrose de $\mathbf{P}^{\mathrm{T}}$. Esta última ecuación asume que tanto, $\mathbf{X}$ como $\mathbf{Y}$ han sido estandarizadas apriori a la predicción. Con el fin de predecir una matriz no estandarizada $\mathbf{Y}$ a partir de una matriz no estandarizada $\mathbf{X}$, se usa la notación, $\mathbf{B}^{\star}_{\mathrm{PLS}}$, el cual se obtiene reintroduciendo las unidades originales dentro de $\mathbf{B}_{\mathrm{PLS}}$ y adicionando la primera columna que corresponde al intercepto (es decir, cuando usamos las unidades originales $\mathbf{X}$ necesita ser aumentada con una primera columna de unos, como en regresión múltiple). Si todas las variables latentes de $\mathbf{X}$ son usadas, esta regresión es equivalente a la regresión por componentes principales. Cuando sólo se usa un subconjunto de las variables latentes, la predicción de $\mathbf{Y}$ es óptima para este número de predictores. La interpretación de las variables latentes frecuentemente se facilitan, mediante gráficas parecidas a las de PCA.

3.5.5. Inferencia Estadística: Evaluación de la Calidad de la Predicción mediante regresión PLS.

La calidad de la predicción obtenida a partir de regresión PLS descrito hasta ahora corresponde al modelo de efectos fijo, es decir, un conjunto de observaciones es considerado como la población de interés y las conclusiones del análisis están restringidas a éste conjunto. En este caso el análisis es *descriptivo* y la cantidad de varianza (de $\mathbf{X}$ y $\mathbf{Y}$) explicada por un vector latente indica su importancia para el conjunto de datos bajo estudio. En éste contexto vale la pena considerar las variables latentes si su interpretación es significativa dentro del contexto investigado. Para un modelo de efectos fijo, la calidad global de un modelo de regresión PLS que usa L-variables latentes es evaluada, primero calculando la matriz predicha de las variables dependientes denotada por $\widehat{\mathbf{Y}}^{[\mathrm{L}]}$ y luego, se mide la similaridad entre $\widehat{\mathbf{Y}}^{[\mathrm{L}]}$ y $\mathbf{Y}$. Varios coeficientes están disponibles para esta tarea. El coeficiente de correlación al cuadrado es algunas veces usado, como también su matriz prima específica, el coeficiente R_V. El coeficiente más popular, es la *suma de cuadrados de residuales*, abreviada por RESS, y dada por:

$$\mathrm{RESS} = \|\mathbf{Y} - \widehat{\mathbf{Y}}^{[\mathrm{L}]}\| \ ,$$

donde, la $\|.\|$, es la norma de $\mathbf{Y}$, es decir, la raíz cuadrada de la suma de cuadrados de los elementos de $\mathbf{Y}$. Entre más pequeño sea RESS, mejor es la predicción, con un valor de 0, indicando una predicción perfecta. Para un modelo de efectos fijos, entre más grande sea L, es decir, el número de variables latentes usado, mejor es la predicción.

En la mayoría de las aplicaciones sin embargo, el conjunto de observaciones es una *muestra* de alguna población de interés. En éste contexto, el objetivo es *predecir* el valor de la variable dependiente para *nuevas* observaciones que se originan a partir de la misma población de la muestra. Esto corresponde a un *modelo aleatorio*. En este caso la cantidad de varianza explicada por una variable latente, indica su importancia en la predicción de $\mathbf{Y}$. En este contexto, una variable latente es relevante únicamente si este mejora la predicción de $\mathbf{Y}$ para nuevas observaciones. Y así, esto nos lleva al problema de cuáles y cuántas variables latentes deberían mantenerse en el modelo de regresión PLS con el fin de lograr una generalización óptima, es decir, predicción óptima para nuevas observaciones. Con el fin de estimar la capacidad de generalización de la regresión PLS, los acercamientos paramétrico estándar no pueden ser usados y por lo tanto el desempeño de un modelo de regresión es evaluado con técnicas de *remuestreo* basadas en computación, tales como el botstrap y técnicas de validación cruzada, en donde los datos son separados dentro de un conjunto de *entrenamiento* (usado para construir el modelo) y un conjunto de *prueba* (usado para probar el modelo). Un ejemplo popular de este último acercamiento, es el *jackknife*. En el jackknife, cada observación, es a su vez, borrada del conjunto de datos, y el resto de observaciones constituyen el conjunto de entrenamiento y son usadas para construir un modelo de regresión PLS que es aplicado para predecir la observación dejada por fuera, la cual constituye el conjunto de prueba. Con este proceso, cada observación es predicha de acuerdo a un modelo de efectos aleatorio. Las observaciones predichas son entonces almacenadas en una matriz denotada por $\widetilde{\mathbf{Y}}$.

Para un modelo de efectos aleatorio, la calidad global de un modelo de regresión PLS que usa L-variables latentes es evaluado usando L-variables para calcular, de acuerdo al modelo aleatorio, la matriz denotada por $\widetilde{\mathbf{Y}}^{[\mathrm{L}]}$, la cual almacena los valores predichos de las observaciones para las variables dependientes. La calidad de la predicción es evaluada como la similaridad entre $\widetilde{\mathbf{Y}}^{[\mathrm{L}]}$ y $\mathbf{Y}$. Como para el modelo de efectos fijos, esto se puede realizar con el coeficiente de correlación al cuadrado, como con el coeficiente R_V. Por analogía con el coeficiente RESS, también se puede usar la *suma de residuales predichos al cuadrado*, denotada por PRESS. La cual es calculada por medio de

$$\mathrm{PRESS} = \|\mathbf{Y} - \widetilde{\mathbf{Y}}^{[\mathrm{L}]}\|^2.$$

Entre más pequeño sea la PRESS, mejor es la predicción para un modelo de efectos aleatorio, con un valor de 0 que indica una predicción perfecta.

Al contrario al modelo de efectos fijos, la calidad de la predicción para un modelo aleatorio no siempre crece con el número de variables latentes usadas en el modelo. Típicamente, la calidad primero crece y luego decrece. Si la calidad de la predicción decrece cuando el número de variables latentes crece, esto indica que el modelo está *sobreajustando* los datos, es decir, que la información útil para ajustar las observaciones a partir del conjunto de entrenamiento no es útil para ajustar *nuevas* observaciones. Por lo tanto, para un modelo aleatorio, es crucial determinar el número óptimo de variables latentes a mantener para construir el modelo. Un acercamiento directo es dejar de adicionar variables latentes cuando la PRESS deje de decrecer. Un acercamiento más elaborado, empieza calculando para l-variables latentes la taza Q_l^2, definida como:

$$Q_l^2 = 1 - \frac{\text{PRESS}_l}{\text{RESS}_{l-1}} \ ,$$

con, PRESS_l (respectivamente RESS_{l-1}) iniciando en los valores de la PRESS (respectivamente RESS) para la l-ésima (respectivamente l-1) variable latente, (donde, $\text{RESS}_0 = K \times (I-1)$). Una variable latente es retenida si su valor de Q_l^2 es mayor que algún valor albitrario generalmente dado igual a $(1 - 0 - 95^2) = 0.0975$, (un conjunto alternativo de valores hace el umbral a 0.05 cuando $I < 100$ y a 0 cuando $I > 100$, Tenenhaus 1998). Obviamente, la elección del umbral es importante desde el punto de vista teórico, pero, desde el punto de vista práctico, los valores que se indicaron anteriormente parecen satisfactorios.

Cuando el número de variables latentes del modelo ha sido decidido, se pueden derivar intervalos de confianza para valores predichos, usando la técnica bootstrap. Cuando se usa el bootstrap, un número grande de muestras es obtenido, extrayendo, para cada muestra, observaciones con reemplazamiento desde el conjunto de entrenamiento. Cada muestra nos proporciona un valor de $\mathbf{B}_{\text{PLS}}$, el cual es usado para estimar los valores de las observaciones dentro del conjunto de prueba. La distribución de los valores de estas observaciones es luego usada para estimar la distribución muestral y derivar intervalos de confianza.

Capítulo 4

Regresión por mínimos cuadrados parciales (PLS) sobre datos variedad-valuados: Una aplicación a matrices simétricas definidas positivas (PD)

4.1. Introducción

En estudios imágenes de resonancia magnética por tensor de difusión (TD-MRI), en cada voxel de una imagen se calcula un tensor de difusión (DT), el cual describe la difusión local de las moléculas de agua en varias direcciones sobre esa región del cerebro. Para medir la difusión se utilizan una secuencia de imágenes, las cuales incluyen un ruido que produce incertidumbre en la estimación de los tensores y en la estimación de ciertas cantidades inherentes a ellos, como lo son los eigenvalores, los eigenvectores, la tasa de fracción anisotrópica (FA) y las trayectorias de fibras que se construyen basadas en estos últimos parámetros. Las imágenes de tensor difusión (DTI) son una herramienta poderosa para evaluar cuantitativamente la integridad de la conectividad anatómica en la materia blanca de poblaciones clínicas. Los métodos que son utilizados para el análisis a nivel grupal de DTI's son análisis estadísticos de ciertas medidas invariantes, como los son, los eigenvalores, los eigenvectores o direcciones principales, la fracción anisotrópica, la difu-

sividad promedio. Pero estas medidas invariantes no capturan toda la información sobre los DT's completos, lo cual lleva a un decrecimiento en el poder estadístico de las DTI's para detectar los cambios sutiles en la materia blanca. Debido a lo anterior se han estado desarrollando nuevos métodos estadísticos para analizar de forma completa los DT's como respuesta y establecer su asociación con un conjunto de covariables, Zhu et al. 2009, Yuan et al. 2012, Li et al. 2009. En algunos de estos desarrollos se ha utilizado la métrica log-euclídea, transformando los DT's de un espacio no-lineal en sus matrices logarítmicas sobre un espacio euclídeo. Se han planteado modelos semi-paramétricos, para estudiar la relación entre el conjunto de covariables y los DT's como respuesta, se han desarrollado procesos de estimación y procesos de pruebas de hipótesis basado en estadísticas de pruebas y en métodos de remuestreo, para evaluar simultáneamente la significancia estadística de hipótesis lineales a través de grandes regiones de interés (ROI). Un análisis estadístico apropiado de los DT's es importante para entender el desarrollo normal del cerebro, las bases neuronales de desordenes neuropsiquiátricos y los efectos conjuntos de factores ambientales y genéticos sobre la función y estructura cerebral. Además cualquier método estadístico para tensores de difusión completo puede ser directamente aplicado a matrices de tensores definidas positivas en anatomía computacional para entender la variación de forma entre imágenes cerebrales estructurales, Grenander and Miller 1998, Lepore et al. 2008.

Los datos matriz simétrica definida-positiva-valuados (DPV), ocurren en una amplia variedad de aplicaciones, como por ejemplo en las DTI's, en donde un DT simétrico definido positivo 3x3, el cual rastrea la difusión efectiva de las moléculas de agua en ciertas regiones del cerebro, es estimado en cada voxel de la imagen. Otra aplicación de datos matriz simétrica DPV, se da en los estudios mediante imágenes de resonancia magnética funcional (fMRI), en donde una matriz de covarianza simétrica DP es calculada para delinear la conectividad funcional entre diferentes ensambles neuronales involucrados en la ejecución de ciertas tareas cognitivas complejas o en procesos de percepción, Fingelkurts and Kahkonen 2005. A pesar de la popularidad de los datos matriz simétrica DPV, existen pocos métodos estadísticos para el análisis de matrices simétricas DP como variables respuestas que viven en una variedad Riemanniana. Entre la literatura que existe para el análisis estadístico mediante modelos de regresión de datos matrices simétricas DP considerados como respuesta y un número pequeño de covariables de interés sobre un espacio euclídeo están: Fletcher and Joshi 2007, Batchelor et al. 2005, Pennec, Fillard, and Ayache 2006, Schwartzman 2006, Kim and Richards 2010, Zhu et al. 2009 y Barmpoutis et al. 2007, pero debido a que los datos matrices simétricas DP no forman un espacio vectorial, no se pueden aplicar directamente las técnicas de regresión multivariada clásicas para establecer la relación entre este tipo de datos y un conjunto de covariables de interés.

En el entorno de un número grande de covariables con alta presencia de multicolinealidad

y pocas observaciones disponibles, no se han planteado métodos de regresión para estudiar la relación entre dichas covariables y variables respuesta de matrices simétricas DP que viven en espacios no-euclídeos. En este Artículo se plantea la regresión PLS, utilizando la estrategia de los mapas exponencial y logaritmo riemanniano para transformar los datos a espacios euclídeos. El desarrollo de la técnica es similar al esquema que existe para la metodología de análisis de datos matriz simétrica DP como respuesta en un modelo de regresión clásico y en un modelo de regresión local polinomial, planteados por Zhu et al. 2009 y Yuan et al. 2012. El modelo de regresión PLS es inicialmente evaluado mediante un conjunto de datos simulados, utilizando técnicas de validación estadística que existen en la actualidad, como lo es la validación cruzada. Se analiza el comportamiento de la técnica de regresión PLS comparado con la técnica de reducción de dimensión clásica PCR.

El capítulo está estructurado de la siguiente forma: en la sección 2, se da una breve revisión acerca de la teoría que existe para modelos de regresión clásicos, PCR y PLSR. En la sección 3, se revisan algunas propiedades sobre la estructura geométrica riemanniana del espacio de matrices simétricas DP, se presentan un bosquejo de los modelos de regresión que existen en la actualidad así como los métodos de estimación de sus coeficientes de regresión. En la sección 4, se presenta nuestro modelo de regresión PLS junto con el proceso de estimación utilizado y su aplicación o evaluación en un conjunto de datos simulados. En la sección 5, se dan algunas conclusiones y recomendaciones para trabajos futuros.

4.2. Métodos de Regresión

4.2.1. Regresión clásica

Dado un conjunto de datos $\{(\mathbf{x}_i, y_i) \ : \ i = 1, 2, \ldots, n\}$, compuestos de una respuesta y_i y un vector $k \times 1$ de covariables $\mathbf{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ik})^T$, en donde la respuesta pueden ser observaciones continuas, discretas o cualitativas y las covariables pueden ser cuantitativas o cualitativas, un modelo de regresión frecuentemente incluye 2-elementos claves: una función enlace $\mu_i(\beta) = E[y|\mathbf{x}_i] = g(\mathbf{x}_i, \beta)$, y un residual $\epsilon_i = y_i - \mu_i(\beta)$, en donde $\beta_{q\times 1}$, es un vector de coeficientes de regresión y $g(.\ ,\ .)$: que va de $\mathbb{R}^k \times \mathbb{R}^q \to \mathbb{R}$, $(\mathbf{x}_i, \beta) \to g(\mathbf{x}_i, \beta)$, con , $q = k+1$, puede ser conocida o desconocida según el tipo de modelo paramétrico, no-paramétrico o semi-paramétrico. El modelo de regresión paramétrico se puede definir de la siguiente forma: $y_i = g(\mathbf{x}_i, \beta) + \epsilon_i$, con $g(\mathbf{x}_i, \beta)$: conocida y $E[\epsilon_i|\mathbf{x}_i] = 0, \forall i = 1, 2, \ldots, n$,

donde la esperanza es tomada con respecto a la distribución condicional de ϵ dado $\mathbf{x}$. El modelo no-paramétrico se puede definir como $y_i = g(\mathbf{x}_i) + \epsilon_i$, con $g(\mathbf{x}_i)$: desconocida y $E[\epsilon_i|\mathbf{x}_i] = 0$.

Para realizar inferencia sobre β, en el caso paramétrico (o sobre $g(.)$, en el caso no-paramétrico), se necesitan al menos tres procedimientos estadísticos: primero, un método de estimación para calcular el estimador del vector de coeficientes β, denotado por $\hat{\beta}$. Segundo, probar que $\hat{\beta}$ es un estimador consistente de β y que tiene ciertas propiedades asintóticas, y tercero desarrollar estadísticas de pruebas para contrastar hipótesis de la forma:

$$H_0 \ : H\boldsymbol{\beta} = \mathbf{b}_0 \quad \text{v.s} \quad H_a \ : H\boldsymbol{\beta} \neq \mathbf{b}_0,$$

en donde, generalmente $H_{r\times s}$, $\beta_{s\times 1}$ y $\mathbf{b_0}_{r\times 1}$.

4.2.2. Regresión en sub-espacios de variables

En muchas ocasiones prácticas el número de variables es mucho mayor a la cantidad de observaciones disponibles en el conjunto de datos para un modelo de regresión, causando el problema de la multicolinealidad en los predictores. Entre las opciones disponibles para hacer frente a este problema se encuentran: Las técnicas basadas en sub-espacios explícitos o implícitos y el enfoque Bayesiano, el cual incluye información adicional a cerca de los parámetros del modelo. En el caso de sub-espacios, la regresión se realiza en un espacio factible de menor dimensión. El sub-espacio se puede construir de forma explícita con una motivación de tipo geométrico derivada del uso de variables latentes, o de forma implícita usando técnicas de regularización para evitar el problema de multicolinealidad en los predictores. La introducción de variables latentes nos permite capturar la información más relevante de la matriz de covariables X o información acerca de la estructura de la interacción entre X y la matriz de variables respuesta Y.

En este enfoque se introducen las variables latentes no-correlacionadas denotadas por $T_1, T_2, \ldots, T_a$ y $U_1, U_2, \ldots, U_a$. La utilización de variables latentes es una factorización de bajo rango de la matriz de predictores y/o respuestas, la cual permite ajustar un modelo de regresión lineal mediante mínimos cuadrados sobre este conjunto de variables latentes.

Los vectores X-cargas P_i y Y-cargas Q_i, generan espacios a-dimensionales donde los coeficientes Ti $_{n\times 1}$ y Ui $_{n\times 1}$ son considerados como variables latentes. Dentro de los enfoques basados en variables latentes se encuentran: la PCR y la PLSR, entre otros, los cuáles se

describen brevemente a continuación.

En PCR, la cual fue introducida por Massy 1965, las variables latentes llamadas componentes principales, son obtenidas a partir de la matriz de correlación de X, denotada por $\mathbf{R}$. La PCR evita el problema de multicolinealidad reduciendo la dimensión de los predictores. Las X-cargas $\{P_i\}_{i=1}^{k}$, son tomadas como los a-primeros eigenvectores de la descomposición espectral de la matriz $\mathbf{R}$ y dichos vectores son las direcciones que maximizan la varianza de las componentes principales, que son definidas mediante las proyecciones de las X's sobre estas direcciones, es decir, la i-ésima componente principal de X se define como $\mathrm{T}_i = XP_i$ tal que P_i maximiza la varianza de T_i,

$$\max_{P_i} \langle XP_i, XP_i \rangle = \max_{P_i} P_i^T X^T X P_i,$$

con $P_i^T P_i = 1$ y $P_i^T P_k = 0$, $k < i$. Las componentes principales representan la selección de un nuevo sistema de coordenadas obtenido al rotar el sistema original de ejes $X_1, X_2, \ldots, X_p$. A continuación, se obtienen todas las cargas o direcciones principales, $\mathrm{P}= [P_1|P_2|\cdots|P_a]_{p\times a}$ y las proyecciones de las $X_i's$ sobre las $\mathrm{P}_i's$, es decir, todas las componentes principales, $\mathrm{T}= [T_1|T_2|\cdots|T_a]_{n\times a}$, con las restricciones de que $\langle T_i, T_k \rangle = 0$ y $\langle T_i, T_i \rangle = Var(\mathrm{T}_i) = \lambda_i$, con λ_i-los eigenvalores asociados a los eigenvectores P_i con $\lambda_1 \geq \lambda_2 \geq \ldots, \lambda_a$. Luego se ajusta un modelo de regresión de Y contra las variables latentes T y se pasa a la predicción de la respuesta para Y-nuevas asociadas a nuevas observaciones del vector de predictores. En PCR, se usan las componentes principales en el espacio de los predictores X's, sin tener en cuenta la información de las respuestas Y's.

Como se mencionó en la sección 3.5, la PLSR fue introducida por Wold 1975b, para ser aplicada en ciencias económicas y sociales, pero debido a las contribuciones de su hijo Wold et al. 1984, ganó una gran popularidad en el área de la Quimiometría, en donde se analizan datos que se caracterizan por muchas variables predictores, con problemas de multicolinealidad, y pocas observaciones disponibles, lo cual sucede en muchos estudios de análisis de imágenes. La metodología de PLSR generaliza y combina características del Análisis de Componentes Principales (PCA) y Análisis de Regresión Múltiple (MLR). Su demanda y evidencia ha aumentado y está siendo aplicada en muchas ramas de la ciencia. La PlSR es similar al análisis de correlación canónica (CCA), pero en lugar de maximizar la correlación, maximiza la covarianza entre las componentes, es decir, se hallan direcciones p y q tales que

$$\max_{p,q} \langle Xp, Yq \rangle = \max_{p,q} p^T X^T Y q,$$

sujeto a, $\|p\| = \|q\| = 1$.

En general, la PLSR es un proceso de dos etapas. Primero, se transforma la matriz de predictores X, con ayuda del vector de las variables respuestas Y, en una matriz de variables latentes no correlacionadas, T= $(T_1, T_2, \ldots, T_p)$, llamados componentes PLS, lo cual lo distingue de PLSR en la cual los componentes son obtenidos usando sólo la matriz de predictores X. En segundo lugar, se ajusta el modelo de regresión estimado usando el vector de respuestas original y las componentes PLS como predictores, luego se procede con la predicción de respuestas para Y-nuevas asociadas a futuras observaciones del vector de predictores. La reducción de la dimensionalidad se obtiene directamente sobre las componentes PLS, ya que estos son ortogonales y el número de componentes necesarios para el análisis de regresión es mucho menor que el número de predictores originales. El proceso de maximizar la covarianza en lugar de la correlación, evita posibles problemas de inestabilidad numérica que pueden aparecer al usar correlación, debido a la división de las covarianzas por varianzas que pueden ser muy pequeñas. Las direcciones de máxima covarianza p y q entre las componentes PLS, se pueden hallar mediante el siguiente problema de eigen-descomposición:

$$X^T Y Y^T X p = \lambda p \quad \text{y} \quad Y^T X X^T Y q = \lambda q$$

con $\|p\| = \|q\| = 1$. Las variables latentes (o componentes PLS) son calculadas proyectando los datos X y Y en las direcciones p y q, es decir, $t = Xp$ y $u = Yq$ y luego se obtienen todas las componentes latentes T=XP y U=YQ.

4.3. Geometría de $\mathrm{Sym}^+(m)$

Ahora se da un resumen de algunos resultados básicos de Schwartzman 2006 acerca de la estructura geométrica del conjunto $\mathrm{Sym}^+(m)$ como una variedad Riemanniana, ver sección 2.3. El espacio $\mathrm{Sym}^+(m)$ es una sub-variedad del espacio euclídeo $\mathrm{Sym}(m)$. Geo métricamente, los espacios $\mathrm{Sym}^+(m)$ y $\mathrm{Sym}(m)$ son variedades diferenciables de dimensiones $m(m+1)/2$ y están homeomórficamente relacionados mediante la transformación matriz exponencial y logaritmo, ver sección 2.2. Para cualquier matriz $A \in \mathrm{Sym}(m)$, su matriz exponencial está dada por $\exp(A) = \sum_{k=1}^{\infty} \frac{A^k}{k!} \in \mathrm{Sym}^+(m)$, recíprocamente, para cualquier matriz $S \in \mathrm{Sym}^+(m)$, existe un $\log(S) = A \in \mathrm{Sym}(m)$ tal que $\exp(A) = S$. En los modelos de regresión no-paramétricos estándar para respuesta sobre espacios euclídeos se estima a $E[S|X = x]$, ver sección 3.1. Sin embargo, para respuestas sobre un espacio curvado, no se puede definir directamente la esperanza condicional de S dado $X = x$, como en el caso usual de espacios euclídeos.

Para $\mu(x) = E[S|X = x]$, se introduce un vector tangente y el espacio tangente en $\mu(x)$ sobre $\text{Sym}^+(m)$, ver sección 2.2. Para un escalar pequeño $\delta > 0$, se considera el mapa diferencible, $C : (-\delta, \delta) \longrightarrow \text{Sym}^+(m), \quad t \to C(t)$, tal que, $C(0) = \mu(x)$. Un vector tangente en $\mu(x)$ se define como la derivada de la curva suave $C(t)$ con respecto a t evaluada en $t = 0$. El conjunto de todos los vectores tangentes en $\mu(x)$ se llama espacio tangente de $\text{Sym}^+(m)$ en $\mu(x)$ y se denota por $T_{\mu(x)}\text{Sym}^+(m)$, dicho espacio se puede identificar con una copia de $\text{Sym}(m)$. El espacio $T_{\mu(x)}\text{Sym}^+(m)$ es equipado con un producto interno $\langle \ . \ , \ . \ \rangle$, llamado métrica riemanniana, la cual varia suavemente de punto a punto, ver sección 2.3. Por ejemplo, se puede usar la métrica de Frobenoius como métrica riemanniana. Para una métrica riemanniana dada, se calcula $\langle U \ , \ V \rangle$ para cualesquiera U y V en $T_{\mu(x)}\text{Sym}^+(m)$ y luego se calcula la longitud de la curva suave $C(t) : [t_0, t_1] \longrightarrow \text{Sym}^+(m)$, la cual es igual a: $\|C(t)\| = \int_{t_0}^{t_1} \sqrt{\langle \dot{C}(t), \dot{C}(t) \rangle} dt$, donde, $\dot{C}(t)$-es la derivada de $C(t)$- con respecto a t. Una geodésica es una curva suave en $\text{Sym}^+(m)$ cuyos vectores tangentes no cambian en longitud o dirección cuando uno se mueve a lo largo de la curva. Para cualquier $U \in T_{\mu(x)}\text{Sym}^+(m)$, existe una única geodésica, denotada por $\gamma_{\mu(x)}(t; U)$, cuyo dominio contiene al intervalo $[0, 1]$, tal que $\gamma_{\mu(x)}(0; U) = \mu(x)$ y $\dot{\gamma}_{\mu(x)}(0; U) = U$. El mapa exponencial riemanniano se define como

$$\text{Exp}_{\mu(x)} : T_{\mu(x)}\text{Sym}^+(m) \longrightarrow \text{Sym}^+(m) \ \ ; \ U \longrightarrow \text{Exp}_{\mu(x)}(U) = \gamma_{\mu(x)}(1; U). \tag{4.1}$$

La inversa del mapa exponencial riemanniano, llamado el log-riemanniano, se define como

$$\text{Log}_{\mu(x)} : \text{Sym}^+(m) \longrightarrow \ T_{\mu(x)}\text{Sym}^+(m) \ \ ; \ S \longrightarrow \text{Log}_{\mu(x)}(S) = U, \tag{4.2}$$

tal que, $\text{Exp}_{\mu(x)}(U) = S$. Por último la distancia más corta entre 2 puntos $\mu_1(x)$ y $\mu_2(x)$ en $\text{Sym}^+(m)$, es llamada la distancia geodésica y se denota por $g(\mu_1(x), \mu_2(x))$, la cual satisface

$$d_g^2(\mu_1(x), \mu_2(x)) = \langle \text{Log}_{\mu_1(x)}\mu_2(x), \text{Log}_{\mu_1(x)}\mu_2(x) \rangle = \|\text{Log}_{\mu_1(x)}\mu_2(x)\|_g^2. \tag{4.3}$$

Se define el residual de S con respecto a $\mu(x)$, denotado por $\varepsilon_\mu(X)$, como: $\varepsilon_\mu(x) = \text{Log}_{\mu(x)}S \in T_{\mu(x)}\text{Sym}^+(m)$. Para $C = [c_{ij}] \in \text{Sym}(m)$, se define la vectorización de C como $\text{Vecs}(C) = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1m} & c_{22} & \dots & c_{2m} & \dots & c_{mm} \end{bmatrix}^T \in \mathbb{R}^{\frac{m(m+1)}{2}}$. La esperanza condicional de S dado $X = x$, se define como la matriz $\mu(x) \in \text{Sym}^+(x)$, tal que

$$E[\text{Log}_{\mu(x)}S|X = x] = E[\varepsilon_\mu(X)|X = x] = \mathbf{0}_{m\times m}, \tag{4.4}$$

en donde la esperanza, es tomada componente a componente con respecto al vector-aleatorio multivariado $\text{Vecs}[\text{Log}_{\mu(x)}S] \in \mathbb{R}^{\frac{m(m+1)}{2}}$.

4.3.1. Modelo de Regresión Para datos Respuesta en el espacio $\mathbf{Sym}^+(m)$

Debido a que los DT's están en un espacio no lineal, es teórica y computacionalmente difícil de desarrollar un marco estadístico formal que incluya teoría de estimación y pruebas de hipótesis, en donde se use un conjunto de covariables para predecir directamente DT's como respuesta. Usando el desarrollo reciente de la métrica log-euclídea, Arsigny et al. 2006, los DT's se pueden transformar del espacio no-lineal en sus matrices logarítmicas en un espacio euclídeo. Zhu et al. 2009, desarrolló un modelo de regresión con los DT's log-transformados como respuesta. El modelo se basó en un método semiparamétrico, el cual evita especificar distribuciones paramétricas para los DT's aleatorios log-transformados. Se han planteado procesos de inferencia para estimar los coeficientes de regresión de dicho modelo, al igual que estadísticas de prueba para contrastar hipótesis lineales de los parámetros desconocidos y procesos de pruebas basados en métodos de remuestreo para evaluar simultáneamente la significancia estadística de hipótesis lineales a través de grandes ROI. A continuación se describe el procedimiento del planteamiento del modelo de regresión polinomial local intrínseco (RPLI) para matrices simétricas DP como respuesta.

Para estimar a $\mu(x) = E[S|X = x_0]$ en el modelo de RPLI, se procede de la siguiente forma. Debido a que $\mu(x)$ está sobre un espacio curvado, no se puede expandir directamente a $\mu(x)$ en $X = x_0$, usando series de Taylor. En lugar de eso, se considera el mapa Logaritmo Riemanniano de $\mu(x)$ en $\mu(x_0)$ sobre el espacio $T_{\mu(x)}\mathrm{Sym}^+(m)$, es decir, $\mathrm{Log}_{\mu(x_0)}\mu(x) \in T_{\mu(x)}\mathrm{Sym}^+(m)$. Como $\mathrm{Log}_{\mu(x_0)}\mu(x)$ está en un espacio tangente diferente para cada valor distinto de X, se pueden transportar desde $T_{\mu(x)}\mathrm{Sym}^+(m)$ al espacio tangente común $T_{I_m}\mathrm{Sym}^+(m)$, a través del transporte paralelo dado por:

$$\Phi_{\mu(x_0)} : T_{\mu(x_0)}\mathrm{Sym}^+(m) \longrightarrow T_{I_m}\mathrm{Sym}^+(m);\ \mathrm{Log}_{\mu(x_0)}\mu(x) \longrightarrow \Phi_{\mu(x_0)}(\mathrm{Log}_{\mu(x_0)}\mu(x)) = Y(x), \tag{4.5}$$

y su inversa $\mathrm{Log}_{\mu(x_0)}\mu(x) = \Phi^{-1}_{\mu(x_0)}(Y(x)) \in T_{\mu(x_0)}\mathrm{Sym}^+(m)$.

Para $\mathrm{Log}_{\mu(x_0)}\mu(x_0) = O_m \in T_{\mu(x_0)}Sym^+(m)$, debido a que $\Phi_{\mu(x_0)}(O_m) = Y(x_0) = O_m$, y como $Y(x)$ y $Y(x_0)$ están en el mismo espacio tangente $T_{I_m}\mathrm{Sym}^+(m)$, se expande a $Y(x)$ en x_0, usando series de Taylor, de donde se obtiene:

$$\mathrm{Log}_{\mu(x_0)}\mu(x) = \Phi^{-1}_{\mu(x_0)}(Y(x)) \approx \Phi^{-1}_{\mu(x_0)}\left(\sum_{k=1}^{k_0} Y^{(k)}(x_0)(x - x_0)^k\right), \tag{4.6}$$

con k_0-un entero y $Y^{(k)}$-es la k-ésima derivada de $Y(x)$ con respecto a x dividida por $k!$.

Equivalentemente se tiene que

$$\mu(x) = \mathrm{Exp}_{\mu(x_0)}\left(\Phi^{-1}_{\mu(x_0)}(Y(x))\right) = \mathrm{Exp}_{\mu(x_0)}\left(\Phi^{-1}_{\mu(x_0)}\left(\sum_{k=1}^{k_0} Y^{(k)}(x_0)(x-x_0)^k\right)\right) = \mu\left(x, \alpha(x_0), k_0\right), \tag{4.7}$$

donde, $\alpha(x_0)$-contiene todos los parámetros en: $\{\mu(x_0), Y^{(1)}(x_0), \ldots, Y^{(k)}(x_0)\}$.

Para el conjunto de vectores en $T_{\mu(x)}\mathrm{Sym}^+(m)$, se pueden definir varias métricas riemanniana, entre las cuales esta la métrica log-euclídea, para la cual ahora se revisan algunas de sus propiedades básicas.

Se usan las notaciones exp(.) y log(.), para representar las matrices exponencial y logaritmo respectivamente. Mientras que se usa Exp y Log, para representar los mapas exponencial y logaritmo riemanniano respectivamente. Se denota por $\partial_{\mu(x)}\log.(U)$-como la diferencial de la matriz logaritmo en $\mu(x) \in \mathrm{Sym}^+(m)$ que actúa sobre un desplazamiento infinitesimal $U \in T_{\mu(x)}\mathrm{Sym}^+(m)$ y se define la métrica log-euclídea sobre $\mathrm{Sym}^+(m)$ como: para $U, V \in T_{\mu(x)}\mathrm{Sym}^+(m)$

$$\langle U, V\rangle := tr\left[(\partial_{\mu(x)}\log.U)(\partial_{\mu(x)}\log.V)\right]. \tag{4.8}$$

La geodésica $\gamma_{\mu(x)}(t;U)$-esta dada por:

$$\gamma_{\mu(x)}(t;U) := \exp\left[\log(\mu(x)) + t\partial_{\mu(x)}\log.V\right] \ , \quad \forall t \in \mathbb{R}. \tag{4.9}$$

Se denota por $\partial_{\log(\mu(x))}\exp.(A)$-como la diferencial de la matriz exponencial en $\log(\mu(x)) \in \mathrm{Sym}(m) = T_{\mu(x)}\mathrm{Sym}^+(m)$ que actúa sobre un desplazamiento infinitesimal $A \in T_{\log(\mu(x))}\mathrm{Sym}^+(m)$. Los mapas exponencial y logaritmo riemanniano se definen respectivamente de la forma siguiente: para $S \in \mathrm{Sym}^+(m)$

$$\mathrm{Exp}_{\mu(x)}(U) := \exp\left[\log(\mu(x)) + \partial_{\mu(x)}\log.(U)\right] \ ; \ \mathrm{Log}_{\mu(x)}(S) := \partial_{\log(\mu(x))}\exp\left[\log(S) - \log(\mu(x))\right]. \tag{4.10}$$

Para $\mu(x), S \in \mathrm{Sym}^+(m)$, la distancia geodésica esta dada por:

$$d^2_g(\mu(x), S) := tr\left[(\log\,\mu(x) - \log(S))^{\otimes 2}\right], \tag{4.11}$$

con $\mathrm{a}^{\otimes 2} = \mathrm{aa}^T$, con a-vector. Ahora para dos matrices $\mu(x)$ y $\mu(x_0) \in \mathrm{Sym}^+(m)$ y cualquier $U_{\mu(x_0)} \in T_{\mu(x_0)}\mathrm{Sym}^+(m)$, el transporte paralelo se define como sigue:

$$\Phi_{\mu(x_0)} : T_{\mu(x_0)}\mathrm{Sym}^+(m) \longrightarrow T_{I_m}\mathrm{Sym}^+(m) \ ; \ U_{\mu(x_0)} \longrightarrow \Phi_{\mu(x_0)}(U_{\mu(x_0)}) := \partial_{\mu(x_0)}\log.\left(U_{\mu(x_0)}\right).$$

Luego con $U_{\mu(x_0)} = \text{Log}_{\mu(x_0)}\mu(x) \in T_{\mu(x_0)}\text{Sym}^+(m)$, entonces

$$Y(x) = \Phi_{\mu(x_0)}\left(\text{Log}_{\mu(x_0)}\mu(x)\right) = \log\,\mu(x) - \log\,\mu(x_0), \tag{4.12}$$

y $\mu(x) = \exp\left[\log\,\mu(x_0) + Y(x)\right]$.

El residual de S con respecto a $\mu(x)$, es definido como: $\varepsilon_\mu(X) := \log(S) - \log(\mu(x))$, y $E[\log S|X = x] = \log\,\mu(x)$, y se define el modelo de RPLI como:

$$\log(S|X) = \log(\mu(x)) + \varepsilon_\mu(X), \tag{4.13}$$

con $E[\varepsilon_\mu(X)] = 0$, es decir que, $E[\log S|X = x] = \log(\mu(x))$.

4.4. El Modelo de Regresión PLS

Suponga que tenemos $n - DT$'s, denotados por $T_i \;:\; i = 1, 2, \ldots, n$ obtenidos a partir de un voxel correspondiente de las DTI's normalizadas y re-orientadas espacialmente de n-sujetos. A continuación se obtiene la log-transformación de los T_i, que se denotan por $\log(T_i) = L^i_{T_{(j,k)}}$, y un vector 6-dimensional dado por:

$$L_{T,i} = (L^i_{T_{(1,1)}}, L^i_{T_{(1,2)}}, L^i_{T_{(1,3)}}, L^i_{T_{(2,2)}}, L^i_{T_{(2,3)}}, L^i_{T_{(3,3)}})^T, \tag{4.14}$$

donde, $L^i_{T_{(j,k)}}$ −denota el (j,k)-elemento de la matriz logaritmo de T_i. Para cada individuo, se observa un conjunto de covariables de interés.

En estudios de imágenes médicas generalmente se consideran muchas medidas demográficas o clínicas sobre los distintos pacientes considerados en un cierto estudio, en donde la cantidad de información disponible es muy grande, presentándose posibles problemas de dependencias lineales entre las covariables de interés, que origina el problema de multicolinealidad, y además generalmente se cuenta con pocos individuos disponibles para el análisis de la información. Para los DT's log-transformados, se considera un modelo lineal dado por:

$$\underset{1\times 6}{L_{T,i}} = \underset{1\times pp\times 6}{\mathbf{x}_i\ \beta} + \underset{1\times 6}{\varepsilon_i}\ , \quad i = 1, 2, \ldots, n \tag{4.15}$$

o

$$\underset{n\times 6}{\mathbf{L}_T} = \underset{n\times pp\times 6}{\mathbf{X}\ \mathbf{B}} + \underset{n\times 6}{\boldsymbol{\varepsilon}}\ . \tag{4.16}$$

con $E[\boldsymbol{\varepsilon}|\mathbf{x}] = \mathbf{0}_{n\times p}$ y $\text{Cov}(\boldsymbol{\varepsilon}|\mathbf{x}) = \Sigma_{np\times np}$, donde $\mathbf{X}$, $\mathbf{Y}$=$\mathbf{L}$, $\mathbf{B}$, y $\boldsymbol{\varepsilon}$, son matrices que

representan las covariables, respuestas, coeficientes de regresión y errores del modelo. Comparado con los modelos lineales generales, el modelo de (4.16), basado en la media condicional y covarianza, no asume suposiciones distribucionales para las medidas de imágenes.

Si $\boldsymbol{\theta}_{(6p+21)\times 1}$-es el vector de parámetros desconocidos contenidos en $\boldsymbol{\beta}$ y Σ, entonces para estimar a $\boldsymbol{\theta}$- se maximiza la función objetivo dada por

$$l_n(\boldsymbol{\theta}) = -\frac{1}{2}\sum_{i=1}^{n}\left(\log|\Sigma|(L_{T,i} - \beta\mathbf{x}_i)^T\Sigma^{-1}(L_{T,i} - \beta\mathbf{x}_i)\right), \tag{4.17}$$

utilizando el algoritmo iterativo planteado por Li et al. 2009 para obtener a θ.

El modelo de regresión 4.16 se ha ajustado usando los algoritmos que existen para PCR y para PLSR, siguiendo los pasos descritos en la sección 2.2, y teniendo en cuenta las log-transformaciones realizadas a los datos originales para llevarlos a un espacio euclídeo. En este espacio euclídeo el modelo de regresión 4.16 hereda las propiedades del modelo de regresión PLS estándar para datos datos no transformados sobre espacios euclídeos, lo cual puede ser justificado mediante el regreso a las variables respuestas originales a partir de los predictores igualmente regresados a sus valores originales.

4.4.1. Evaluación del modelo de Regresión PLS mediante Datos simulados

Mediante conjuntos de datos simulados se evalúan los comportamientos del modelo de regresión PLS, comparando sus resultados de predicción con los arrojados por la técnica de PCR en el caso de matriz diseño de rango incompleto y con el modelo de regresión clásico, en el caso de una matriz diseño de rango completo.

Los entornos considerados para simular los datos fueron los siguientes. Primero se simuló una muestra de matrices simétricas PD de tamaño n=20 con k=15 covariables generadas de una distribución normal multivariada con media cero y estructura de covarianza dada por $\Sigma = 0.6I_6$, luego se incrementó el tamaño muestral a n=30 y se incrementó el número de covariables de k=15 a k=45 con una estructura de covarianza dada por $\Sigma = 0.3I_6 + 0.61_n1_n^T$, con 1_n: vector de unos. En ambos entornos se utilizaron valores para los coeficientes de betas datos por la matriz de orden $p \times 6$, $\beta_k = [1 + 0.1 \times (k-1)]^T$, se calculó la exponencial de Σ para asegura la definidez positiva de esta.

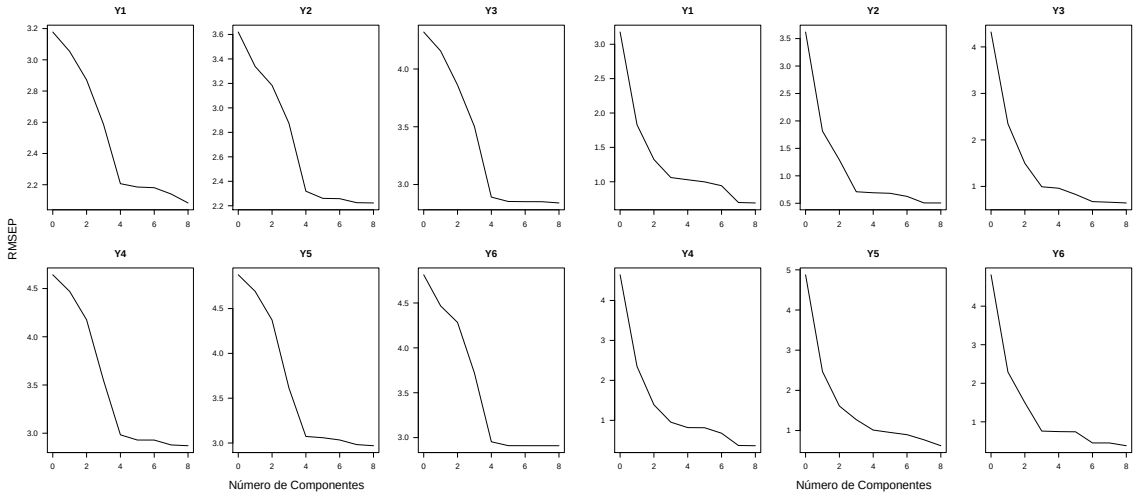
A continuación se exponen los resultados obtenidos en cada uno de los escenarios considerados en el estudio simulado que se llevo a cabo. Para el primer entorno se tiene en la Tabla uno, los porcentajes de varianza explicados por cada una de las componentes latentes mediante PCR y PLSR, se observa que PCR explica más de la variabilidad de X que PLSR, lo cual siempre sucede, mientras que al observar la Tabla dos, se tiene que las componentes PLS explican un mayor porcentaje de la variabilidad de Y que las componentes PCR, con dos componentes se alcanza mas del 80 % de la variabilidad en Y, y aproximadamente un 20 % de la variabilidad en X. En la figura uno, se tienen las gráficas de la raíz cuadrada del error cuadrático medio de predicción (RMSEP) contra el número de componentes usando validación cruzada (CV), a partir de la cual se puede observar que en PCR se necesitarían alrededor de 4 componentes para explicar la mayor parte de la variabilidad de los datos mientras que en PLSR se necesitan 3 componentes en la mayoría de los casos. Aunque en general en esta ilustración hay poca diferencia entre los resultados obtenidos por ambos métodos, lo cual se debe al proceso de simulación llevado a cabo. En la figura dos, se observan gráficas de los datos predichos junto a los valores observados de las respuestas, observándose un mayor precisión en el ajuste cuando se utiliza PLSR. Para el segundo entorno se tiene en a Tabla tres, se tienen los porcentajes de varianza explicados por cada una de las componentes latentes mediante PCR y PLSR, nuevamente se tiene que PCR explica más de la variabilidad de X que PLSR, y en la Tabla cuatro, se tiene que las componentes PLS explican un mayor porcentaje de la variabilidad de Y que las componentes PLS, con cinco componentes se alcanza mas del 60 % de la variabilidad en Y, y aproximadamente un 35 % de la variabilidad en X. En la figura tres, se tienen las gráficas de los RMSEP contra el número de componentes, y se puede observar que en PCR se necesitarían alrededor de 7 componentes para explicar la mayor parte de la variabilidad de los datos mientras que mediante PLSR se necesitan 5 componentes en la mayoría de los casos. En la figura cuatro, se observan gráficas de los datos predichos junto a los valores observados de las respuestas, observándose un mayor precisión en el ajuste cuando se utiliza PLSR.

Tabla 4.1: Porcentaje de Variabilidad de X explicada por cada componente.

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7	Comp 8
PCR	17.57	15.55	13.59	12.46	11.16	9.16	6.81	4.64
PLSR	14.27	9.93	10.16	13.45	12.60	5.75	4.46	7.07

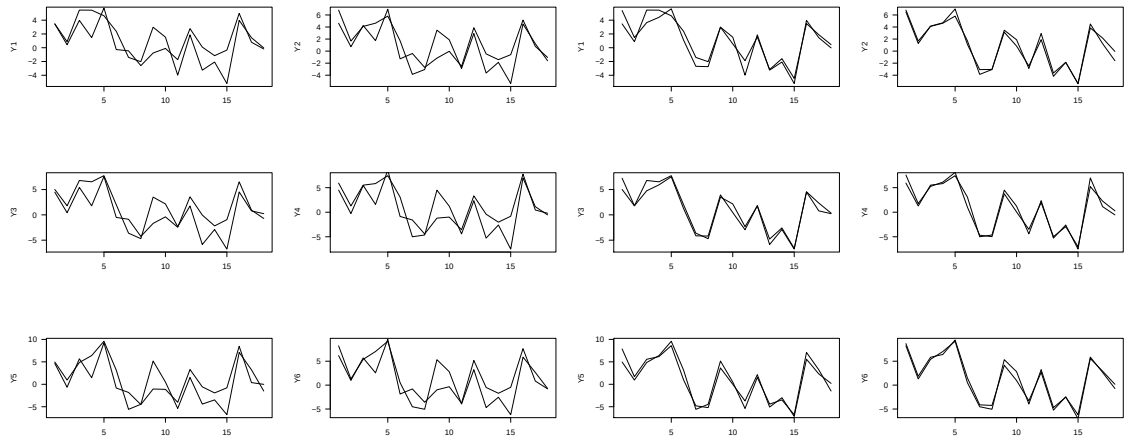
Tabla 4.2: Porcentajes de Varianza explicada acumuladas de X y Y por las componentes seleccionadas mediante PCR y PLSR.

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps	8 comps
X	17.57	33.11	46.70	59.16	70.32	79.48	86.28	90.93
X	14.27	24.20	34.37	47.82	60.43	66.17	70.64	77.70
Y1	7.69	18.38	33.74	51.79	52.72	52.91	54.64	57.04
Y1	66.85	82.64	88.85	89.51	90.13	91.21	95.17	95.26
Y2	14.95	22.65	36.98	58.96	60.99	61.09	62.21	62.29
Y2	74.87	87.30	96.16	96.35	96.46	97.01	98.04	98.05
Y3	7.45	20.12	34.30	55.21	56.38	56.43	56.44	56.77
Y3	70.51	88.00	94.72	95.05	96.33	97.57	97.67	97.78
Y4	7.30	19.10	41.57	58.71	60.19	60.20	61.57	61.78
Y4	74.39	91.05	95.78	96.90	96.92	97.87	99.36	99.39
Y5	7.44	19.65	45.13	60.30	60.66	61.30	62.61	62.93
Y5	74.38	89.10	93.22	95.70	96.19	96.62	97.51	98.38
Y6	13.89	20.83	40.31	62.35	63.45	63.46	63.46	63.47
Y6	77.35	90.32	97.51	97.60	97.63	99.12	99.12	99.38



(a) PCR **(b)** PLSR

Figura 4.1: RMSEP v.s Número de componentes mediante PCR y PLSR



(a) PCR **(b)** PLSR

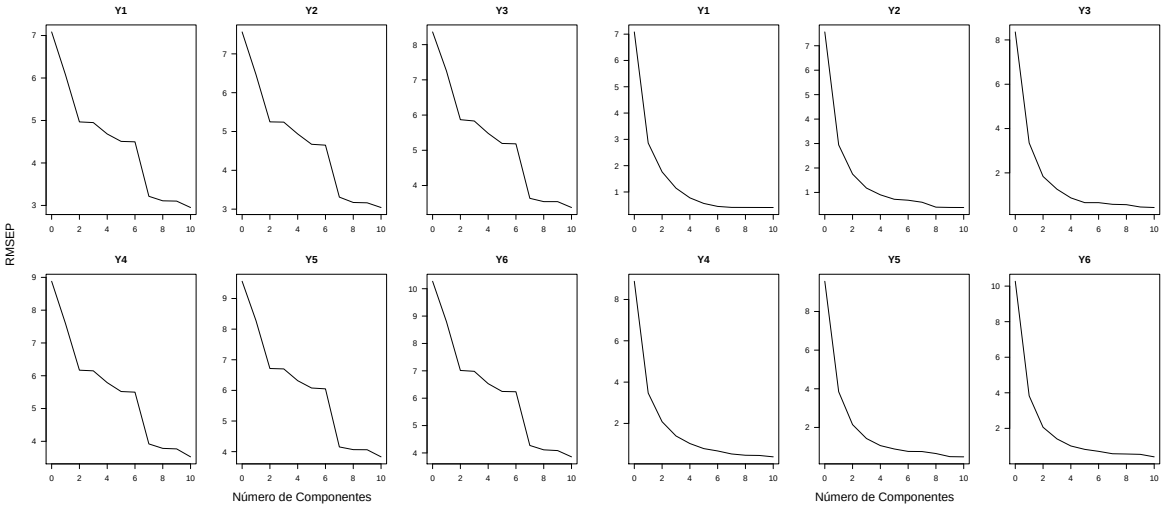
Figura 4.2: Gráfico de Valores Predichos Junto a valores observados mediante PCR y PLSR.

Tabla 4.3: Porcentaje de Variabilidad de X explicada por cada componente, entorno 2.

	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7	Comp8	Comp9	Comp10
PCR	12.81	9.33	8.74	7.42	7.22	6.39	6.33	5.12	4.97	4.44
PLSR	10.63	8.65	6.39	5.21	3.85	5.34	4.88	5.36	5.32	5.00

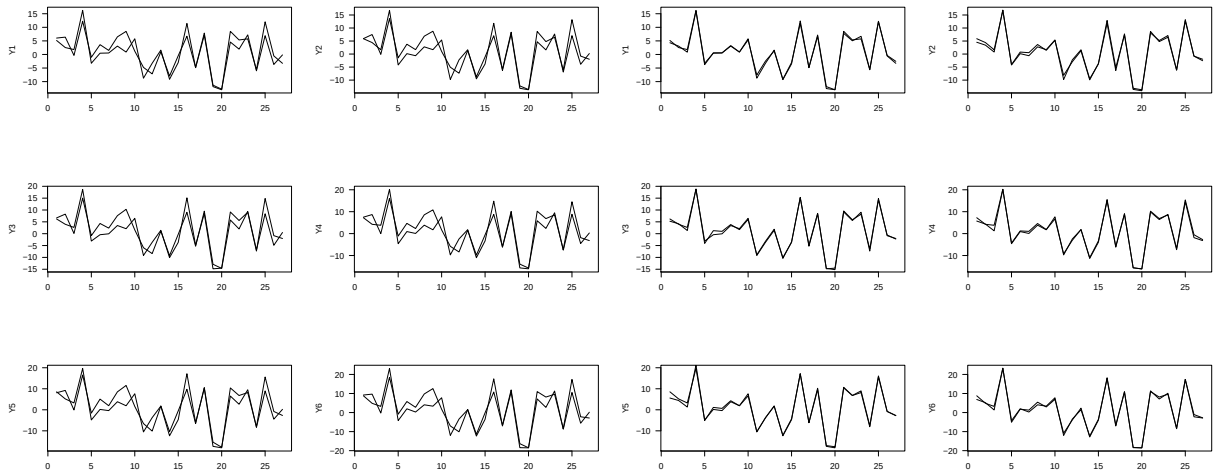
Tabla 4.4: Porcentajes de Varianza explicada acumuladas de X y Y por las componentes seleccionadas mediante PCR y PLSR, entorno 2.

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps	8 comps	9 comps	10 comps
X	12.81	22.14	30.88	38.30	45.52	51.90	58.23	63.35	68.33	72.77
X	10.63	19.28	25.67	30.88	34.73	40.07	44.95	50.32	55.64	60.64
Y1	26.52	50.85	51.17	56.31	59.51	59.69	79.40	80.74	80.83	82.65
Y1	83.70	93.81	97.39	98.80	99.37	99.59	99.66	99.66	99.66	99.67
Y2	26.97	51.87	51.99	57.41	61.87	62.25	80.86	82.42	82.52	83.83
Y2	84.85	94.65	97.57	98.58	99.09	99.19	99.37	99.72	99.74	99.74
Y3	24.82	50.72	51.34	57.02	61.40	61.56	81.08	82.05	82.05	83.70
Y3	83.92	95.16	97.72	98.91	99.38	99.38	99.52	99.54	99.70	99.73
Y4	27.00	51.74	52.05	57.50	61.39	61.65	80.51	81.84	81.99	84.23
Y4	84.74	94.50	97.54	98.67	99.23	99.44	99.66	99.73	99.74	99.81
Y5	25.11	50.70	50.90	56.36	59.61	59.97	81.14	81.93	81.96	83.96
Y5	83.80	94.97	97.77	98.77	99.14	99.37	99.38	99.54	99.74	99.75
Y6	26.75	53.38	53.80	59.58	63.02	63.15	82.70	83.96	84.18	85.90
Y6	86.10	95.97	98.12	99.03	99.37	99.53	99.69	99.71	99.73	99.85



(a) PCR **(b)** PLSR

Figura 4.3: RMSEP v.s Número de componentes mediante PCR y PLSR, entorno 2.



(a) PCR **(b)** PLSR

Figura 4.4: Gráfico de Valores Predichos Junto a valores observados mediante PCR y PLSR, entorno 2.

En este estudio simulado se ha ajustado en modelo de regresión PLS lineal para estudiar la relación entre un conjunto grande de predictoras de interés con un conjunto de variables respuesta que viven en un espacio simétrico Riemanniano, para lo cual se ha utilizado la teoría de mapas Exponenciales y Riemannianos para transformar los datos del espacio no euclídeo al espacio euclídeo de matrices simétricas, en donde se ha desarrollado la metodología. Los resultados ha mostrado un apoyo a la metodología propuesta debido a que ha arrojado mejores predicciones de la respuesta en comparación a la técnica de regresión por componentes principales, como sucede en situaciones clásicas de análisis de datos en espacios euclídeos con matrices de predictoras que presentan alta multicolinealidad o problemas de bajo número de observaciones y muchas predictoras. Los resultados han apoyado el hecho de que es fundamental tener buenas predicciones del vector de variables respuesta, que en este caso no esta dado por un escalar en cada voxel de la imágen, sino por una matriz simétrica PD 3×3 que representa a la matriz de varianzas covarianzas del movimiento aleatorio Bronwiano que simula al movimiento de las moléculas de agua a través del cerebro en cada uno de los voxeles correspondiente de la imágen completa sobre ciertas regiones de interés. Esta matriz simétrica PD 3×3 frecuentemente se representa de forma geométrica mediante un elipsoide tri-dimensional.

Capítulo 5

Conclusiones y Trabajos Futuros

En esta investigación se propuesto un modelo de regresión PLS lineal para estudiar la relación entre un conjunto grande de predictoras de interés que viven en un espacio euclídeo con un conjunto de variables respuesta que viven en una variedad Riemanniana, o más exactamente en un Espacio Simétrico Riemanniano, para lo cual se ha utilizado la teoría de mapas Exponenciales y Riemannianos para transformar los datos del espacio no euclídeo al espacio euclídeo de matrices simétricas, en donde se ha desarrollado la metodología. Los resultados han mostrado un apoyo a la metodología propuesta en comparación a la técnica de regresión por componentes principales, como sucede en situaciones clásicas de análisis de datos en espacios euclídeos con matrices de predictoras que presentan alta multicolinealidad o problemas de bajo número de observaciones y muchas predictoras. Para trabajos futuros esperamos plantear modelos más realistas, como por ejemplo modelos PLS no lineal para este tipo de datos matrices simétricas PD y otros tipos de datos variedad valuados como lo son, los datos obtenidos mediante representaciones geométricas de objetos vía la representación medial axial (m-rep), el grupo de rotaciones ortogonales, entre otros. La ilustración presentada en esta tesis para datos simulados, nos da una luz de cómo se podrían comportar este tipo de modelos en aplicaciones de datos reales y nos orienta a seguir trabajando en esta dirección. Como trabajo futuro inmediato se tiene la ampliación y aplicación de este tipo de modelos a datos reales los cuáles generalmente presentan algún tipo de dificultad a la hora de obtenerlos pero que con algo paciencia y dedicación es posible lograr obtener en distintas entidades del sector de la salud.

Referencias

Arsigny, V., P. Fillard, X. Pennec, and N. Ayache. 2006. "Log-euclidean metrics for fast and simple calculus on diffusion tensors." *Magnetic Resonance in Medicine,* 56:411–421.

Auslander, L., and R. E. MacKenzie. 1977. *Introduction to Differentiable Manifolds.* Dover.

Barmpoutis, A., B. C. Vemuri, T. M. Shepherd, and J. R. Forder. 2007. "Tensor splines for interpolation and approximation of dt-mri with applications to segmentation of isolated rat hippocampi." *IEEE Transations on Medical Imaging,* 26:1537–1546.

Batchelor, P., M. Moakher, D. Atkinson, F. Calamante, and A. Connelly. 2005. "A rigorous framework for diffusion tensor calculus." *Magnetic Resonance in Medicine,* 53:221–225.

Boothby, W. M. 1986. *An Introduction to Differentiable Manifolds and Riemannian Geometry.* Academic Press.

Burnham, A. J., R. Viveros, and J. F. MacGregor. 1996. "Frameworks for latent variable multivariate regression." *Journal of Chemometrics,* 10:31–45.

Curtis, M.L. 1984. *Matrix Groups.* Springer-Verlag.

Duistermaat, J. J., and J. A. Kolk. 2000. *Lie Groups.* Springer.

Fingelkurts, A. A., and S. Kahkonen. 2005. "Functional connectivity in the brain-is it an elusive concepts?" *Neuroscience and Biobehavioral Reviews,* 28:827–836.

Fletcher, P. T., and S. Joshi. 2007. "Riemannian geometry for the statistical analysis of diffusion tensor data." *Signal Processing,* 87:250–262.

Garthwaite, H. Paul. 1994. "An interpretation of Partial Least Square." *American Statistical Association,* 89:122–127.

González, J., D. Peña, and R. Romera. 2009. "A robust partial least squares method with applications." *Journal of Chemometrics,* 23:78–90.

Grenander, U., and M. I. Miller. 1998. "Computational anatomy: an emerging discipline." *Quarterly of Applied Mathematics,* 56:617–694.

Hall, B. C. 2003. *Lie groups, Lie algebras, and representations: an elementary introduction*. Springer-Verlag.

Höskuldsson, A. 1988. "PLS Regression Methods." *Journal of Chemometrics,* 2:211–228.

———. 2001. "Causal and path modelling." *Chemometrics and Intelligent Laboratory Systems,* 58:287–311.

Helgason, S. 1978. *Differential Geometry, Lie Groups, and Symmetric Spaces*. Academic Press,.

Herstein, I. N. 1975. *Topics in Algebra*. John Wiley and Sons.

Kawakubo, K. 1991. *The Theory of Transformation Groups.* Oxford University Press.

Kim, P. T., and D. S. Richards. 2010. "Deconvolution density estimation on spaces of positive definite symmetric matrices." *IMS Lecture Notes Monograph Series. A Festschrift of Tom Hettmansperger*.

Lee, J. M. 1997. *Riemannian Manifolds: An Introduction to Curvature.* Springer.

Lepore, N., C. A. Brun, Y. Chou, M. Chiang, R. A. Dutton, K. M. Hayashi, E. Luders, O. L. Lopez, H. J. Aizenstein, A. W. Toga, J. T. Becker, and P. M. Thompson. 2008. "Generalized tensor-based morphometry of hiv/aids using multivariate statistics on deformation tensors." *IEEE Transactions in Medical Imaging,* 27:129–141.

Li, Y., Zhu H., Y. Chen, J. G. Ibrahim, H. An, W. Lin, C. Hall, and D. Shen. 2009. "RADTI: Regression Analysis of Diffusion Tensor Images." *Medical Imaging*, vol. 7259.

Martens, H. 2001. "Reliable and relevant modelling of real world data: a personal account of the development of PLS Regression." *Chemometrics and Intelligent Laboratory Systems,* 58:85–95.

Martens, H., and T. Naes. 1989. *Multivariate Calibration*. John Wiley & Sons.

Martin, H., F. Westad, D. Folkenberg, and M. Martens. 2001. "Analysis of designed experiments by stabilised PLS Regression and jack-knifing." *Chemometrics and Intelligent Laboratory Systems,* 58:151–170.

Massy, W. F. 1965. "Principal Components Regression in Exploratory Statistical Research." *Journal of the American Statistical Association,* 64:234–246.

McIntosh, A.R., and Lobaugh N.J. 2004. "Partial least squares analysis of neuroimaging data: applications and advances." *Neuroimage,* 23:250–263.

Milnor, J. W. 1997. *Topology from the Differentiable Viewpoint.* Princeton University Press.

Munkres, J. R. 1975. *Topology: A First Course.* Prentice-Hall.

Pennec, X., P. Fillard, and N. Ayache. 2006. "A riemannian framework for tensor computing." *International Journal of Computer Vision,* 66:41–66.

Rojo, J., and S. Tuan. Nguyen. 2009. "Dimension Reduction of Microarray Data in the Presence of a Censored Survival Response: A Simulation Study." *Statistical Applications in Genetics and Molecular Biology,* 8 (1): 1–40.

Rudin, W. 1976. *Principles of Mathematical Analysis.* McGraw-Hill.

Schwartzman, A. 2006. "Random ellipsoids and false discovery rates: Statistics for diffusion tensor imaging data." Ph.D. diss., Stanford University.

Spivak, M. 1999. *A Comprehensive Introduction to Differential Geometry.* Publish or Perish.

Tenenhaus, M. 1998. *La Regression PLS, Theorie et Pratique.* Editions Technip.

Wold, H. 1975a. "Soft Modeling by Latent Variables; the Non-linear Iterative Partial Least Squares Approach." *Perspectives In Probability and Statistics,*, pp. 1–2.

———. 1975b. "Soft Modeling by Latent Variables; the Non-linear Iterative Partial Least Squares Approach." *Perspectives In Probability and Statistics,*, pp. 1–2.

———. 1985. "Partial least squares." *Encyclopedia of Statistical Sciences,* 6:581–591.

Wold, H. 1982. "Estimation of Principal Components and Related Models by Iterative Least Squares." *In Krishnaiah, P (ed.), Multivariate Analysis, Academic Press, New York,*, pp. 391–420.

Wold, S. 2001. "Personal memories of the early PLS development." *Chemometrics and Intelligent Laboratory Systems,* 58:109–130.

Wold, S., C. ALbano, III W. J. Dunn, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg, and M. Sjöström. 1984. "Multivariate Data Analysis in Chemestry, in Chemometrics, Mathematics ans Statistics in Chemestry." *Reidel Publishing Company, Dordrecht,*, pp. 17–18.

Wold, S., M. Sjöström, and L. Eriksson. 2001. "PLS-Regression: a basic tool of chemometrics." *Chemometrics and Intelligent Laboratory Systems,* 58:109–130.

Wold, S., J. Trygg, A. Berglund, and H. Anti. 2001. "Some recent developments in PLS modelling." *Chemometrics and Intelligent Laboratory Systems,* 58:83–84.

Worsley, K.J. 1997. "An overview and some new developments in the statistical analysis of PET and fMRI data." *Human Brain Mapping,* 5:254–258.

Referencias

Yuan, Y., H. Zhu, W. Lin, and J. S. Marron. 2012. "Local polynomial regression for symmetric positive-definite matrices." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 74.

Zhu, H. T., Y. S. Chen, J. G. Ibrahim, Y. M. Li, and W. L. Lin. 2009. "Intrinsic regression models for positive-definite matrices with applications to diffusion tensor imaging." *Journal of the American Statistical Association* 104:1203–1212.