Centro de Investigación en Matemáticas A.C

ANÁLISIS ESTADÍSTICO DE TRAYECTORIAS SOBRE LA ESFERA: UN CASO DE ESTADÍSTICA SOBRE VARIEDADES

T E S I S

QUE PARA OBTENER EL GRADO DE:

MAESTRO EN CIENCIAS CON ESPECIALIDAD EN PROBABILIDAD Y ESTADÍSTICA

PRESENTA:

LILIA KAREN RIVERA ESCOVAR

DIRECTOR DE TESIS:

DR. MIGUEL NAKAMURA SAVOY

2016

Datos del jurado.

1. **Datos del tutor.**
   Dr. Miguel Nakamura Savoy.
   Institución: CIMAT.
   Departamento: Probabilidad y estadística.

2. **Datos del sinodal 1.**
   Dr. Rolando Biscay Lirio
   Institución: CIMAT
   Departamento: Probabilidad y estadística

3. **Datos del sinodal 2.**
   Dr. Luis Hernández Lamoneda
   Institución: CIMAT
   Departamento: Matemáticas básicas

Datos del trabajo escrito.

Análisis estadístico de trayectorias sobre la esfera: un caso de estadística sobre variedades.
118 págs.
2016.

De pronto tuve conciencia de que ese momento,
de que esa rebanada de cotidianidad,
era el grado máximo de bienestar, era la Dicha.
Nunca había sido tan plenamente feliz
como en ese momento...
Mario Benedetti, La tregua.

Agradecimientos

A mis padres y hermano, los cuales siempre me recibieron con los brazos abiertos y me ayudaron a leventarme en los momentos más difíciles de mi vida. Siempre han sido y serán mi fuente de inspiración. Los amo con toda el alma.

Al Dr. Miguel Nakamura, por haber asesorado la presente tesis y trabajar conmigo el desarrollo y entendimiento de una parte de la teoría estadística sobre variedades, particularmente la que referió al análisis estadístico de trayectorias. De esa misma forma le agredezco sus invaluables consejos académicos y personales, porque siempre fue más allá de su labor como académico y docente.

A los sinodales Rolando Biscay y Luis Hernández, por sus observaciones y comentarios que enriquecieron y refinaron la teoría desarrollada en el presente trabajo. Principalmente le agradezco a Luis Hernández su tiempo, paciencia y conocimientos, pues desde un principio me ayudó a asentar y delimitar la teoría concerniente a geometría diferencial.

A los Doctores Rogelio Ramos, Victor Rivero, Juan Carlos Pardo, Enrique Villa, Johan Van Horebeek y Daniel Hernández ya que cada uno de ellos de distinta manera me escuchó, apoyó, animó y brindó su ayuda académica siempre que lo requerí. De manera especial agradezco al Dr. Rogelio Ramos quien fuese mi tutor durante la maestría, así como al Dr. Victor Rivero el cual fungió como mi asesor de tesis en la licenciatura y mi tutor en la especialidad.

A todos los profesores del CIMAT que me impartieron clases, gracias por formarme como persona, estudiante y profesionista, por dejar un pedazo de su sabiduría y conocimiento en mí. Al CIMAT, el cual me dio la oportunidad de hacer una maestría y me ofreció un pedazo de primer mundo, por permiterme conocer a investigadores de talla internacional, los cuales siempre me mostraron la belleza de las matemáticas puras y aplicadas.

Al Consejo Nacional de Ciencia y Tecnología, CONACYT, por darme todas las facilidades económicas para poder realizar mis estudios de posgrado.

A Dolores Aguilera, Claudia Vega, Eduardo Aguirre y Jannet Vega, los cuales representan al departamento de servicios escolares del CIMAT, gracias por tenerme toda la paciencia del mundo para aclarame dudas administrativas y apoyarme con el proceso de titulación.

A mi pequeña tertulia conformada por Manuel Pedraza, Emmanuel Ambriz, Germán Ayala, Rodrigo Hernández, Héctor Juárez y Gerónimo Rojas. Muchas gracias muchachos por haber formado parte de mi vida y haberme permitido ser parte de la suya, por todos los inigualables y preciosos momentos que transcurrieron a su lado. De esa misma forma agradezco a Miguel Pluma y César de Alba el haber compartido conmigo buenos y malos momentos, ser mis confidentes y consejeros.

A mi compañero Jorge Dávila quien me auxilió con sus conocimientos en todo lo que requerí para el entendimiento y desarrollo de la parte que refiere a geometría diferencial abordada en la presente tesis.

A Jessica Pérez y Delia Avellaneda por ser mis amigas y estar conmigo a lo largo de diez años; son las mejores amigas que cualquiera pudiera desear, las adoro.

A todas las personas que están y estuvieron en mi vida, gracias por todas las experiencias vividas.

2-dic-2015.

Índice general

Índice de figuras

Resumen

El análisis estadístico sobre variedades es un tema de actualidad que se encuentra en la frontera de la estadística moderna, principalmente por las diversas aplicaciones que comprende. Ejemplos diversos se han desarrollado recientemente en el área de medicina y de biología, así como en otras ramas de la ciencia (ver Nielsen & Barbaresco [2015], *Geometric Science of Information*, Springer). Sin embargo, el asunto presenta ciertas dificultades teóricas, en virtud de que la metodología de $\mathbb{R}^n$ no es aplicable. Esto es consecuencia de la estructura del espacio en donde se encuentran los datos de interés. Por consiguiente, se plantea el reto de extender las nociones estadísticas—tanto básicas como avanzadas—y con ello habilitar un proceso de inferencia formal basado en datos que se presentan en estos espacios.

La temática de análisis estadístico sobre variedades es extensa, pues abarca nociones de probabilidad y estadística formales tales como modelos probabilísticos con soportes no convencionales, teoría asintótica, estimadores, y otros. La exposición planteada en la tesis se especializará en el análisis estadístico de trayectorias sobre variedades riemannianas, con un enfoque desarrollado principalmente sobre la esfera. Para fines de incursionar en la temática, se analizó el artículo de Su et al. [2014] titulado "Statistical analysis of trajectories on Riemannian manifolds: bird migration, hurricane tracking and video surveillance", *The Annals of Applied Statistics*, 530–552. Éste proporciona una visión actual de la temática así como nuevas herramientas de modelación, además de poner en práctica la teoría que desarrolla a datos reales.

El presente trabajo proporcionará una breve motivación y una introducción al análisis estadístico sobre variedades, con la finalidad de afianzar la noción e importancia de esta temática. La tesis plantea constituirse en un recurso de primera instancia para acceder a algunos conocimientos de geometría diferencial. Aunado a lo anterior, conceptualizará un resumen *in extenso* del artículo base, complementado con algunos ejemplos de simulación de trayectorias de huracanes. Finalmente, se aportarán comentarios críticos acerca de la metodología propuesta por el artículo base y se identificarán bibliografía y temas indispensables para el entendimiento de esta rama de la estadística.

Keywords: Variedades, variedades riemannianas, análisis estadístico sobre variedades, análisis estadístico de trayectorias, warping function, time warping, trayectorias de huracanes, esfera.

Capítulo 1

Introducción al análisis estadístico sobre variedades

A lo largo de la historia el ser humano ha intentado entender el entorno que le rodea, con la finalidad de poder hacer pronósticos y tener herramientas para la mejora de toma de decisiones desde ámbitos sociales hasta ambientales. Es por lo anterior que se ha dado a la tarea, particularmente en los últimos años, de analizar datos "comunes" con otras perspectivas, pues se ha percatado de que hay datos que en sí mismos poseen cierta complejidad y por ende ha visto la necesidad de tratarlos con teoría distinta a la que se conoce para $\mathbb{R}^n$. Estos datos, los cuales son conocidos como datos complejos, se caracterizan por ser elementos de espacios más abstractos que el n-dimensional.

Un juego de datos puede ser complejo por la dimensión que posee o por el espacio donde yacen sus elementos. El primer caso se puede ejemplificar cuando el número de variables es mayor que el número de datos; es decir $n << p$, donde n representa el número de datos y p el número de variables. Este caso también es conocido como un problema de altas dimensiones y puede ser consultado a fondo en Hastie et al. [2009], que es considerado uno de los pioneros en esta temática. El segundo caso se puede ejemplificar con datos que son funciones y por tanto yacen en el espacio de funciones, los cuales también son conocidos como datos funcionales y las técnicas para su análisis se abrevian como FDA, por sus siglas en inglés. Estos datos pueden ser abordados en una primera instancia en Ramsay [2006], que es considerado el libro base para su tratamiento. Otro ejemplo de datos que son complejos por la estructura del espacio en el que yacen son los datos topológicos, los cuales se encuentran ligados con una nueva rama de la estadística conocida como TDA, por sus siglas en inglés. El análisis topológico de datos trata principalmente de describir el espacio en el que se encuentran los datos; por ejemplo, describir si el espacio en el que se está trabajando tiene hoyos y de ser así cuántos tiene. Un artículo considerado seminal, el cual introdujo y motivó esta temática fue Carlsson [2009]. La complejidad que introducen los datos expuestos anteriomente, radica en el hecho de que las técnicas usuales de estadístca, como por ejemplo PCA, no funcionan o bien no son pertinentes. Por lo tanto imponen el reto de desarrollar, analizar y extender nueva teoría con la cual puedan ser estudiados.

Esta tesis versará principalmente sobre el análisis estadístico de datos que se caracterizan por ubicarse en espacios no lineales. De esta forma, el objetivo del presente capítulo es introducir, motivar y exhibir la importancia de estudiar este tipo de datos desde un enfoque estadístico y probabilístico, así como dar un breve esbozo de lo que tratarán los capítulos subsecuentes de la tesis.

1.1. Motivación al análisis estadístico sobre variedades

En esta sección se incentivará la importancia y complejidad que pueden tener los datos que son complejos por el espacio en el que se encuentran. Para lograr lo anterior se abordarán algunos ejemplos, en los cuales se exhiba la complejidad del dato y el interés por parte de algún agente en tratar de estudiarlo e interpretarlo.

- Dirección del movimiento de tortugas.
  Supóngase que se tiene un grupo de treinta y seis tortugas terrestres, como las de la Figura 1.1, las cuales después de haber desovado toman alguna dirección, tal y como se muestra en la Figura 1.2. El dato que se registra en este caso es la dirección que toma cada tortuga, de tal forma que ésta también se puede ver como un punto sobre el círculo unitario. Por tanto, el dato con el que se trabaja no es lineal, en el sentido de que se encuentra en un espacio cuya curvatura es distinta de cero y por ende no posee la estructura de un espacio vectorial. Esto último implica que el dato no pertenece a $\mathbb{R}$, ya que la curvatura del círculo no es cero.

  Analizar las direcciones de las tortugas es de interés para los biólogos, ya que en general el estudio de patrones de conducta conduce a un mejor entendimiento de especies y la relación con su entorno. Los datos comentados en este ejemplo se conocen como datos circulares o datos direccionales y se puede conocer más de ellos en Fisher [1995] y Mardia and Jupp [1999].

Figura 1.1: Tortuga terrestre.

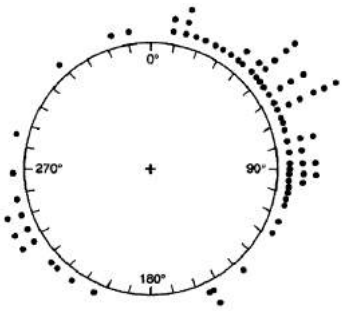


Figura 1.2: Direcciones de 36 tortugas.

- Temperatura de la superficie marina.
  Se mide con la dupla $({}^oC, (\theta, \phi))$, donde ${}^oC$ denota la temperatura del mar en

la superficie y (θ, ϕ) representan la posición geográfica—latitud, longitud—del lugar en el que se está midiendo la temperatura. En este ejemplo el dato consiste de $({}^oC, (\theta, \phi))$ y pertenece al espacio $\mathbb{R} \times S^2$. Esto se debe a que ${}^oC \in \mathbb{R}$ por ser una medición numérica, mientras que (θ, ϕ) está en S^2—que es la esfera unitaria—por ser un posicionamiento geográfico. Por lo tanto, de acuerdo con los argumentos anteriores, el dato no es lineal ya que $({}^oC, (\theta, \phi)) \notin \mathbb{R} \times \mathbb{R} = \mathbb{R}^2$.

Es importante para metereólogos y oceanólogos cuantificar la temperatura del mar, pues son los mares y oceános los que moderan la temperatura de la Tierra. La Figura 1.3 muestra la temperatura promedio de la superficie marina alrededor del mundo, para ilustrar el tipo de variación que se menifiesta con este tipo de datos.

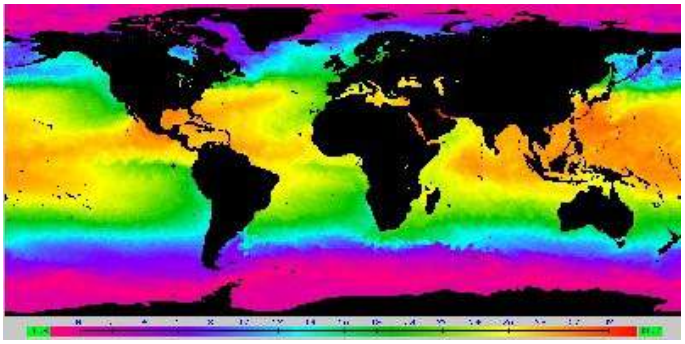


Figura 1.3: Temperatura media anual de la superficie del mar.

- Migración del halcón de Swainson.
  El halcón de Swainson, Figura 1.4, migra de Norteamérica a Sudamérica. Dicha migración dura alrededor de dos meses, y es considerada por los biólogos una de las migraciónes más largas de entre todas las aves del continente americano. El dato que se identifica, en esta situación, es la trayectoria que deja una parvada que pertenece a esta especie, es decir, la trayectoria recorrida durante el período de migración. Al igual que en los casos anteriores, el dato no es lineal pues la treyectoria no puede representarse como un elemento de $\mathbb{R}^n$. La afirmación anterior obedece al hecho de que cada punto que conforma a la trayectoria en cuestión, cae en un espacio cuya curvatura es distinta de cero, que en este caso concreto es la esfera. Cabe notar que el dato es adicionalmente complejo en el sentido de que realmente, lo que se observa es la trayectoria completa de puntos sobre la esfera.

  Es de interés estudiar estas trayectorias, ya que en el caso de México así como de otros países, hay requerimientos de hábitat para la época reproductiva de varios animales. Por lo tanto, puede ser de vital importancia conocer la trayectoria promedio que siguen dichas aves, pues con base en ella se pueden hacer posibles labores de conservación. En la Figura 1.5 se muestra la trayectoria genérica de la migración de una parvada que pertenece al halcón de Swainson.

Figura 1.4: Halcón de Swainson.

Figura 1.5: Trayectoria de la migración.

- Análisis de imágenes.
  En este tipo de datos el objeto de estudio se centrará en la forma que puede proporcionar una imagen y de la cual se desea obtener información. Una forma, en general, se define como la geometría de un objeto módulo su posición, tamaño y orientación. Como ejemplo se tiene la Figura 1.6, donde los peces tienen la misma forma independientemente de su escala, rotación y posicionamiento.

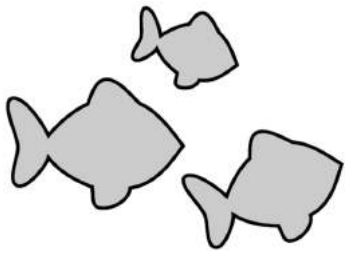


Figura 1.6: Peces con diferente posición, escalamiento y rotación.

Para asentar ideas se puede pensar como un caso de estudio el análisis de un hipocampo. El hipocampo es una parte del cerebro humano, Figura 1.7, el cual desempeña funciones importantes en la memoria así como en el manejo del espacio. El dato con el que se trabaja es la forma del hipocampo que se obtiene a través de una resonancia magnética del cerebro, como se muestra en la Figura 1.8.

Por ser una forma el objeto de estudio, es necesario "estandarizarla"; esto significa suprimir su escalamiento, rotación y posicionamiento. Al eliminar los elementos anteriores se obtiene que el espacio resultante, donde se encuentra la forma, no es el espacio n-dimensional. Por ende, el dato registrado no es lineal. Para revisar más detalles acerca del análisis de formas se puede consultar Dryden and Mardia [1998]. Éste es un libro introductorio en lo que respecta a la teoría de formas, el cual también motiva e introduce de manera didáctica

la importancia de describir y comparar las formas de los objetos.

Para concluir con este ejemplo y con relación en el estudio de un hipocampo, para los médicos es importante poder analizar este tipo de imágenes, pues con ellas pueden saber cuando un paciente es propenso a padecer Alzheimer[1].

Figura 1.7: Hipocampo.

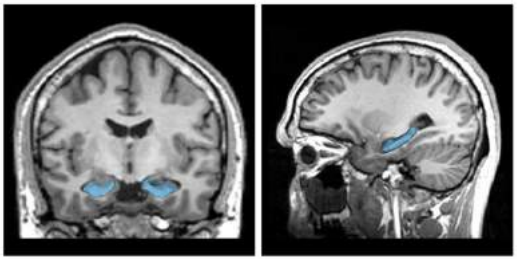


Figura 1.8: Hipocampo obtenido de una resonancia magnética del cerebro.

Los ejemplos anteriores se caracterizan por involucrar datos que por su naturaleza radican en espacios no euclidianos, a los cuales se les conoce como variedades no lineales. Entonces el siguiente paso es definir, de manera general, el concepto de variedad. Esta noción requerirá de ciertas bases en materia de topología, las cuales serán enunciadas brevemente. Es relevante aclarar que dichos conceptos ayudarán a definir correctamente los espacios en los que se trabajará, de forma que estos tengan propiedades que permitan extender la teoría estadística y probabilística que se conoce para $\mathbb{R}^n$. Sin embargo, para aquellos lectores que lo deseen, estas definiciones pueden ser omitidas en una primera lectura, pues en principio el objetivo es que se recoja la intuición de lo que es una variedad. Finalmente los lectores que deseen profundizar en las definiciones subsecuentes, pueden consultar Loring [2008] y Willard [1970].

Definición 1.1.1 *Espacio topológico.*
Es un pareja (X, τ), donde X es un conjunto cualquiera y τ es una familia de subconjuntos de X que satisface las siguientes propiedades:

1. *$X \in \tau$ y $\emptyset \in \tau$.*
2. *Dada una familia $\{U_i \in \tau; i \in I\}$ de elementos de τ, tal que I es un conjunto cualquiera, entonces $\cup_{i \in I} U_i \in \tau$.*

[1]Las Figuras 1.1–1.8 se obtuvieron, respectivamente, de los siguientes sitios de internet:
http://tortugas.anipedia.net.
http://www.windows2universe.
http://hotspotbirding.com.
http://www.birdscalgary.com.
http://www.sci.utah.edu/ fletcher/CS7960.
http://yogafacile.it.
https://nac.spl.harvard.edu.

3. Si $U_1, \ldots, U_n \in \tau$ *entonces* $\cap_{i=1}^{n} U_i \in \tau$.

En tal caso se dirá que τ *es una topología sobre* X *y a sus elementos se les llamará conjuntos abiertos de* (X, τ).

Definición 1.1.2 *Espacio Hausdorff.*
Es un espacio topológico (X, τ)*, tal que para cualesquiera dos puntos* $x, y \in X$*, existen dos abiertos* $U(x), V(y) \in \tau$ *tales que* $U(x) \cap V(y) = \emptyset$.

Definición 1.1.3 *Base.*
Si (X, τ) *es un espacio topológico, una base para* τ *es una colección* $\mathcal{B} \subset \tau$ *tal que*

$$\tau = \{\cup_{B \in \mathcal{C}} B \mid \mathcal{C} \subset \mathcal{B}\}.$$

Definición 1.1.4 *Espacio segundo numerable.*
Un espacio X *es segundo numerable si posee una base a lo sumo numerable de abiertos.*

Las siguientes dos definiciones serán vitales, pues aclararán las bondades—y también dificultades—de trabajar en estos espacios a los cuales, como ya fue mencionado anteriormente, se les denominará como variedades.

Definición 1.1.5 *Espacio localmente euclídeo.*
Un espacio topológico M *es localmente euclídeo de dimensión* n *si todo punto* $p \in M$ *tiene una vecindad* U*, tal que existe un homeomorfismo* ϕ *de* U *sobre un subconjunto abierto de* $\mathbb{R}^n$*. A la pareja* $(U, \phi : U \longrightarrow \mathbb{R}^n)$ *se le llamará carta coordenada. La carta coordenada* (U, ϕ) *estará centrada en* $p \in U$ *si* $\phi(p) = 0$.

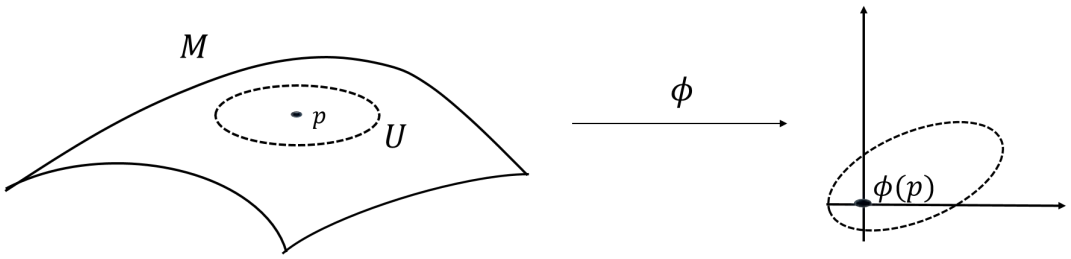


Figura 1.9: Carta coordenada.

La aportación de la Definición 1.1.5 es que introduce la noción de carta coordenada, la cual permitirá extender conocimientos de $\mathbb{R}^n$ a espacios no lineales.

Definición 1.1.6 *Variedad.*
Una variedad topológica M *es un espacio Hausdorff, segundo numerable que localmente es un espacio euclídeo. Se dice que es de dimensión* n *si localmente es un espacio euclídeo de dimensión* n.

La Definición 1.1.6 establece que una variedad es un espacio que localmente se parece a $\mathbb{R}^n$, por lo cual hereda—de manera local—sus propiedades. Por tanto, intuitivamente, una variedad podría entenderse como un espacio conformado exclusivamente por "parches" de $\mathbb{R}^n$. Tómese como ejemplo el toro, Figura 1.10, el cual se encuentra representado por pequeños "parches". El argumento anterior implica que una variedad, en general, no cuenta con espacio externo e interno. Esto quiere decir que una variedad, en principio, no posee espacio ambiente. Por ejemplo si se colocara un grupo de hormigas sobre un toro, Figura 1.11, éstas podrían moverse solamente por los parches que conforman a dicho espacio. El hecho de que una variedad no cuente con espacio ambiente, es uno de los factores que agregan complejidad al análisis estadístico, puesto que para medir distancias entre observaciones que estén sobre una variedad, habrá que considerar una métrica distinta a la euclidiana, la cual contemple la curvatura del espacio.

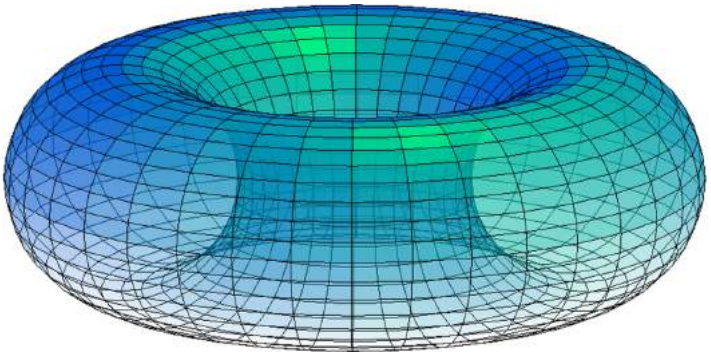


Figura 1.10: Toro.

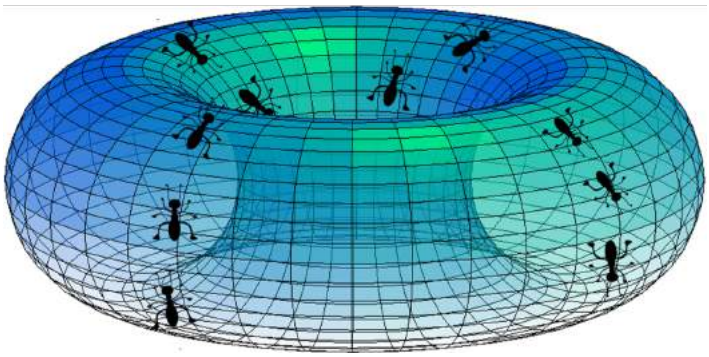


Figura 1.11: Hormigas sobre el toro.

A continuación se muestran algunos ejemplos de variedades, con la finalidad de aclarar las ideas dadas por la Definición 1.1.6.

Ejemplos:

- Espacio n-dimensional.
  También conocido como $\mathbb{R}^n$, se caracteriza por ser un espacio lineal y por ende

un espacio vectorial, ya que su curvatura es cero. Es una variedad pues cada abierto, en este espacio, es localmente euclídeo. Se considera como una de las variedades más estudiadas y por consiguiente una para las cuales se conocen más resultados.

- Círculo unitario.
  Es un espacio que de manera local se parece a $\mathbb{R}$. Esto se debe a que si se toma un punto $p \in S$—siendo S el círculo unitario—y en torno a dicho punto se traza una vecindad de radio ϵ, se podrá observar que dicha vecindad es homeomorfa a una linea recta, como se exhibe en la Figura 1.12. Del hecho anterior se sigue que el círculo unitario es una variedad de dimensión uno.

- Esfera unitaria.
  Es un espacio que de manera local se parece a $\mathbb{R}^2$, ya que si se toma una bola abierta en torno a un punto q que pertenezca a este espacio, se aprecia que esta vecindad es homeomorfa a un pequeño plano que se puede interpretar como un pequeño $\mathbb{R}^2$. El hecho anterior queda ejemplificado mediante la Figura 1.13. Por ende, la esfera es una variedad de dimensión dos.

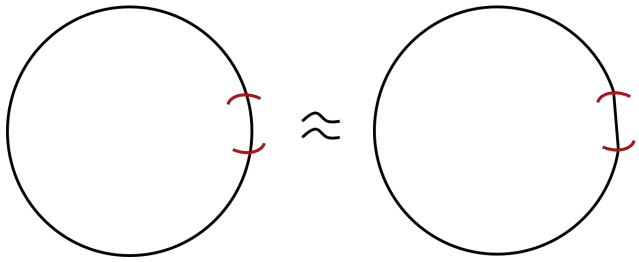


Figura 1.12: Círculo que de manera local se parece a $\mathbb{R}$.

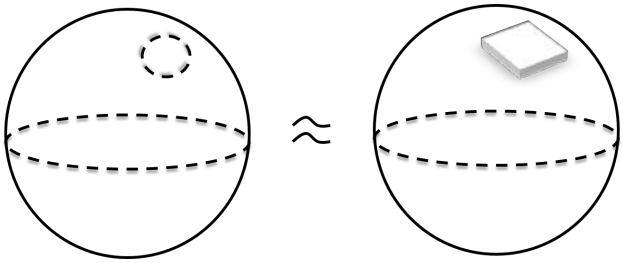


Figura 1.13: Esfera que de manera local se parece a $\mathbb{R}^2$.

Es importante destacar que no todas las variedades poseen una representación gráfica; un ejemplo de ello son SE(2) y el espacio de formas. SE(2) es el grupo de transformaciones rígidas en $\mathbb{R}^2$, tal que dichas transformaciones corresponden a traslaciones y rotaciones en el plano. Por otro lado, el espacio de formas es el que se encuentra definido por todas las rotaciones, traslaciones y escalamientos que puede tener una

forma. Por lo tanto, a diferencia de $\mathbb{R}^n$, en los espacios no lineales se pierde intuición e interpretabilidad de resultados estadísticos, pues estos no son directamente visualizables. Por consiguiente, éste es otro punto que añade dificultad al análisis estadístico sobre variedades. Para una lectura más profunda del espacio de formas y de SE(2), se recomienda leer Dryden and Mardia [1998] y Gallier [2001], respectivamente.

Ya que se ha introducido la noción de variedad y se han exhibido algunos ejemplos, sigue abordar otro tipo de variedades que tienen una estructura más rica, en el sentido de que es posible definir una distancia. Estas variedades son las riemannianas. Una variedad riemanniana es una variedad diferenciable, la cual está equipada con un producto interno. De esta manera, es diferenciable en el sentido de que la variedad posee una variación suave, es decir, no tiene picos. Por otro lado, el producto interno permitirá medir distancias sobre la variedad. La importancia que poseen las variedades riemannianas es que ayudarán a extender la metodología de cálculo diferencial a espacios más abstractos que $\mathbb{R}^n$. Por ende, trabajar nociones probabilísticas y estadísticas será más accesible, por la estructura que éstas poseen.

Algunos ejemplos de variedades riemannianas son el círculo y la esfera, ya que ambas varían de manera suave y la métrica que se les asocia es la de $\mathbb{R}$ y $\mathbb{R}^2$, respectivamente. Contrariamente a los ejemplos anteriores, el triángulo y el cono no son variedades riemannianas, pues no varían suavemente debido al pico que poseen. Algunos libros en los que se puede profundizar la teoría referente a variedades riemannianas son Do Carmo Valero [1992], Amari and Nagaoka [2007] y Lee [2006]. Todos ellos son libros introductorios, que comienzan con las nociones básicas de geometría diferencial para luego abordar conceptos más complejos referentes a esta temática.

1.2. Relevancia y complejidad del análisis estadístico sobre variedades

De acuerdo con la secuencia de ideas que se ha presentado y desarrollado hasta este punto, se desea motivar y mostrar que el tema de estadística sobre variedades posee relevancia en la actualidad. Para ello, basta hacer una búsqueda de las palabras "manifolds and statistics" en las tres revistas más importantes que refieren a temas de matemáticas así como de probabilidad y estadística, como lo son: Annals of Mathematics, Annals of Statistics y Journal of the Royal Statistical Society. Por ejemplo, Annals of Statistics muestra 50 artículos relacionados con este tema en lo que va del presente año; adicionalmente, en Google Scholar se pueden encontrar alrededor de 68800 referencias relacionadas con las palabras ya citadas. Lo anterior muestra la considerable actividad que tiene la presente área. Además, los temas de estadística que abordan estas publicaciones son muy variados, ya que van desde la estadística descriptiva hasta la inferencia paramétrica y no paramétrica considerando diversos enfoques, como el frecuentista y el bayesiano.

Para evidenciar la complejidad que puede existir al hacerse estadística sobre espacios no lineales, se mostrará en primer lugar la ineficacia de las herramientas estadísticas que se suelen utilizar para $\mathbb{R}^n$. Posteriormente y como segundo punto, se comentarán algunas nociones de estadística sobre variedades, con la finalidad de esbozar y ejemplificar el alcance de esta teoría.

Para exhibir el primer punto, se tomará como ejemplo la media muestral, que es uno de los conceptos base de estadística. Se define como $\bar{X} = \sum_{i=1}^{n} x_i/n$, donde $\{x_i\}_{i=1}^{n}$ representa un conjunto de observaciones. Supóngase que se cuenta con dos direcciones, 359° y 1°, las cuales están representadas en la Figura 1.14. Al tomar la media de estos dos datos se esperaría que diera una dirección representativa, como la de la Figura 1.15. Sin embargo, lo que se obtiene es una dirección como la que se muestra en la Figura 1.16, ya que $(359+1)/2 = 180$. Como segundo ejemplo tómese los puntos $(0°, 90°)$ y $(0°, -90°)$, los cuales representan el polo norte y polo sur en la esfera unitaria, respectivamente. Al promediar dichos puntos, entrada por entrada, se tiene que el punto resultante cae fuera de la esfera, como muestra la Figura 1.17, en la cual los puntos rojos representan los polos y el punto negro la suma de éstos.

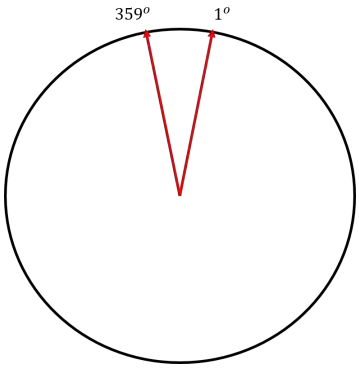


Figura 1.14: Dos datos direccionales, 359° y 1°.

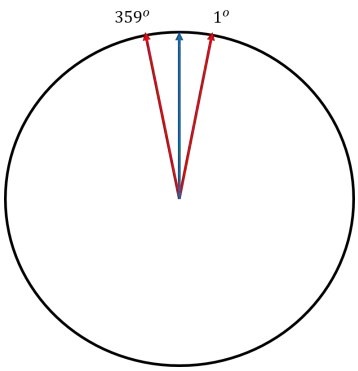


Figura 1.15: Expectativa de la media de dos direcciones.

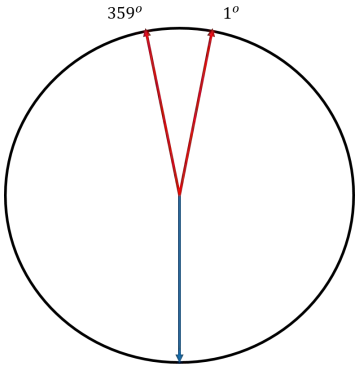


Figura 1.16: Realidad de la media de dos direcciones.

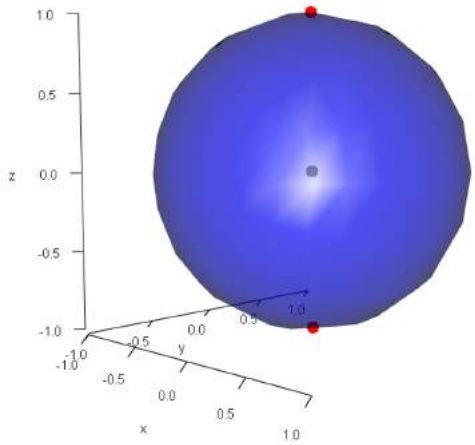


Figura 1.17: Suma de los puntos que están en los polos de la esfera.

La poca representatividad de la media es consecuencia de que las variedades, en general, no son espacios vectoriales. Por ende, las herramientas que han sido desarrolladas para $\mathbb{R}^n$ no funcionan en estos espacios, que son más complejos. Este punto es vital, pues aquí se esclarece que toda la intuición así como nociones estadísticas que se tienen de $\mathbb{R}^n$, pierden sentido en las variedades no lineales.

Para reforzar la idea expuesta en el párrafo anterior así como para exhibir el segundo punto—que es mostrar la dificultad de hacer estadística sobre variedades—se muestran a continuación algunos ejemplos de conceptos estadísticos en $\mathbb{R}^n$ que se han logrado extender a espacios no lineales.

1. Media.
   Conocida como media de Fréchet o Karcher. Tiene la siguiente definición.

   Definición 1.2.1 *Sea M una variedad y sea $\{x_i\}_{i=1}^n$ una colección de puntos tales que $x_i \in M$ para $i = 1, \ldots, n$. La media de Fréchet se define como*

   $$\mu = \underset{p \in M}{\text{argmín}} \sum_{i=1}^{n} d(p, x_i)^2, \tag{1.1}$$

   donde $d(\cdot, \cdot)$ representa la distancia definida en M.

   En otras palabras, $p \in M$ es el punto que minimiza la distancia entre todos los datos $x_i \in M$. Esta noción de media surge con Fréchet [1948], que es el artículo pionero en definir el concepto de media sobre variedades, mientras que Karcher [1977] es el primero en ofrecer un estudio acerca de sus propiedades.

   Dado que la media de Fréchet implica un proceso de minimización, se sigue que la media puede ser no única a diferencia de la media que se conoce en $\mathbb{R}^n$. Tómese como ejemplo la esfera y suponga que hay un conjunto de puntos que están sobre el ecuador, como muestra la Figura 1.18. La media en tal caso no

sería única, ya que los puntos que están en los polos son los que cumplen la Definición 1.2.1. Otro ejemplo sobre la esfera, en el cual la media sí es única, se encuentra representado mediante la Figura 1.19. En este caso se obtuvo la media de un conjunto de 14 puntos, tal que la media es el punto negro que se encuentra sobre la esfera.

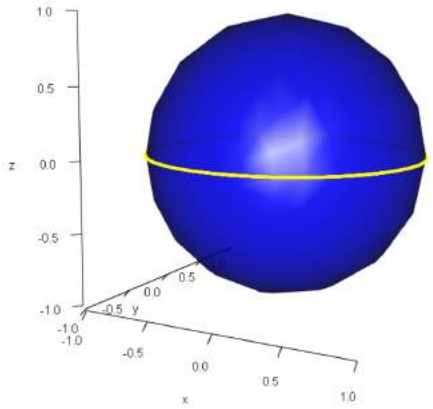


Figura 1.18: Ecuador en la esfera.

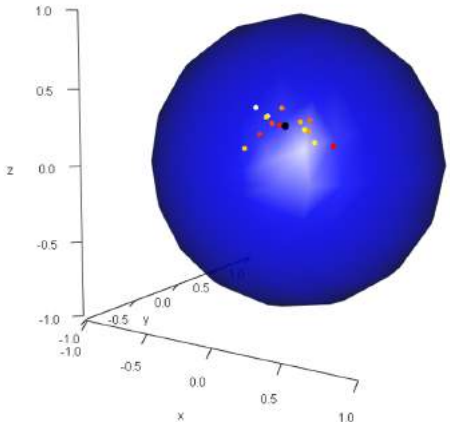


Figura 1.19: Media de Fréchet para 14 puntos.

Con los dos ejemplos exhibidos se muestra que una noción tan básica, como es la media, puede complicarse en espacios más abstractos. Por consiguiente, queda comprobado que hacer un análisis estadístico sobre variedades representa un reto. Algunos artículos en los cuales se pueden revisar diferentes aplicaciones de la media sobre variedades son: Kaziska and Srivastava [2008], Kume and Le [2003] y Rentmeesters and Absil [2011].

2. Clustering.
   También conocido como manifold clustering, consiste en crear grupos de datos, tales que pueden provenir de una o más variedades. Un artículo considerado

seminal en esta temática fue Souvenir and Pless [2005], ya que logró introducir teoría para clasificar datos que radican en múltiples variedades y a su vez logró hacer contribuciones a la teoría clásica de clustering. Una referencia que muestra la utilidad de hacer clustering en variedades mediante ejemplos reales y sintéticos es Tu et al. [2014].

3. PCA.
   Conocido como PGA por sus siglas en inglés (Principal Geodesic Analysis), consiste en reducir la dimensionalidad de los datos que están en una variedad M. Esta teoría puede ser revisada en el artículo de Fletcher et al. [2004] que es considerado el pionero en hacer PCA sobre variedades, ya que en éste logra describir la variabilidad de datos sobre variedades.

   Algunos artículos que refinaron la teoría propuesta por Fletcher son Jung et al. [2011], Jung et al. [2012] y de manera más reciente Pennec [2015]. Cabe mencionar que este último artículo, viene a resumir las metodologías que se han propuesto para PCA sobre variedades. Además, ofrece una nueva teoría, que abarca cualquier espacio no lineal. Esto implica un gran avance, pues las herramientas que se habían desarrollado de manera previa sólo contemplaban variedades riemannianas.

4. Estadística no paramétrica.
   Como su nombre lo indica, consiste en implementar estadística no paramétrica sobre variedades. Surge con el trabajo de Hendriks and Landsman [1996] titulado *Asymptotic tests for mean location on manifolds*, el cual sería seguido por Patrangenaru [1998] con su tesis doctoral *Asymptotic Statistics on Manifolds*.

   La estadística no paramétrica ha sido una de las herramientas más usadas para el análisis de datos no lineales, pues al estar éstos en espacios más complejos de los cuales no se posee intuición por su curvatura, se busca una herramienta que permita a los datos expresar la estructura que les gobierna. Es en este sentido que la estadística no paramétrica arroja resultados más nítidos, en comparación de su contraparte paramétrica. Por lo tanto, resulta muy conveniente el que los resultados derivados del análisis estadístico no dependan de la distribución elegida. Algunos libros que abordan de manera completa esta temática son Bhattacharya and Bhattacharya [2012] y Patrangenaru [2015].

Con los ejemplos anteriores se muestra que la tarea de inferencia estadística sobre variedades es un tema de actualidad. Sin embargo, esto conlleva un reto teórico y computacional para extender las nociones probabilísticas y estadísticas que se conocen de $\mathbb{R}^n$ a variedades. Lo anterior se traduce en uno de los principales objetivos de la presente tesis, así como en una parte fundamental de ella.

1.3. Importancia del análisis estadístico sobre variedades

El análisis estadístico sobre variedades es una temática no convencional, la cual surge con Rao [1945]. Es considerada como una metodología joven que ha despuntado en los últimos años, debido al auge computacional de la última década. Es por ello que, de manera reciente, se ha profundizado en la teoría del análisis estadístico sobre variedades, pues la cantidad de aplicaciones que tiene son muy diversas. Incluyen, por ejemplo, las que se presentaron en la Sección 1.1 del presente capítulo.

La incursión de la estadística en el marco de geometría diferencial, ha sido abordada por algunos libros. Uno de ellos es Shun-ichi [1985], quien ofrece una de las primeras referencias en tratar esta sinergia. Este libro es muy cuidadoso y esmerado en muchos aspectos, pues aporta un marco histórico acerca de cómo ocurrió dicha sinergía, además de ofrecer nociones de estadística así como de geometría diferencial, y explicar y motivar la importancia de la geometría diferencial en la estadística.

No obstante, a pesar de la existencia de libros como el ya comentado, todavía no existe una cantidad considerable de libros que aborden el análisis estadístico sobre variedades. Más aún, que aborden esta temática de una forma alcanzable para personas que no poseen conocimientos en probabilidad y estadística o en geometría diferencial. Para complementar la afirmación anterior, se resumirán a continuación algunos libros, los cuales abordan el análisis estadístico sobre variedades.

1. Shun-ichi [1985]. *Differential-Geometrical methods in statistics.*
   Es una monografía que está dividida en dos partes. La primera parte consta de la teoría referente a geometría diferencial, mientras que la segunda refiere a la teoría estadística sobre variedades. Esta última se encuentra especializada a las distribuciones que pertenecen a la familia exponencial.

   A pesar de que inicia con las nociones básicas de geometría diferencial, desde una perspectiva intuitiva, es necesario contar con cierta intuición geométrica y topológica para alcanzar a entender los conceptos que aborda. En lo que respecta a la parte de teoría de probabilidad y estadística, se requieren los conocimientos básicos de inferencia estadística. Este texto es ideal para alumnos de licenciatura quienes ya poseen cierto bagaje en las temáticas de geometría diferencial e inferencia estadística.

2. Fisher et al. [1987]. *Statistical analysis spherical data.*
   Es uno de las primeros libros en abordar el tema de análisis estadístico sobre variedades. Se caracteriza por ofrecer un resumen de los métodos estadísticos y probabilísticos que existen para trabajar y simular datos puntuales que yacen en la esfera, para luego abordar teoría moderna con la que pueden ser tratados. Así mismo, trata algunas técnicas matemáticas para trabajar vectores y matrices con coordenadas polares y estándar .

Curiosamente este libro nunca hace alusión a la esfera, vista como una variedad. Por lo tanto, la teoría desarrollada es exclusivamente de índole estadístico. Por consiguiente, este ejemplar puede ser leído por cualquier persona que tenga conocimientos básicos en álgebra matricial y bases sólidas en inferencia estadística.

3. Mardia and Jupp [1999]. *Directional statistics.*
   Trata la metodología estadística y probabilística de datos direccionales. Primero ahonda en datos que se encuentran sobre el círculo y después hace la extensión a datos que están sobre la esfera. Un ejemplo de este tipo de datos fue visto en la Sección 1.1, con las direcciones que toma un grupo de 36 tortugas.

   Lo interesante del texto es que motiva, con comentarios esporádicos, la idea de que es posible trabajar y extender la teoría desarrollada a espacios más generales que $\mathbb{R}^n$. Primero aborda la teoría clásica de datos direccionales y luego plantea la teoría moderna con la que pueden ser tratados estos datos, incluyendo el análisis estadístico sobre variedades. Este libro se puede considerar como una referencia base para todo aquel que desee conocer y aprender la teoría estadística de datos direccionales, pues como conocimiento previo se requiere únicamente una parte básica de inferencia estadística.

4. Amari and Nagaoka [2007]. *Methods of information geometry.*
   Básicamente trata la relación que hay entre la estadística y la geometría diferencial. Dedica los cuatro primeros capítulos a dar las herramientas necesarias de geometría diferencial y estadística. Los capítulos subsecuentes tratan las diversas aplicaciones que puede tener la geometría diferencial, como inferencia estadística, redes neuronales y sistemas dinámicos.

   En la medida de lo posible, esta referencia ofrece la intuición de los conceptos geométricos que va planteando. Sin embargo, entra de lleno en materia de geometría diferencial, lo cual puede tornarse complicado para aquellas personas que buscan un primer acercamiento a esta rama de las matemáticas. De igual forma, en lo que respecta a inferencia estadística es necesario contar con una formación superior a la básica, pues llega a obviar ciertas definiciones que pueden resultar cruciales para el entendimiento del material cubierto. Por tanto, la lectura de dicho ejemplar es accesible para alumnos de posgrado, que tengan conocimientos en las ramas ya citadas.

5. Bhattacharya and Bhattacharya [2012]. *Nonparametric inference on manifolds: with applications to shape spaces.*
   Es el primer libro en ofrecer un tratado de inferencia no paramétrica en variedades, con aplicaciones al espacio de formas. Se caracteriza por abordar un enfoque clásico y bayesiano, así como por ofrecer nuevas herramientas teóricas en lo que respecta a esta temática. Además, muestra ejemplos de cómo se implementa esta teoría con datos reales y sintéticos.

Para la lectura de este texto se requieren conocimientos sólidos en lo que respecta a geometría diferencial, estadística y teoría asintótica de probabilidad. Por tanto, la lectura de este libro puede resultar poco accesible para estudiantes de licenciatura así como para algunos alumnos de posgrado, ya que el material que presupone y ofrece es avanzado.

6. *Nonparametric statistics on manifolds and their applications to object data analysis.*
   Es el libro más reciente en lo que refiere al análisis estadístico sobre variedades, publicado el 25 de septiembre de 2015. Para consultar su contenido se puede revisar la siguiente liga:

   https://www.crcpress.com/ Nonparametric-Statistics-on-Manifolds-and-Their-Applications-to-Object/Patrangenaru-Ellingson/9781439820506.

Por consiguiente y en conformidad con la estructura de ideas expuesta, se tiene que la tesis cobra relevancia e importancia, ya que por una parte ofrecerá un texto autocontenido accesible para aquellas personas que no poseen conocimientos de geometría diferencial, y además aportará una concepción estadística y probabilística del análisis de datos sobre variedades. En este trabajo dicho análisis estará particularizado al estudio de trayectorias.

1.4. Análisis estadístico de trayectorias sobre variedades

El análisis estadístico de trayectorias tiene su origen con Trouvé and Younes [2000]. Sin embargo, es hasta Su et al. [2014a] con *Statistical analysis of trajectories on Riemannian manifolds: bird migration, hurricane tracking and video surveillance*, que surge el primer artículo en abordar un estudio estadístico de trayectorias sobre variedades riemannianas. El presente artículo se caracteriza por usar nociones maduras de probabilidad y estadística, así como por concebir a la trayectoria como un dato. Además, logra una conjunción del marco teórico de geometría diferencial con el de probabilidad y estadística. Lo anterior se traduce en la implementación de la teoría abordada y con ello en el estudio de algunos casos, tales como el análisis de trayectorias de vehículos y de actividad humana. En otras palabras, Su et al. [2014a] es un artículo que innovó la representación y estudio de trayectorias sobre variedades. Por consiguiente y después de una extensa búsqueda bibliográfica, se adoptó esta referencia como base para el desarrollo de la presente tesis.

Para destacar la trascendencia que puede poseer un estudio estadístico de trayectorias sobre variedades, se mostrarán a continuación algunos ejemplos. Se expondrán de forma que estos también presenten las herramientas estadísticas que se pueden emplear y que a su vez es necesario extender.

1. Movimiento humano, como el seguimiento e interpolación de la trayectoria que puede tener una o varias partes del cuerpo. Su principal aplicación es en el área

de rendimiento deportivo, así como para el diagnóstico médico. La estadística que hay detrás de dicha aplicación tiene por objetivo detectar el movimiento óptimo, que puede hacer un golfista o un beisbolista por ejemplo, para lograr una anotación y calcular la probabilidad de que en efecto sea exitosa la acción.

Esta aplicación representa un reto estadístico, pues para lograr los objetivos mencionados es necesario establecer una métrica útil, en el sentido de que incorpore la estructura subyacente del espacio en el que se encuentran los datos. Con dicha métrica se habilita un análisis de reconocimiento de patrones, así como una extensión del análisis de regresión o de interpolación para variedades. Posteriormente, se procede a ajustar un modelo de probabilidad y con la ayuda de técnicas Monte Carlo, calcular la probabilidad de ocurrencia de una trayectoria. Las Figuras 1.20, 1.21 y 1.22 muestran un ejemplo de interpolación para el movimiento de brazo de un golfista. La Figura 1.20 deja un espacio entre imágenes de las cuales se desea obtener la interpolación, mientras que las Figuras 1.21 y 1.22 muestran los resultados que se obtuvieron con los procedimientos ya mencionados en este párrafo.

Figura 1.20: Interpolación del braceo de un golfista.

Figura 1.21: Interpolación lineal.

Figura 1.22: Interpolación via variedades.

2. Trayectorias de personas infectadas de alguna enfermedad, la cual puede contagiarse por la picadura de algún insecto o por contagio directo; un ejemplo lo sería el dengue. Ésta es una enfermedad que se transmite por picadura de mosco y es considerada como una de las enfermedades epidemiológicas más peligrosas, según la OMS.

   El rol que juega la estadística en este contexto es encontrar una trayectoria que represente los lugares que visitan de manera frecuente las personas que se encuentran infectadas. Ulteriormente, poder estimar el número de veces que una persona debe estar expuesta a un posible foco de infección, para determinar si contrae la enfermedad o no. La complejidad estadística en este problema radica en el hecho de encontrar esa trayectoria representativa, pues ésta debe respetar la forma que poseen calles y avenidas por donde pasan las personas contempladas en el estudio. En la Figura 3.41 se muestra un mapa de los lugares, en la República Mexicana, donde hay mayor presencia del mosquito del dengue. Las líneas grises y puntos verdes representan una propuesta de distribución de patrullas sanitarias, de tal forma que se maximice la cobertura médica en las zonas de mayor suceptibilidad al dengue.

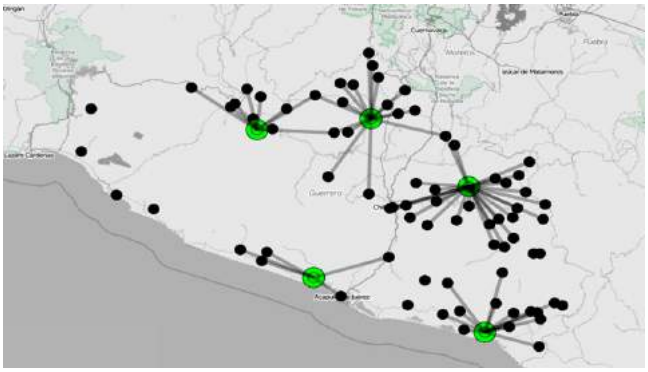


Figura 1.23: Mapa de la República Mexicana con mayor incidencia del dengue.

3. Trayectorias de automóviles. El movimiento en general de un vehículo se puede clasificar en cuatro grandes grupos. Éstos son una vuelta a la izquierda, a

la derecha, un movimiento en "U" o simplemente una línea recta. Sin embargo, estos movimientos poseen cierta variación por las diferentes velocidades de desplazamiento que tienen los vehículos. Dichas variaciones pueden deberse a diversos factores, siendo uno de ellos las alteraciones que presenta el tráfico. Debido a lo anterior, es que la clasificación de la trayectoria de un vehículo en movimiento se puede complicar.

La aplicación estadística, en este caso, consiste en estimar la variación de un conjunto de trayectorias así como clasificarlas. Por consiguiente, el reto es encontrar una métrica que incorpore la velocidad con la que se recorre cada trayectoria, permitiendo que en el análisis se logre dicernir cuándo la trayectoria observada pertenece a un peatón y no a un autómovil. La Figura 1.24 muestra del lado izquierdo un conjunto de trayectorias tomadas con una cámara de tránsito, tal que dicho conjunto está conformado por dos automóviles y un peatón. En el lado derecho, de la misma figura, se muestra el resultado de un proceso de aprendizaje automatizado aplicado al conjunto de trayectorias en estudio. Los resultados obtenidos fueron los diferentes lugares de localización y dirección que pueden presentar los vehículos en cuestión[2].

Figura 1.24: Trayectorias de automóviles y peatón.

Hasta este punto se ha esbozado la estadística que puede hacerse para un conjunto de trayectorias que yacen en una variedad M. Sigue comentar, a grandes rasgos, el tipo de análisis estadístico que será estudiado en el presente trabajo. Dado un conjunto de trayectorias, se plantea encontrar una trayectoria media que sea representativa de dicho conjunto, en el sentido de que logre capturar una forma que sea representativa e interpretable. Posteriormente se propone encontrar la varianza asociada a la muestra de trayectorias. De esta manera, con estos parámetros y un modelo de probabilidad—a decir una distribución normal—se obtiene une representación matemática para describir y simular trayectorias.

Para concluir la presente sección, se aclara que el análisis estadístico de trayectorias tratado en la presente tesis se verá restringido a la esfera. La motivación para

[2]Las Figuras 1.20, 1.21, 1.22, 3.41 y 1.24 se obtuvieron de los siguientes sitios de internet:
https://www.cs.cmu.edu.
http://www.conacytprensa.mx.
http://people.csail.mit.edu.

ello es que los resultados del proceso estadístico se pueden visualizar, por lo cual son más sencillos de interpretar y entender. Por otra parte, con dicha restricción se facilitará el cómputo, pues al ser la esfera una de las variedades más estudiadas, se cuenta con expresiones analíticas cerradas para algunas nociones geométricas de interés. De manera que estas expresiones serán de utilidad al momento de realizar ciertas implementaciones.

1.5. Estructura de la tesis

Se esbozará, de manera concisa, los objetivos de la tesis y el contenido que posee cada capítulo.

1.5.1. Objetivos

Los objetivos de la tesis son identificar y recomendar literatura base, así como incursionar en la metodología para estudiar trayectorias sobre variedades, particularmente en la esfera. Por lo tanto, la tesis plantea las siguientes metas y aportaciones:

1. Ofrecer un texto autocontenido.

2. Abordar un caso de estudio.

3. Desarrollar un breve ensayo de simulación.

1.5.2. Capítulo 2

Trata los elementos técnicos de geometría diferencial que son necesarios para entender el resumen del artículo base—el cual será abordado en el Capítulo 3— de manera que dichas nociones de geometría diferencial serán especializadas a la esfera. El capítulo contendrá las siguientes secciones:

1. Espacio tangente a un punto.

2. Curvas geodésicas.

3. Mapeo exponencial.

4. Mapeo logarítmico

5. Transporte paralelo.

Además, éste se caracterizará por ofrecer expresiones analíticas cerradas sobre la esfera, de las nociones de geometría mencionadas anteriormente. Así mismo, ofrecerá algunas pruebas didácticas. La finalidad es familiarizar y aportar intuición, al lector, sobre los conceptos geométricos y cómo se enlazan entre sí.

1.5.3. Capítulo 3

Es un resumen estructurado del artículo *Statistical analysis of trajectories on Riemannian manifolds: Bird migration, hurricane tracking and video surveillance*, el cual estará dividido en tres grandes secciones. El propósito de dicha estructuración es procurar que sea entendible el procedimiento estadístico que hay de por medio, para el manejo de trayectorias sobre variedades. Las secciones contempladas son:

1. Trayectorias. La intención es que el lector alcance a percibir la complejidad que caracteriza al dato.

2. Trayectorias como objeto matemático. Introduce la necesidad de usar la geometría diferencial como herramienta y con ella caracterizar las trayectorias.

3. Análisis estadístico de trayectorias. Combina algunas nociones de geometría diferencial y estadística, para lograr el objetivo de hacer inferencia sobre la esfera. Además, abordará un breve ejemplo de simulación—cuyo índole es principalmente didáctico—de trayectorias de huracanes. Los datos que se usarán se pueden encontrar en el siguiente sitio

$$http://weather.unisys.com/hurricane/atlantic/.$$

El objetivo de este capítulo es ofrecer un resumen asequible del artículo base, así como rellenar detalles técnicos que se dan como presupuestos. Además, identificará el rol que juegan las nociones de geometría diferencial en el desarrollo estadístico, y finalmente hará alcanzable la teoría descrita con la implementación de los algoritmos desarrollados a los datos de huracanes mencionados anteriormente.

Capítulo 2

Elementos técnicos para estadística sobre variedades

2.1. Introducción

Para poder hacer y entender la teoría estadística sobre variedades es necesario tener herramientas técnicas adecuadas, en este caso geometría diferencial, que se especializará en la esfera. La esfera es el conjunto de vectores en $\mathbb{R}^3$ cuya norma satisface ser igual a uno y se denota como S^2, de forma que

$$S^2 = \{x \in \mathbb{R}^3 : \ \|x\| = 1\}.$$

Además, cabe decir que la esfera es una variedad riemanniana, como fue visto en el Capítulo 1. Un atributo que destaca en estas variedades es tener un producto interno definido, y por ende una distancia. En el caso de la esfera, la distancia es una medida que se toma a lo largo de la superficie y es la más corta para cualesquiera dos puntos $p, q \in S^2$. La distancia en la esfera se define como

$$d(p, q) = \arccos(\langle p, q \rangle), \tag{2.1}$$

tal que $\langle \cdot, \cdot \rangle$ denota el producto interno del espacio euclidiano. A la ecuación (2.1) se le considerará como la distancia intrínseca de la esfera, la cual cobrará importancia en definiciones que serán tratadas más adelante.

Los objetivos de este capítulo son tres: familiarizar al lector con algunos conceptos de geometría diferencial, estudiar teoría preliminar supuesta en el artículo base y especializar los conceptos de geometría diferencial en la esfera. Todo esto tiene la finalidad de hacer accesible el resumen del artículo base que será abordado en el Capítulo 3.

Las aportaciones del capítulo son dar un orden lógico a la intuición geométrica y ofrecer cierta heurística de los conceptos geométricos que serán tratados en secciones posteriores. Lo anterior se logrará mediante la definición y ordenamiento de conceptos—cuya dificultad vaya en orden creciente—, así como la explicación e interpretación de los mismos. Además, se mostrarán representaciones gráficas y

demostraciones didácticas.

Como lecturas generales se recomiendan Su et al. [2014a] y Fletcher [2010]. En la primera referencia se encontrarán algunas formulaciones de la hiperesfera, mientras que en la segunda se podrán hallar nociones heurísticas y técnicas del análisis estadístico en variedades.

2.2. Nociones básicas de geometría diferencial

En esta sección se abordarán algunos conceptos básicos de esta teoría. Primero se definirán de manera general y luego se especializarán a la esfera, con la finalidad de contar con expresiones cerradas de conceptos que serán tratados posteriormente. Dichas expresiones serán de vital importancia en el Capítulo 3, debido a que el modelo estadístico para trayectorias en variedades recaerá por completo en nociones de geometría diferencial.

Las definiciones dadas en el presente capítulo se obtuvieron de Do Carmo Valero [1992], Lee [2006], Fletcher et al. [2004], Loring [2008] y Do Carmo [1976]. Las expresiones analíticas de la esfera se obtuvieron de Bhattacharya and Bhattacharya [2012] y Su et al. [2014a].

2.2.1. Espacio tangente a un punto

Se denota como T_pM, donde p es un punto que pertenece a una variedad M. Para poder formalizar este concepto, primero se abordará la definición de curva parametrizada, curva y vector tangente.

Definición 2.2.1 *Curva parametrizada o trayectoria:*
Sea M una variedad diferenciable, I un intervalo abierto y $\alpha : I \subset \mathbb{R} \to M$ una función diferenciable, entonces α será conocida como curva parametrizada.

Definición 2.2.2 *Curva:*
Una curva (en M) es un subconjunto $C \subset M$ que admite una parametrización $\alpha : I \to M$; i.e. existe α diferenciable con $\alpha(I) = C$ tal que α es una función regular, $\alpha'(t) \neq 0$ para todo t.

A continuación se muestra un ejemplo de los puntos comentados en este párrafo. Se tienen dos curvas parametrizadas $\alpha(t) = (\sin(t), \cos(t))$ y $\beta(t) = (\cos(t), \sin(t))$, representadas en la Figura 2.1 respectivamente, tal que $\alpha(t) \neq \beta(t)$; sin embargo, ambas curvas parametrizadas imprimen la misma curva o traza, que es el círculo. Por tanto la palabra curva parametrizada hará alusión a una función $\alpha(t)$ mientras que curva se referirá a la misma imagen o traza que dejan varias fuinciones, en este caso $\alpha(t)$ y $\beta(t)$.

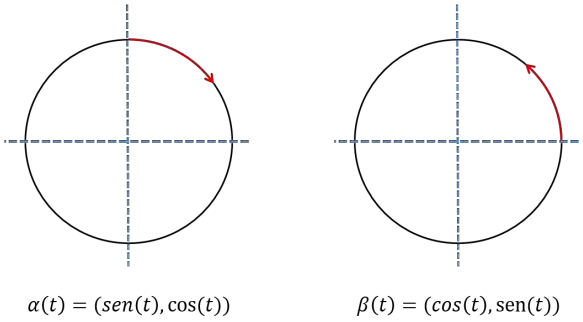


Figura 2.1: Círculo unitario, parametrizado de dos formas diferentes.

Definición 2.2.3 *Vector tangente.*
Sea M una variedad diferenciable, $p \in M$ y α una curva parametrizada en M. Supóngase $\alpha(0) = p$, y sea D elconjunto de funciones sobre M que son diferenciables en p. El vector tangente a una curva α en $t = 0$ es una función $\alpha'(0) : D \longrightarrow \mathbb{R}$ dada por

$$\alpha'(0)f = \left.\frac{d(f \circ \alpha)}{dt}\right|_{t=0}, \quad f \in D.$$

Un vector tangente en p, es el vector tagente en $t = 0$ de alguna curva $\alpha : (-\epsilon, \epsilon) \to M$ con $\alpha(0) = p$.

La Definición 2.2.3, permite extender a variedades diferenciables la noción que se tiene de vector tangente en $\mathbb{R}^n$ y con ello la noción de vector velocidad. Lo anterior es relevante ya que las variedades no cuentan con un espacio ambiente, como se mencionó en el Capítulo 1.

Definición 2.2.4 *Espacio tangente a M en un punto.*
Dado un punto $p \in M$, el conjunto de todos los vectores tangentes a M en p, se llamará espacio tangente a M en p.

El plano tangente en términos geométricos, se puede interpretar como un conjunto vectores que están ligados a un cierto conjunto de curvas parametrizadas, las cuales pasan por un punto $p \in M$. Además, se caracteriza por tener la misma dimensión que la variedad M y por ser un espacio vectorial. Es conveniente aclarar que en este caso el neutro aditivo, de dicho espacio vectorial, está dado por el vector tangente a la curva constante $\alpha(t)$ tal que $t \longrightarrow p$, donde p es el punto en el que se define el plano tangente.

Por otro lado, el que el espacio tangente a un punto de la variedad sea un espacio vectorial tiene cierta importancia, y es que se puede entender como una "linealización" de la variedad. La utilidad de este hecho es que se prodrán aprovechar conocimientos de $\mathbb{R}^n$; por ejemplo, la noción de media muestral.

Una condición suficiente para que el espacio tangente a un punto exista es que la variedad sea diferenciable. En la Figura 2.2 se muestran dos variedades donde el espacio tangente no existe; esto se debe a que en ambos casos, el cono y el triángulo tienen un pico, punto en el cual no es posible definir el espacio tangente. Es por ello

que es importante considerar a las variedades riemannianas, pues al ser diferenciables el espacio tangente siempre existe. Ambos conceptos, variedad riemanniana y variedad diferenciable, fueron revisados brevemente en el Capítulo 1.

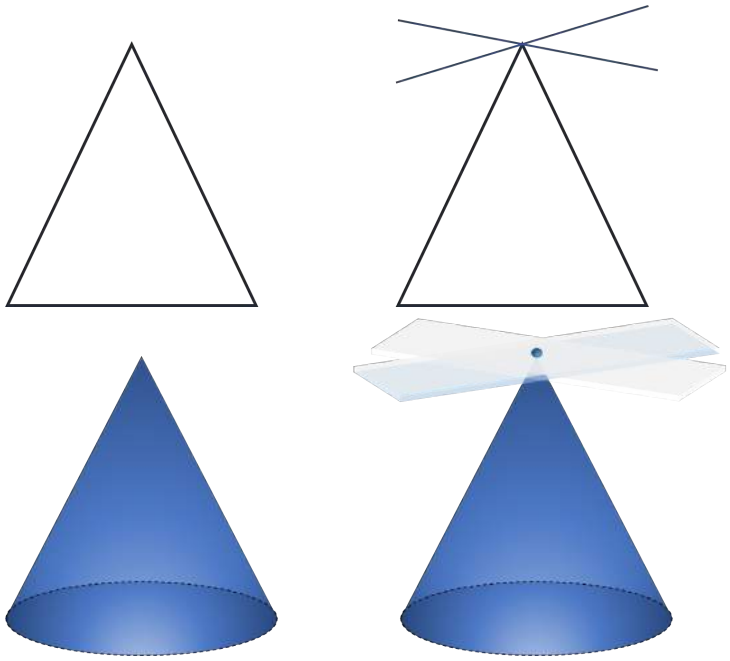


Figura 2.2: Triángulo y cono con línea y plano tangente respectivamente.

Espacio tangente a la esfera en un punto

En el caso de la esfera, es un plano el cual se define como

$$T_pS^2 = \{v \in \mathbb{R}^3 : \langle v, p\rangle = 0\}, \quad \forall p \in S^2. \tag{2.2}$$

Gráficamente se puede representar como se exhibe en la Figura 2.3.

Algunos comentarios importantes que surgen a partir de (2.2) son los siguientes:

- El plano tangente a la esfera es de dimensión dos, por lo cual existe un isomorfismo con $\mathbb{R}^2$. La relevancia de esto es que se podrán emplear conocimientos y métricas del espacio de funciones, de manera específica una modificación de la norma L^2, como se verá en el Capítulo 3.

- Todo punto de la esfera posee un plano tangente, el cual se define de manera única. Por lo tanto, para cualquier punto $p \in S^2$ existe una linealización de la esfera.

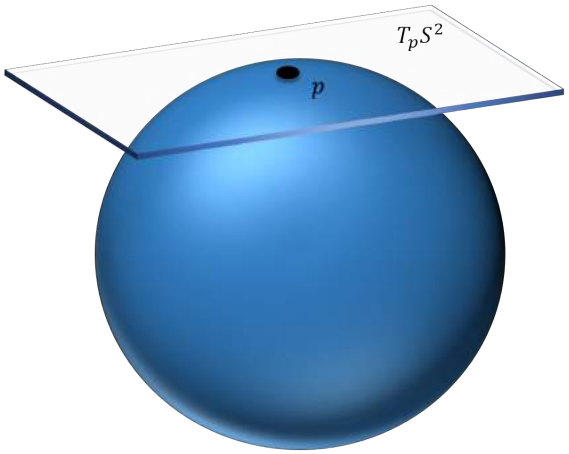


Figura 2.3: Plano tangente a un p en la esfera.

2.2.2. Curva geodésica

Es una curva que localmente minimiza la longitud entre dos puntos de una variedad M. Se denota como $\gamma_{p,v}(t)$, donde $p \in M$, v representa la dirección que toma la curva γ y t denota el tiempo que cubrirá la curva.

Definición 2.2.5 *Sea $\gamma : I \longrightarrow M$, I cualquier intervalo abierto contenido en $\mathbb{R}$, γ es geodésica en $t_0 \in I$ si $\frac{D}{\partial t}\left(\frac{\partial \gamma}{\partial t}\right) = 0$ en t_0; si γ es geodésica en t para toda $t \in I$, entonces se dice que γ es geodésica.*

Es importante aclarar que en el contexto del presente trabajo, el operador $\frac{D}{\partial t}\left(\frac{\partial \gamma}{\partial t}\right)$, se entenderá como una "segunda derivada". Para tener la formalidad y percibir la intuición de este operador, así como la analogía que posee con la segunda derivada usual, se recomienda consultar Do Carmo Valero [1992] y Sánchez Morgado and Palmas Velasco [2007].

De la Definición 2.2.5 se tienen las siguientes consecuencias:

- Las geodésicas son curvas con velocidad constante y aceleración cero.
- Si $p \in M$ y $v \in T_pM$ entonces existe una única geodésica $\gamma_v(0) = x$ y $\gamma_v'(0) = v$.

Para asentar la noción de curva geodésica a continuación se muestran los siguientes ejemplos: $\mathbb{R}^2$ y el toro. Para cualesquiera dos puntos $p, q \in \mathbb{R}^2$, la curva geodésica que los une es una línea recta, como lo muestra la Figura 2.4. Por otro lado, si p y q son dos puntos en el toro, entonces la curva geodésica que los une es aquella que tiene la menor distancia en el toro, como se ejemplifica en la Figura 2.5.

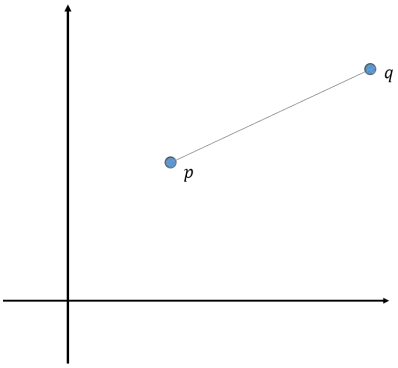


Figura 2.4: Curva geodésica en $\mathbb{R}^2$.

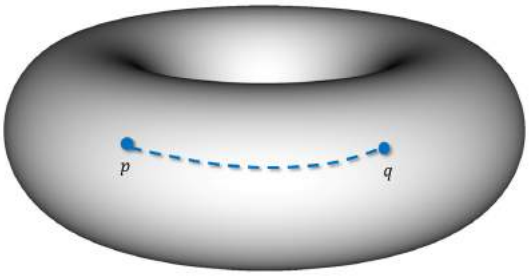


Figura 2.5: Curva geodésica en el toro.

Curvas geodésicas en la esfera

Las curvas geodésicas en la esfera son grandes círculos, que pueden ser parametrizados de diversas formas. La primera parametrización es

$$\gamma_{p,v}(t) = \cos(t)p + \sin(t)v, \text{ tal que } -\pi < t \leq \pi. \tag{2.3}$$

Esta curva geodésica empieza en p cuando $t = 0$ y toma la dirección del vector v, cuya norma es igual a uno.

Una segunda parametrización es

$$\gamma_{p,v}(t) = \cos(t\|v\|)p + \sin(t\|v\|)\frac{v}{\|v\|}, \text{ tal que } \frac{-\pi}{\|v\|} < t \leq \frac{\pi}{\|v\|} \text{ y } v \neq 0. \tag{2.4}$$

Esta representación también obedece el hecho de que $\gamma_{p,v}(0) = p$ y toma la dirección dirección del vector v, el cual tiene norma $\|v\|$.

La parametrización (2.3) se caracteriza por tener una velocidad unitaria. En contraste (2.4) se caracteriza por llevar una velocidad v. Por tanto, el elemento que cambia en cada representación es la velocidad con la que se recorre la curva en cuestión. Se puede hacer la comprobación obteniendo la derivada con respecto a t de la curva geodésica $\gamma_{p,v}(t)$, asociada a cada parametrización, y luego calculando la norma al cuadrado de dicha derivada. El resultado que se obtendrá será uno y $\|v\|^2$, respectivamente.

A continuación se mostrará que las parametrizaciones exhibidas radican en la esfera de radio uno. La prueba consiste en verificar que la norma al cuadrado de la curva geodésica, bajo cada parametrización, es uno. Para las pruebas se usarán propiedades del producto interno, y los siguientes hechos:

- $\|p\| = 1$; esto es cierto, ya que p es un punto que pertenece a la esfera unitaria.
- $\langle v, p\rangle = 0$, lo cual se sigue de la definición del plano tangente a un punto en la esfera.
- La norma del vector v para la primera parametrización es uno.

Caso 1: ecuación (2.3)

$$\begin{aligned}
\|\gamma_{p,v}(t)\|^2 &= \langle \gamma_{p,v}(t), \gamma_{p,v}(t)\rangle \\
&= \langle \cos(t)p + \sin(t)v, \cos(t)p + \sin(t)v\rangle \\
&= \langle \cos(t)p, \cos(t)p\rangle + 2\langle \cos(t)p, \sin(t)v\rangle + \langle \sin(t)v, \sin(t)v\rangle \\
&= \cos^2(t)\langle p, p\rangle + \sin^2(t)\langle v, v\rangle \\
&= \cos^2(t) + \sin^2(t) \\
&= 1.
\end{aligned}$$

Caso 2: ecuación (2.4)

$$\begin{aligned}
\|\gamma_{p,v}(t)\|^2 &= \langle \gamma_{p,v}(t), \gamma_{p,v}(t)\rangle \\
&= \langle \cos(t\|v\|)p + \sin(t\|v\|)\frac{v}{\|v\|}, \cos(t\|v\|)p + \sin(t\|v\|)\frac{v}{\|v\|}\rangle \\
&= \langle \cos(t\|v\|)p, \cos(t\|v\|)p\rangle + 2\langle \cos(t\|v\|)p, \sin(t\|v\|)\frac{v}{\|v\|}\rangle + \\
&\quad \langle \sin(t\|v\|)\frac{v}{\|v\|}, \sin(t\|v\|)\frac{v}{\|v\|}\rangle \\
&= \cos^2(t\|v\|)\langle p, p\rangle + 2\frac{\cos(t\|v\|)\sin(t\|v\|)}{\|v\|}\langle p, v\rangle + \frac{\sin(t\|v\|)}{\|v\|^2} \\
&= \cos^2(t\|v\|)\langle p, p\rangle + \frac{\sin^2(t\|v\|)}{\|v\|^2}\langle v, v\rangle \\
&= \cos^2(t\|v\|)\langle p, p\rangle + \frac{\sin^2(t\|v\|)}{\|v\|^2}\|v\|^2 \\
&= \cos^2(t\|v\|) + \sin^2(t\|v\|) \\
&= 1.
\end{aligned}$$

Por lo tanto queda comprobado que $\gamma_{p,v}(t)$, bajo las parametrizaciones dadas, está sobre la esfera unitaria.

En la Figura 2.6 se muestran algunos ejemplos de curvas geodésicas, para el esbozo de éstos se consideró $t \in (-pi, 0)$. Del lado izquierdo se tiene una curva geodésica que pasa por los puntos p y q, tal que $q \neq -p$; del lado derecho se muestran varias curvas geodésicas que pasan por p y $q = -p$. Es importante notar que en el segundo caso hay una infinidad de curvas geodésicas que pasan por p y $-p$, lo cual se debe a que $-p$ es el punto antípodo de p; es decir, $-p$ es el punto diametralmente opuesto a p. Aparentemente lo anterior es un hecho inocuo; sin embargo, adquirirá relevancia en un concepto geométrico que será tratado más adelante, así como en el Capítulo 3.

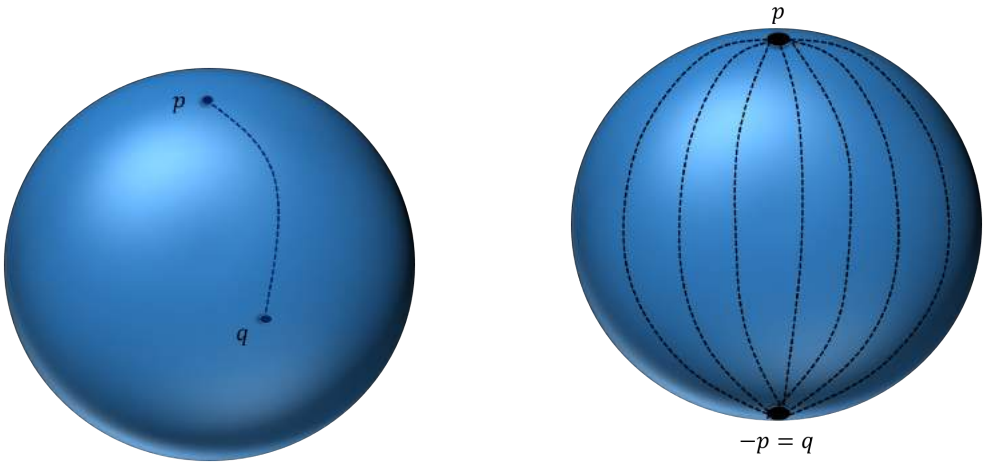


Figura 2.6: Curvas geodésicas en la esfera.

Para finalizar esta sección es importante comentar que existe una relación entre el plano tangente y las curvas geodésicas. Dicha relación es que las funciones diferenciables $\gamma(t)$ que ayudan a definir el plano tangente, son curvas geodésicas.

2.2.3. Mapeo exponencial

Esta noción geométrica permitirá llevar un punto del plano tangente a una variedad M. Se denota como $\exp_p(v)$, donde p es un punto que pertenece a la variedad M y v es un vector que pertenece al plano tangente T_pM. Formalmente, el mapeo exponencial se define a continuación.

Definición 2.2.6 *Sea $v \in T_pM$ y $p \in M$, entonces existe una única geodésica tal que*

$$\gamma_{p,v}(0) = p,\ \gamma'_{p,v}(0) = v\ y\ \exp_p(v) = \gamma_{p,v}(1) = \gamma_{p,\frac{v}{\|v\|}}(\|v\|). \tag{2.5}$$

Algunas propiedades que posee son:

- Preserva distancias, $d(p, \exp_p(v)) = \|v\|$, donde $d(\cdot,\cdot)$ representa la distancia intrínseca de la variedad.
- Es diferenciable y $\exp_p(0) = p$.
- Es un difeomorfismo en una vecindad alrededor de cero.

A nivel geométrico, el mapeo exponencial es un punto de la variedad M. Este punto se obtiene mediante el mapeo de una curva geodésica que inicia en un punto $p \in M$, de forma que la curva se recorre con una velocidad v en una unidad de tiempo.

Mapeo exponencial en la esfera

Está dado por la siguiente formulación

$$\exp_p(v) = \cos(\|v\|)p + \sin(\|v\|)\frac{v}{\|v\|},\ v \neq 0,$$

de forma que dicha expresión cumple con la Definición (2.5). Lo anterior se debe a que $\exp_p(v) = \cos(\|v\|)p + \sin(\|v\|)\,(v/\|v\|) = \gamma_{p,v}(1)$, donde $\gamma_{p,v}(1)$ corresponde a la parametrización (2.4) de las curvas geodésicas en la esfera. A continuación se probará que la parametrización del mapeo exponecial en la esfera produce puntos en la esfera unitaria. La prueba consiste básicamente en verificar que la norma al cuadrado del mapeo exponencial es uno, ya que la norma de cualquier punto $p \in S^2$ es uno. Para ésta se utilizarán las siguientes afirmaciones:

- Sea $v \in T_pS^2$, tal que $v \neq 0$ y $p \in S^2$; entonces, $\langle v, p\rangle = 0$.

- $\|p\| = 1$ para todo punto $p \in S^2$.

Se tienen las siguientes igualdades,

$$\begin{aligned}
\|\exp_p(v)\|^2 &= \langle \exp_p(v), \exp_p(v)\rangle \\
&= \langle \cos(\|v\|)p + \sin(\|v\|)\frac{v}{\|v\|}, \cos(\|v\|)p + \sin(\|v\|)\frac{v}{\|v\|}\rangle \\
&= \langle \cos(\|v\|)p, \cos(\|v\|)p\rangle + \langle \cos(\|v\|)p, \sin(\|v\|)\frac{v}{\|v\|}\rangle + \\
&\quad \langle \sin(\|v\|)\frac{v}{\|v\|}, \cos(\|v\|)p\rangle + \langle \sin(\|v\|)\frac{v}{\|v\|}, \sin(\|v\|)\frac{v}{\|v\|}\rangle \\
&= \cos^2(\|v\|)\langle p, p\rangle + 2\frac{\cos(\|v\|)\sin(\|v\|)}{\|v\|^2}\langle p, v\rangle + \frac{\sin^2(\|v\|)\langle v, \rangle}{\|v\|^2} \\
&= \cos^2(\|v\|)\langle p, p\rangle + \frac{\sin^2(\|v\|)\|v\|^2}{\|v\|^2} \\
&= \cos^2(\|v\|) + \sin^2(\|v\|) \\
&= 1.
\end{aligned}$$

Por tanto, queda comprobado que el mapeo exponencial produce puntos en S^2.

Resulta oportuno comentar que en el caso de la esfera, el mapeo exponencial está definido para todo punto p. La utilidad de hecho anterior se verá en Capítulo 3. Para concluir esta sección, la Figura 2.7 muestra geométricamente el mapeo exponencial en la esfera.

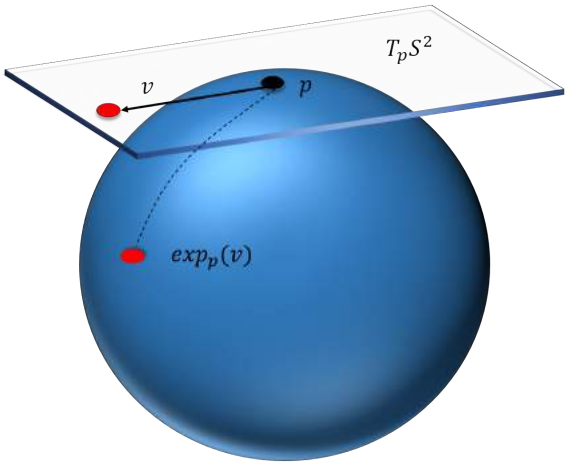


Figura 2.7: Mapeo exponencial en la esfera.

2.2.4. Mapeo logarítmico

También conocido como log-mapeo, se define como el inverso del mapeo exponencial; va de una variedad M al espacio T_pM y tiene las siguientes propiedades:

- $\log_p(p) = 0$ para todo punto $p \in M$.
- $d(p, q) = \|v\|$ para todo $p, q \in M$, donde $d(\cdot, \cdot)$ denota la distancia intrínseca de la variedad.

Se denota como $\log_p(q)$ o $\exp_p^{-1} q$, donde $p, q \in M$. En la presente tesis, y con la finalidad de evitar ambigüedades, se adoptará la notación $\log_p(q)$.

Intuitivamente, el mapeo logarítmico es un vector en el espacio tangente a un punto. Esta aseveración es natural, pues al ser la función inversa del mapeo exponencial, se sigue que esta formulación produzca vectores en dicho espacio.

Mapeo logarítmico en la esfera

Se formula como

$$\log_p(q) = \frac{\arccos(p'q)}{\sqrt{1-(p'q)^2}} \left[q - (p'q)p\right], \text{ tal que } q \neq p, -p, \tag{2.6}$$

donde $p'q = \langle p, q \rangle$. A continuación se verificará que la formulación (2.6) origina vectores en el plano tangente, para lo cual se utilizarán los siguientes resultados preliminares:

- Sea $p \in S^2$, entonces $\|p\|^2 = 1$.
- $\|p\|^2 = p \cdot p = \langle p, p \rangle$.

La prueba consiste en verificar que $\log_p(q) \cdot p = 0$. Utilizando propiedades del producto punto y los resultados preliminares, se tienen las siguientes igualdades:

$$\begin{aligned}
\log_p(q) &= \frac{\arccos(p'q)}{\sqrt{1-(p'q)^2}}\,[q-(p'q)p] \\
\log_p(q)\cdot p &= \left(\frac{\arccos(p'q)}{\sqrt{1-(p'q)^2}}\,[q-(p'q)p]\right)\cdot p \\
&= \frac{\arccos(p'q)}{\sqrt{1-(p'q)^2}}\,[q\cdot p-(p'q)(p\cdot p)] \\
&= \frac{\arccos(p'q)}{\sqrt{1-(p'q)^2}}\,[q\cdot p-(p\cdot q)(p\cdot p)] \\
&= \frac{\arccos(p'q)}{\sqrt{1-(p'q)^2}}\,[q\cdot p-(p\cdot q)] \\
&= 0.
\end{aligned}$$

Por lo tanto, $\log_p(q)$ da origen a vectores en T_pS^2.

Para finalizar la presente sección, en la Figura 2.8 se exhibe la representación del mapeo logarítmico en la esfera.

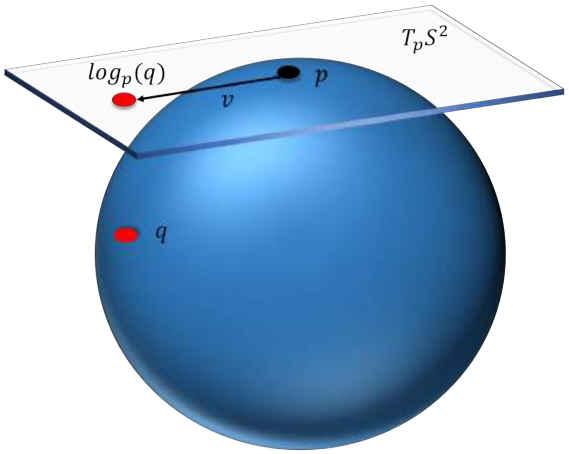


Figura 2.8: Mapeo logarítmico en la esfera.

2.2.5. Transporte paralelo

Será uno de los conceptos clave en el Capítulo 3. Éste permitirá llevar "paralelamente" vectores de un punto $p \in M$ a un punto $q \in M$, o bien, representar vectores de un espacio a otro. Dicha representación se caracterizará por tener la misma longitud y orientación que el vector original.

Para ejemplificar la intuición de este concepto, en la Figura 2.9 se muestra un transporte paralelo en $\mathbb{R}^2$. Del lado izquierdo se tiene un vector v cuyo origen es el punto

p; éste se desea transportar de manera "paralela" hacia el punto q, a lo largo de la recta definida por dichos puntos. Por otra parte, del lado derecho se tiene el transporte paralelo del vector v a lo largo de dicha recta.

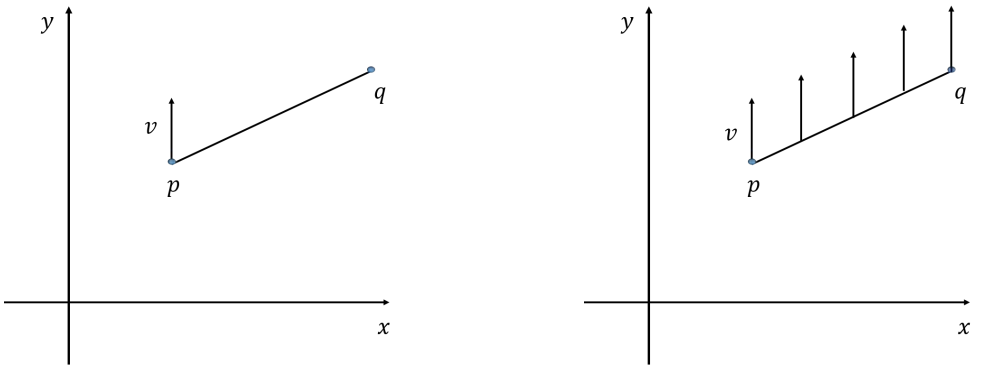


Figura 2.9: Transporte paralelo de un vector v a lo largo de una recta, cuyos extremos son los puntos p y q.

Como se puede notar, el transporte paralelo dio origen a un conjunto de vectores; éstos tienen la misma magnitud y dirección, además de ser paralelos entre sí. Por lo tanto, el transporte paralelo puede entenderse como mover un vector v de un punto $p \in M$ a un punto $q \in M$, de manera paralela a lo largo de una curva parametrizada definida en la variedad M.

En una variedad tan sencilla, como $\mathbb{R}^2$, es asequible entender la noción geométrica de transporte paralelo. Sin embargo en espacios más abtractos, como es el caso de las variedades no lineales, no es fácil entender dicha formulación.

Por otro lado, la herramienta teórica es más complicada, comparada con lo que se ha desarrollado hasta este momento. Por tal motivo, sólo se dará la intuición geométrica de qué es lo que permite hacer el transporte paralelo. Para dar la teoría a nivel intuitivo se introducirán las siguientes definiciones.

Definición 2.2.7 *Campo vectorial.*
Un campo vectorial X sobre una variedad diferenciable M, es una correspondencia que asocia a cada punto $p \in M$ un vector $X(p) \in T_pM$.

Definición 2.2.8 *Un campo vectorial X a lo largo de una trayectoria $\alpha(t)$, es una aplicación diferenciable $X : (-\epsilon, \epsilon) \longrightarrow \mathbb{R}^3$, tal que $X(t) \in T_{\alpha(t)}M$.*

Un ejemplo de campo vectorial a lo largo de una curva parametrizafa $\alpha(t)$, es el que se encuentra definido por $\dot{\alpha}(t)$, es decir, la derivada con respecto a t de $\alpha(t)$. Este campo vectorial se encuentra representado en la Figura 2.10.

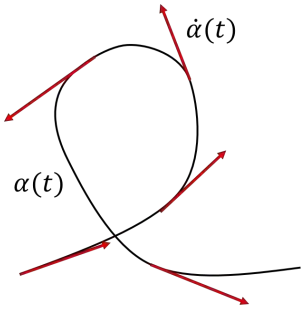


Figura 2.10: Campo vectorial definido por los vectores tangentes de la curva $\alpha(t)$.

En el caso de la presente tesis, será de vital importancia transportar paralelamente los vectores velocidad de una curva parametrizada $\alpha(t) \in M$ a través de geodésicas, hacia algún espacio tangente a un punto. Los elementos que usará este transporte paralelo son los siguientes:

1. Campo velocidad.
   Es un campo vectorial, el cual es el conjunto de vectores que será transportado. Este campo vectorial se obtendrá mediante la derivada con respecto a t de la trayectoria $\alpha(t)$.

2. Curvas geodésicas.
   Son las curvas parametrizadas sobre las que se realizará el transporte paralelo del campo velocidad.

3. Punto de referencia.
   Es un punto $c \in M$. Se caracteriza por ser el lugar donde se definirá el espacio tangente T_cM.

4. Espacio tangente a un punto.
   Es el espacio T_cM, lugar donde se transportará el campo velocidad.

A continuación se abordará un esbozo gráfico de cómo funciona el transporte paralelo, así como la forma en la que intervienen los elementos anteriomente enunciados. Por facilidad dichas representaciones se harán en la esfera.

Esbozo
Primero se hará la representación del transporte paralelo. En la Figura 2.11 se muestra una curva "parametrizada" $\alpha(t) \in S^2$ cuyos extremos son los puntos $p, q \in S^2$; también se muestra el vector $v \in T_pS^2$, cuyo origen es el punto p. El vector v es el que se desea transportar de manera paralela, al plano tangente que se definirá en el punto q.

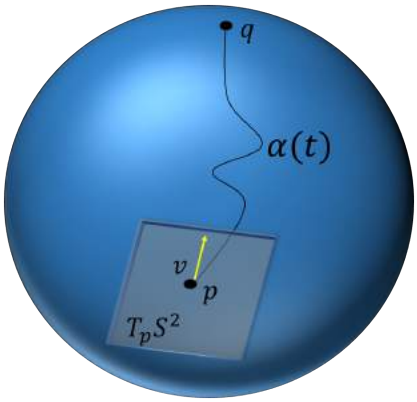


Figura 2.11: Vector $v \in T_pS^2$, el cual será transportado paralelamente a lo largo de la curva parametrizada $\alpha(t)$.

En la Figura 3.10 se muestra el resultado del transporte paralelo del vector v. En este caso, el transporte consistirá en encontrar una representación del vector v en T_qS^2, a la cual se le denotará como $\varphi(v)$ tal que $\varphi(v) \in T_qS^2$. Para encontrar $\varphi(v)$, se usará la curva parametrizada $\alpha(t)$ ya que ésta representa la conexión entre los puntos p y q, por lo cual, a lo largo de $\alpha(t)$ se irá identificando el vector v, mediante planos tangentes, hasta llegar al plano T_qS^2, como se muestra en la Figura 3.11.

La representación $\varphi(v)$ existe gracias a que hay un isomorfimo

$$\varphi : T_pS^2 \longrightarrow T_qS^2,$$

tal que φ es una función que preserva ángulos, longitudes y orientación. Es decir, para $u, v \in T_pS^2$ existe $\varphi(u), \varphi(v) \in T_qS^2$ tal que $\varphi(v) \cdot \varphi(v) = u \cdot v$, como se ejemplifica en la Figura 2.14. Es primordial notar que bajo este contexto, el transporte paralelo también definió un campo vectorial, el cual se encuentra representado por el conjunto de vectores amarillos en la Figura 3.11.

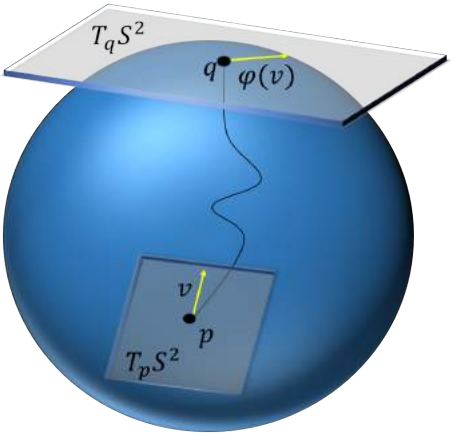


Figura 2.12: Representación del vector v en T_qS^2.

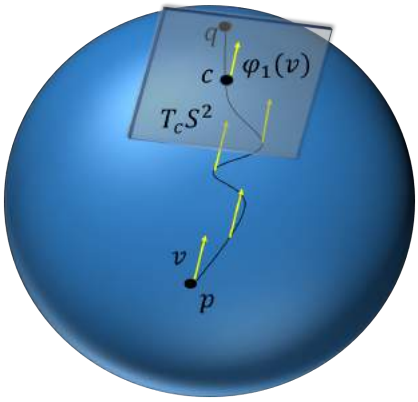


Figura 2.13: Transporte paralelo del vector v.

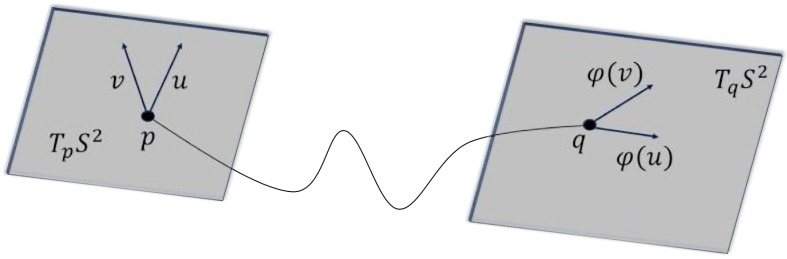


Figura 2.14: Representación de los vectores v y u.

A continuación se muestra cómo intervienen de manera conjunta, el campo velocidad $\dot{a}(t)$, las curvas geodésicas, el punto de referencia y el plano tangente en el transporte paralelo de una curva parametrizada $\alpha(t) \in S^2$. La Figura 2.15 ejemplifica $\alpha(t) \in S^2$ y $\dot{\alpha}(t)$ su campo velocidad representado por los vectores amarillos. Por otro lado, la Figura 2.16 muestra el plano tangente T_cS^2, lugar en el que se transportará el campo velocidad $\dot{\alpha}(t)$.

Es importante notar los dos detalles siguientes. Primero, recordar que $\alpha(t)$ necesita ser suave, ya que el campo velocidad se encuentra definido mediante la derivada de ésta. Segundo, el punto c sobre el cual se define el plano tangente, puede ser cualquier punto de la esfera.

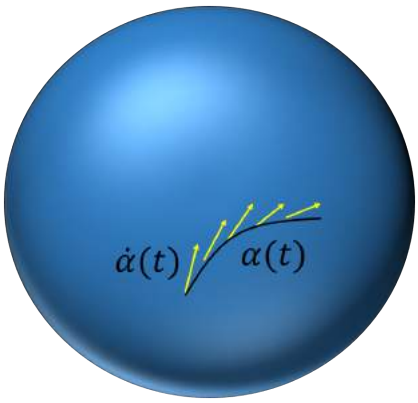


Figura 2.15: Curva $\alpha(t)$ y campo velocidad $\dot{\alpha}(t)$.

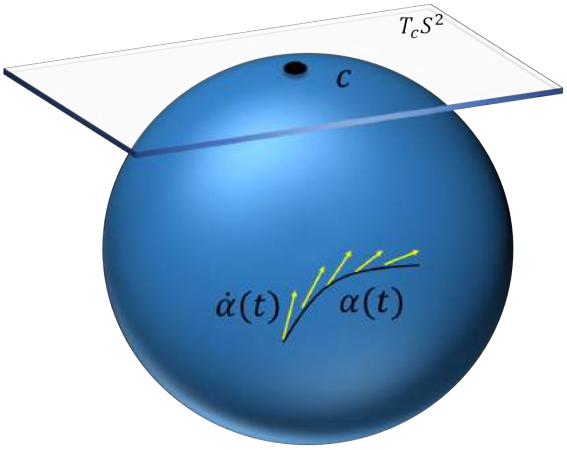


Figura 2.16: Plano tangente al punto c.

La Figura 2.17 ejemplifica el conjunto de curvas geodésicas que se usarán para transportar el campo $\dot{\alpha}(t)$, mientras que la Figura 2.18 muestra el transporte paralelo de los vectores velocidad a lo largo de las curvas geodésicas. Es relevante comentar que, para lograr este transporte paralelo, las curvas geodésicas deben de tener el mismo punto de fin, c.

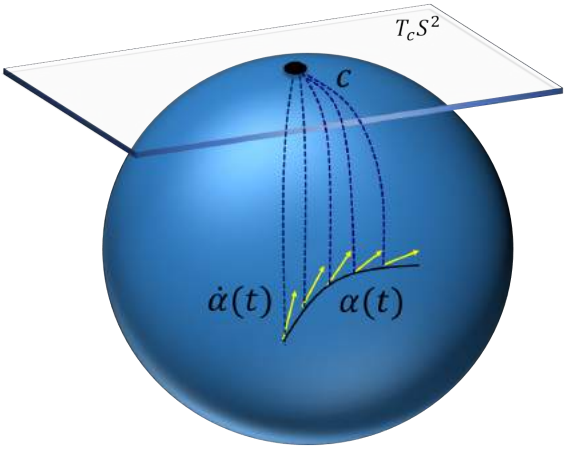


Figura 2.17: Curvas geodésicas con un punto c en común.

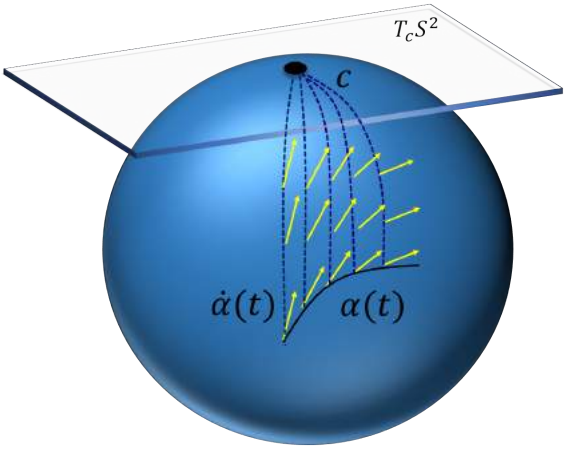


Figura 2.18: Transporte paralelo del campo velocidad $\dot{\alpha}(t)$.

Para finalizar el presente esbozo, la Figura 2.19 ejemplifica el transporte paralelo del campo velocidad $\dot{\alpha}(t)$ en T_cS^2, representado por los vectores rojos. Por otro lado la Figura 2.20 muestra una curva negra, la cual es la representación de la curva parametrizada $\alpha(t)$ en T_cS^2. Es importante notar que en este caso, cuando se transporta paralelamente un vector v en la esfera a lo largo de geodésicas, la representación del vector v en T_cS^2 queda ligeramente rotada. Es en este sentido que la noción de transportar "paralelamente" cambia de acuerdo con la variedad con la que se esté trabajando.

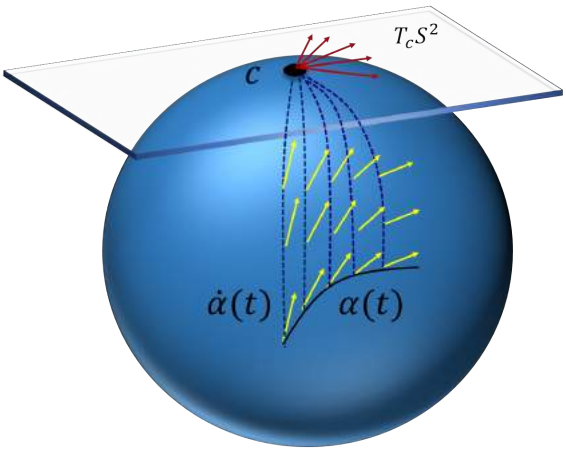


Figura 2.19: Representación del campo $\dot{\alpha}(t)$ en T_cS^2.

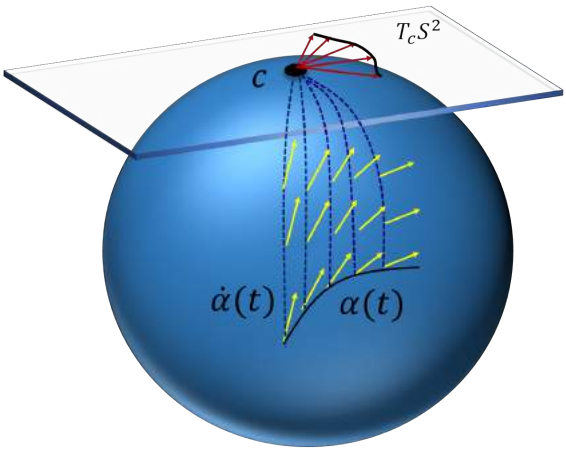


Figura 2.20: Representación de la curva parametrizada α en T_cS^2.

Una vez que se ha dado la intuición y elementos que usa el transporte paralelo, se abordará una modificación de éste, cuya utilidad será vista en el Capítulo 3. Dicha modificación es el *Transported Square Root Vector Field o TSRVF*. El TSRVF se puede interpretar como un transporte paralelo escalado el cual, de manera análoga al transporte paralelo anteriormente esbozado, da origen a campos vectoriales.

El TSRVF surge a partir de una extensión conceptual de $\mathbb{R}^n$, la cual puede ser revisada en Srivastava et al. [2011b]. En este artículo también se podrá encontrar la intuición del cómo y por qué surge ésta noción geométrica, así como su utilidad en la parte computacional. A este tipo de transporte se le denotará como $h_\alpha(t)$, donde $\alpha(t)$ es una trayectoria suave sobre la variedad M. Formalmente, el TSRVF se define como se muestra a continuación.

Definición 2.2.9 *Para cualquier trayectoria suave $\alpha(t) \in M$, el TSRVF es el transporte paralelo del campo vectorial de velocidades escaladas de una trayectoria $\alpha(t)$ a un punto de referencia $c \in M$ de acuerdo con*

$$h_\alpha(t) = \frac{\dot{a}(t)_{\alpha(t)\longrightarrow c}}{\sqrt{|\ \dot{\alpha}(t)\ |}} \in T_c M. \tag{2.7}$$

De la definición anterior, $|\cdot|$ denota la norma relacionada con la métrica intrínseca de la variedad M, $\dot{a}(t)$ denota la derivada de la curva $\alpha(t)$ con respecto a t y $\alpha(t) \longrightarrow c$ repesenta la geodésica que va de $\alpha(t)$ a c.

Es conveniente y relevante aclarar que lo que se transporta paralelamente, no es la posición de la curva parametrizada, sino su velocidad; por ende, lo que se tiene en el plano tangente es una representación de la velocidad de la trayectoria. Por tal motivo, para recuperar la posición de $\alpha(t)$ es necesario resolver una ecuación diferencial, la cual incorporará el punto donde inicia dicha curva parametrizada y el transporte paralelo $h_\alpha(t)$. La ecuación diferencial a resolver es

$$\dot{\beta}(t) =| \ V_{\beta(t)}(t) \ | \ V_{\beta(t)}(t), \tag{2.8}$$

tal que $V_{\beta(t)} = (h_\alpha(t))_{c\longrightarrow\beta(t)}$. Es decir, $V_{\beta(t)}$ es el campo vectorial inducido por el transporte paralelo $h_\alpha(t)$, a través de la curva geodésica que empieza en c y termina en $\beta(t)$, tal que $\beta(0) = \alpha(0) \in M$. De esta forma es que la curva parametrizada resultante $\beta(t)$ será exactamente la curva parametrizada original $\alpha(t)$. En otras palabras, lo que se está haciendo es un transporte paralelo—TSRVF—de regreso.

A continuación se ejemplifica el regreso del TSRVF en la esfera. La Figura 2.21 muestra el transporte paralelo $h_\alpha(t)$, el campo vectorial $V(t)$ que dibuja y el punto $\alpha(0)$, tal que $V(t)$ se encuentra representado por los vectores amarillos. Por otro lado, la Figura 2.22 exhibe un conjunto de curvas geodésicas que parten del punto c, y con éstas se identificará el campo vectorial $V_{\beta(t)}$ que coincide con $V(t)$. De esta forma, al resolver la ecuación (2.8) se obtiene la curva parametrizada $\beta(t)$.

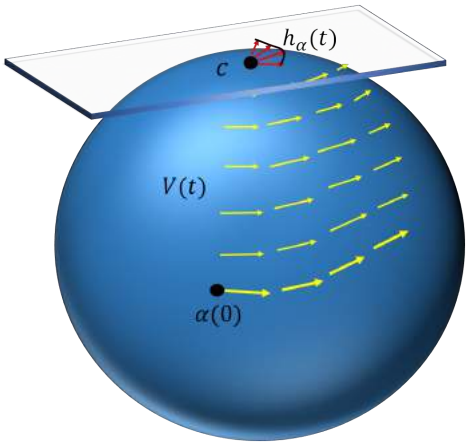


Figura 2.21: Campo vectorial $V(t)$.

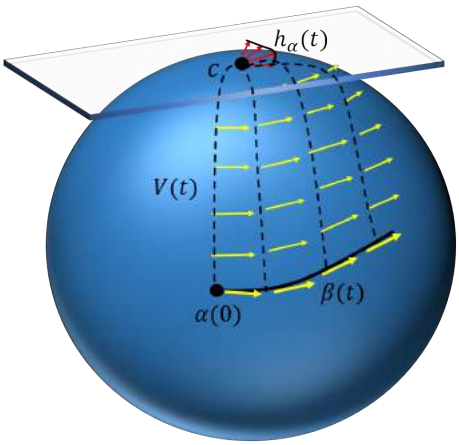


Figura 2.22: Regreso del TSRVF.

Transporte paralelo en la esfera

Tiene la siguiente definición:

Definición 2.2.10 *Sean p y q dos puntos en S^2, tal que $p \neq q$ y v un vector en T_pS^2. El transporte paralelo $v_{p\longrightarrow q}$, a lo largo de la curva geodésica que va de p a q, está dado por*

$$v - \frac{2\langle v, q\rangle}{\mid p+q \mid^2}(p+q). \tag{2.9}$$

En la presenta definición $\mid \cdot \mid$ representa a la norma euclidiana.

Es oportuno notar que, en este contexto, se desea hacer el transporte paralelo de T_pS^2 a T_qS^2. Por lo tanto, la Definición 2.2.10 produce vectores en T_qS^2.

Por otro lado, para que el transporte paralelo sea único, $\alpha(t)$ no debe pasar por $-q$. Este hecho se debe a que $-q$ es el punto antípodo del lugar donde se definió T_qS^2, que es el plano donde se transportará el vector velocidad v. Para aclarar ideas, es importante recordar que hay una infinidad de curvas geodésicas que van de $-q$ a q. Por lo tanto, existiría una infinidad de posibles representaciones del vector v en T_qS^2, lo cual conllevaría que el transporte paralelo no sea único. A continuación se tiene la prueba de que (2.9) ofrece vectores que viven en T_qS^2. Para dicha tarea se tienen algunos resultados preliminares:

1. $\mid q \mid^2 = \langle q, q\rangle = 1$.

2. $\mid p+q \mid^2 = \langle p+q, p+q\rangle = 2 + 2p \cdot q$.

La prueba consiste en verificar que $\left(v - (2\langle v, q\rangle / \mid p+q \mid^2)\,(p+q)\right)\cdot q = 0$. Entonces, utilizando propiedades del producto punto y los resultados preliminares, se tienen

las siguientes igualdades:

$$\begin{aligned}
w &= v - \frac{2\langle v, q\rangle}{\mid p+q \mid^2}(p+q) \\
w \cdot q &= v \cdot q - \frac{2\langle v, q\rangle}{\mid p+q \mid^2}(p \cdot q + q \cdot q) \\
&= v \cdot q - \frac{2(v \cdot q)}{\mid p+q \mid^2}(p \cdot q + \mid p+q \mid^2) \\
&= v \cdot q \left(1 - \frac{2(p \cdot q+1)}{\mid p+q \mid^2}\right) \\
&= v \cdot q \left(\frac{\mid p+q \mid^2 -2(p \cdot q+1)}{\mid p+q \mid^2}\right) \\
&= v \cdot q \left(\frac{2+2p \cdot q-2p \cdot q-2}{\mid p+q \mid^2}\right) \\
&= 0.
\end{aligned}$$

Por lo tanto, w es un vector que está en T_qS^2. Con esto queda comprobado que el transporte paralelo, en la esfera, produce vectores en el plano tangente.

Para finalizar esta sección, en la Figura 2.23 se muestra el transporte paralelo de dos curvas geodésicas en la esfera, mientras que en la Figura 2.24 se exhibe el transporte de una trayectoria paralela en la esfera. Para la realización de las curvas geodésicas, en la Figura 2.23, se usó la expresión (2.3) la cual se encuentra en la Sección 2.2.2 de la presente tesis. La curva geodésica roja requirió los parámetros $v_1 = \left(1/\sqrt{2}, 0, 1/\sqrt{2}\right)$ y $p_1 = (.0028, .9999, .000116)$, mientras que la azul necesitó $v_2 = (0, 1, 0)$ y $p_2 = (1, 0, 0)$. Para el transporte paralelo de ambas trayectorias se usó la ecuación (2.9) tomando $q = (0, 0, 1)$, se derivó la expresión (2.3) para obtener los vectores v y los puntos p se tomaron de la evaluación de los parámetros v_1, p_1, v_2 y p_2 en la ecuación (2.3).

La Figura 2.24 muestra el transporte paralelo de la curva parametrizada

$$\alpha(t) = \frac{1}{2}\left(\sin(t), \cos(t), \sqrt{3}\right), \quad -\pi \leq t \leq \pi.$$

Para realizarlo se derivó $\alpha(t)$, con la finalidad de obtener el campo velocidad. Para los puntos p se tomó la evaluación de $t \in [-\pi, \pi]$ en $\alpha(t)$, y al igual que en el caso de las curvas geodésicas se consideró $q = (0, 0, 1)$. El algoritmo de cómo se programó el transporte de las trayectorias geodésicas y paralelas puede ser consultado en el Anexo A de la presente tesis.

Curvas Geodésicas.

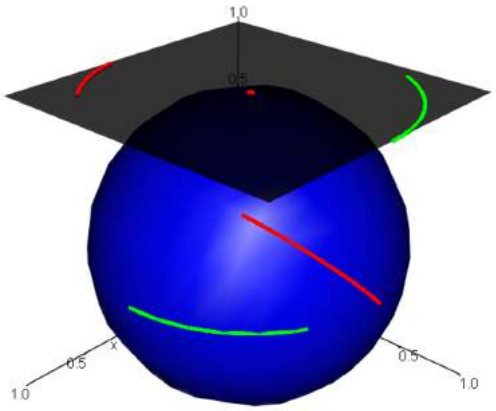


Figura 2.23: Transporte paralelo de dos curvas geodésicas.

Curva paralela.

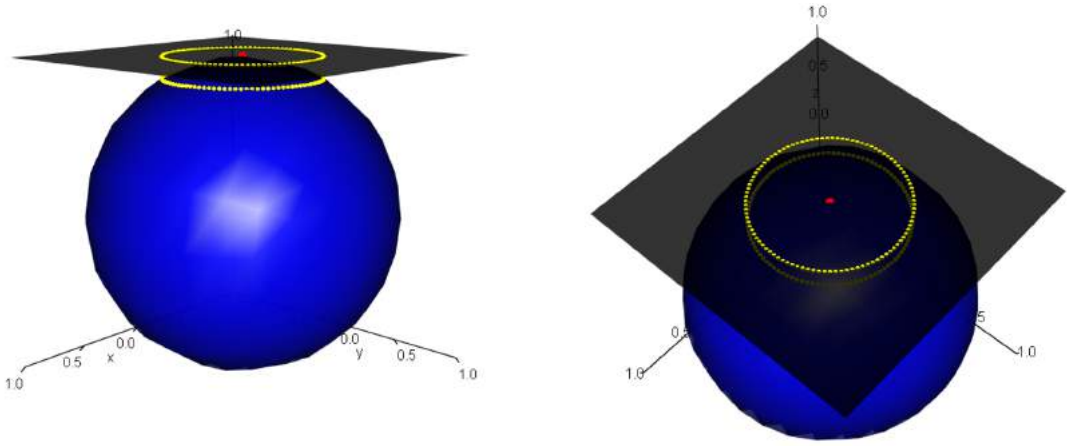


Figura 2.24: Transporte paralelo de una curva paralela.

2.3. Epílogo

Para cerrar este capítulo, se tienen los siguientes comentarios:

- La teoría abordada en el presente capítulo se puede extender a variedades más complejas que la esfera. Por ejemplo, SE(2) y el espacio de formas, variedades que fueron vistas en el Capítulo 1.
- Mientras más compleja sea la variedad con la que se esté trabajando, más difícil será obtener expresiones analíticas para los conceptos anteriormente tratados.

Éste es uno de los elementos que complica el estudio estadístico en variedades.

Algunas lecturas adicionales que se recomiendan, para profundizar la teoría vista, son:

1. Do Carmo Valero [1992] *Riemannian geometry.*
   Principalmente aborda nociones maduras de geometría diferencial. Además cuenta con un capítulo introductorio, el cual contiene todas las nociones básicas necesarias para entender el contenido del mismo. Por otro lado, ofrece una introducción a la teoría de variedades riemannianas y sus propiedades. En ésta se puede revisar el tema de curvas geodésicas y espacio tangente a un punto.

2. Do Carmo [1976] *Differential geometry of curves and surfaces.*
   Ofrece una introducción a la teoría de variedades desde principios básicos, usando herramientas de cálculo diferencial en $\mathbb{R}^n$. En la presente referencia se puede revisar de manera detallada la parte de transporte paralelo.

3. Lee [2006] *Riemannian manifolds: an introduction to curvature.*
   Es una introducción a la teoría de variedades; sin embargo, aborda y usa nociones más profundas que Do Carmo [1976]. Esta referencia es excelente para aquellos que han llevado cursos de topología y tienen conocimientos básicos de variedades en $\mathbb{R}^n$. En dicho texto se puede revisar lo concerniente a mapeo exponencial.

4. Loring [2008] *An introduction to manifolds.*
   La presente cita da una introducción a la teoría de variedades. Comienza con una breve recapitulación de conceptos de cálculo diferencial en $\mathbb{R}^n$, para luego abordar la teoría de geometría diferencial desde principios básicos, ayudándose con ejemplos ilustrativos. Esta referencia se distingue de las otras por abarcar nociones complejas de geometría diferencial, de una manera accesible e intuitiva para el lector. También cuenta con una parte histórica que ameniza la lectura de los capítulos. En ésta referencia se puede revisar la parte que corresponde a campos vectoriales.

Finalmente, basta comentar que en el presente capítulo se dieron las nociones teóricas primordiales, tanto a nivel técnico como intuitivo, de geometría diferencial. Como se verá, dichas nociones serán vitales para entender el desarrollo del Capítulo 3.

Capítulo 3

Análisis estadístico de trayectorias sobre la esfera

3.1. Introducción

Este capítulo contiene un resumen estructurado del artículo Su et al. [2014a] que lleva por título *Statistical analysis of trajectories on Riemannian manifolds: bird migration, hurricane tracking and video surveillance*. El artículo se considera como base para el desarrollo de la presente tesis, por su novedosa incursión en el análisis estadístico de trayectorias sobre variedades. Éste es innovador en el sentido de que ofrece un cambio de paradigma para el análisis estadístico de trayectorias, al incorporar tiempos aleatorios y trabajar sobre variedades. Además, dicho artículo logra una sinergia entre nociones de geometría diferencial con probabilidad y estadística, para luego proponer un análisis estadístico sobre variedades. Esta propuesta consiste de dos estapas. La primera es encontrar una trayectoria media y cuantificar la varianza asociada a un conjunto de trayectorias. La segunda es considerar dichos parámetros en un modelo de probabilidad con la finalidad de realizar inferencia estadística mediante simulaciones.

El resumen que será expuesto a continuación aplicará ideas y conceptos tratados en el Capítulo 2. Se ofrecerán comentarios esporádicos, que complementarán y aclararán conocimientos obviados en el artículo base. Finalmente, informará sobre recomendaciones bibliográficas que afianzarán la teoría desarrollada. El resumen se organiza en tres grandes secciones. Esta estructura obedece a una propuesta propia, que resulta de analizar el contenido del artículo con la intención de facilitar su presentación. Las secciones son:

- Trayectorias[1]: Aborda brevemente el entendimiento de la complejidad del dato. Esta sección comentará las ventajas y desventajas que existen al hacer un análisis estadístico clásico de trayectorias, comparado con el análisis propuesto por el artículo base. La importancia que tiene esta sección es perfilar lo que se

[1]En este capítulo se hablará de trayectoria bajo la Definición 2.2.1, establecida en el Capítulo 2.

ha desarrollado para el análisis estadístico de trayectorias y exhibir las ventajas estadísticas que se ganan al considerar un nuevo enfoque.

- Trayectorias como objeto matemático: Introduce la notación necesaria para estudiar las trayectorias desde la perspectiva de geometría diferencial. Así mismo, explicará los pasos previos al análisis estadístico, que a la postre serán cruciales para la comparación de trayectorias a través de cierta medida. La importancia de esta sección es, motivar la necesidad de abordar nociones de geometría diferencial para incorporarlas en el estudio estadístico de trayectorias y mostrar cómo esta herramienta determina y permite la estadística sobre variedades.

- Análisis estadístico de trayectorias: Será una combinación de lo que se desarrolló en las dos secciones previas. De manera concreta se tratarán los algoritmos para obtener la media de un conjunto de trayectorias, así como la varianza asociada a éste. Además, se enunciará un algoritmo para obtener la densidad de probabilidad de una trayectoria y se explicarán las aplicaciones que puede tener. Por lo tanto, y con el objetivo de consolidar dichos algoritmos, se desarrollará un pequeño ejemplo de simulación de trayectorias de huracanes. La finalidad de esta sección es exponer los pasos que se deben seguir para implementar un estudio estadístico de trayectorias sobre variedades riemannianas, particularmente sobre la esfera.

Es relevante comentar que el resumen que será desarrollado a continuación no aborda el tratamiento de algunos casos de estudio, los cuales son clustering de trayectorias de vehículos y análisis de clasificación de siluetas de video. Lo anterior se debe a la especialización temática en la esfera, la cual fue adoptada en el Capítulo 1.

3.2. Trayectorias

Como fue establecido en el Capítulo 1, analizar trayectorias desde una perspectiva estadística posee relevancia y dificultad. La relevancia dependerá del problema que se quiera resolver, mientras que la dificultad radicará principalmente en la estructura del dato. Un ejemplo de esto es la cantidad de observaciones que tiene la trayectoria, la velocidad con la que fue recorrida, el tiempo que se dejó entre el asiento de cada observación, etc. Por tanto estas características imponen un reto, pues se requiere una metodología que permita estudiar un conjunto de trayectorias con las características enunciadas, de tal manera que ésta no pierda de vista la estructura intrínseca que tienen los datos; por ejemplo, su forma. Por ende es necesario encontrar un enfoque que habilite el estudio estadístico de trayectorias, de manera que se desperdicie la menor cantidad de información.

El análisis estadístico de trayectorias ha sido emprendido con diferentes perspectivas. Una de ellas versa en el estudio del tiempo con el fue recorrida la trayectoria. Este enfoque, a su vez, se divide en dos vertientes: considerar tiempos aleatorios o no aleatorios en el estudio estadístico. La segunda vertiente es la más común y se

clasifica dentro del análisis estadístico tradicional de trayectorias. Por tanto, a continuación se exponen las ventajas y desventajas que se obtienen al realizar un análisis estadístico de trayectorias, considerando tiempos no aleatorios. La finalidad de dicha exposición es evaluar las facilidades y contratiempos que ofrece dicho planteamiento.

Ventajas:

1. El análisis estadístico es sencillo, ya que no existe necesidad de recurrir a nueva teoría que vaya más allá de las nociones estándar de probabilidad y estadística.

2. La parte computacional es accesible, pues existen paqueterías implementadas. Por ejemplo, la paquetería Trajectories del software R.

Desventajas:

1. La trayectoria media, o *cross sectional mean*, no es representativa. Esta trayectoria se puede interpretar como una media puntual de un conjunto de trayectorias. Se consigue tomando k puntos representativos de cada una de las trayectorias en estudio—la elección de tales puntos dependerá del experto estadístico o del espacialista en el área—y posteriormente se promedia el n-ésimo punto de todas las trayectorias, tal que $n = 1, \ldots, k$.

2. La varianza puntual, o *cross sectional variance*, se encuentra inflada. Ésta cuantifica, puntualmente, qué tan distantes están las trayectorias entre sí. Para su cálculo se requiere la trayectoria media y el conjunto de trayectorias observadas. De manera general, los pasos son:

   a) Considerar k puntos representativos en cada trayectoria, así como en la trayectoria media.

   b) Tomar como lugar de referencia el i-ésimo punto de la trayectoria media y calcular la distancia de este punto al i-ésimo punto de cada trayectoria.

   c) Obtener el promedio de las distancias calculadas.

   Éstos tres pasos se repiten para los $k - 1$ puntos restantes.

3. El análisis estadístico es pobre, debido a que la media y la varianza no son representativas, en el sentido de que no capturan el comportamiento de los datos; por ejemplo la forma intrínseca de éstos.

El origen de las desventajas anteriores es que las trayectorias en estudio no transcurren a la misma velocidad. Por ende, cada trayectoria está constituida por una cantidad de observaciones diferentes. Para hacer comparables las trayectorias se eligen puntos representativos de éstas, y como resultado de ello es que todas las trayectorias tienen la misma cantidad de observaciones. Sin embargo, tal procedimiento conlleva a una pérdida de información y por lo tanto una pérdida respecto a la estructura de los datos. El hecho anterior es el factor que influye en que la trayectoria media no refleje el comportamiento de las trayectorias individuales, así como en el incremento de la varianza.

En este contexto la trayectoria media es el equivalente a la media muestral de un conjunto de observaciones, tal que las observaciones en este caso son trayectorias, de manera análoga ocurre con la varianza puntual de un conjunto de trayectorias.

A raíz de los problemas anteriores es que surgió la necesidad de estudiar otras herramientas y puntos de vista, como el que expone el artículo de Su et al. [2014a]. El enfoque que considera es el estudio de trayectorias ocupando tiempos aleatorios. Dicho planteamiento se puede motivar con la migración de aves y el seguimiento de huracanes. En el caso de la migración de aves, a pesar de que una parvada siga la misma curva, no necesariamente vuela con la misma velocidad. Lo mismo ocurre con los huracanes; dos huracanes pueden tener la misma curva, y sin embargo pueden estar asociados a diferentes intensidades de recorrido y corresponder a diferentes años de registro. Esto quiere decir que se involucra cierta aleatoriedad temporal al observar las trayectorias. En consecuencia, al incorporarla en un estudio estadístico, se observan resultados que hacen más sentido con la intuición. No obstante, dado el reciente desarrollo de esta teoría, presenta algunas dificultades las cuales serán enlistadas junto con sus bondades. Es importante mencionar que esta perspectiva de estudio constituye una de las principales aportaciones del artículo.

Ventajas:

1. La trayectoria media es representativa.

2. La varianza puntual es menor, comparada con la del análisis clásico.

3. Se deriva una caracterización probabilística de una trayectoria, con base en los dos parámetros anteriores.

Desventajas:

1. El análisis estadístico es complicado, inducido principalmente por los diferentes registros de hora y fecha, así como por las diferentes velocidades que caracterizan a las curvas.

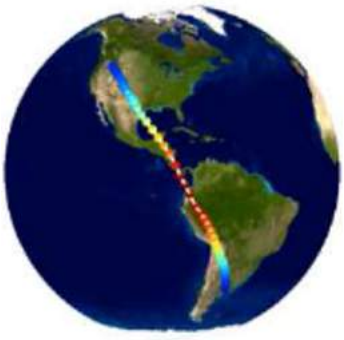
2. Hay poca teoría desarrollada. Por lo tanto existe la necesidad de explorar nuevas herramientas, las cuales en su mayoría no están adecuadas para ser usadas en un estudio estadístico.

3. El cómputo es más complejo, pues requiere la implementación de conceptos y teoría desde principios básicos.

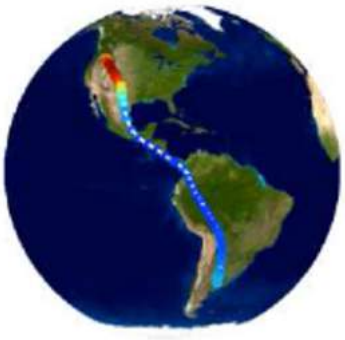
A continuación se muestra un análisis gráfico comparativo de los dos enfoques expuestos. Se tomó como caso de estudio un conjunto conformado por 35 trayectorias que representan la migración del halcón de Swainson. Estas trayectorias fueron observadas durante el período que comprende de 1995 a 1997. Las Figuras 3.1, 3.2 y 3.3 muestran, respectivamente, el conjunto de trayectorias de esta especie durante

su período de migración, la trayectoria media y la varianza asociada a dicho conjunto. Es relevante aclarar que antes de aplicar las metodologías ya comentadas, las trayectorias en cuestión pasaron por un proceso previo de interpolación[2].

Figura 3.1: Conjunto de trayectorias del halcón de Swainson durante su época de migración.



(a) Trayectoria media con el análisis tradicional.



(b) Trayectoria media considerando tiempos aleatorios.

Figura 3.2: Trayectoria media del halcón Swainson.

[2]Las imágenes que se utilizaron en el análisis comparativo de metodologías fueron tomadas del artículo Su et al. [2014a].

(a) Elipses que representan la varianzas puntuales con el análisis tradicional.

(b) Círculos que representan la varianzas puntuales considerando tiempos aleatorios.

Figura 3.3: Varianzas puntuales asociadas al conjunto de trayectorias del halcón Swainson.

Como se puede apreciar, para esta muestra de trayectorias, el análisis propuesto por Su et al. [2014a] arroja resultados que concuerdan con la intuición estadística. Lo anterior es en el sentido de que la curva o traza asociada a la trayectoria media se encuentra acorde con la curva de las trayectorias individuales, a diferencia de la media que se obtuvo vía el análisis clásico. De esa misma forma, las varianzas puntuales crecen conforme las trayectorias se van desfasando entre sí, contrariamente a las varianzas que se obtienen con el enfoque tradicional. Éstos resultados muestran que en un estudio estadístico de trayectorias—las cuales poseen variabilidad temporal y una forma particular—el desarrollo de la teoría propuesta por el artículo citado es pertinente.

Para concluir esta sección se anotan algunas recomendaciones bibliográficas, en las cuales se puede consultar más acerca de los enfoques expuestos.

1. Jupp and Kent [1987]. *Fitting smooth paths to speherical data.*
   Explica las limitantes que tiene el análisis clásico de trayectorias. Por otra parte, aborda la problemática que existe al interpolar puntos en una trayectoria discreta, por ejemplo vía splines, cuando hay errores de medición o los tiempos son desconocidos. Su principal aportación es lograr interpolar datos sobre la esfera.

2. Liu and Müller [2004]. *Functional convex averaging and synchronization for time-warped random curves.*
   Aclara las desventajas de considerar un análisis estadístico con la *cross sectional mean* y la *cross sectional variance*, cuando los datos de interés contemplan una variación temporal. Por tal motivo proporciona algunas referencias, en las cuales se puede profundizar por qué un análisis puntual es inadeacuado para este tipo de datos. Bajo esta línea de pensamiento, aborda algunas metodologías para tratar la variabilidad temporal en trayectorias, con el objetivo de encontrar el mejor enfoque para el tratamiento de éstas.

3.3. Trayectorias como objeto matemático

Un conjunto de trayectorias puede radicar en diversos espacios, y ejemplos diversos de ello fueron esbozados en el Capítulo 1. Éstos espacios se caracterizaban por ser no lineales. Por tal motivo, para estudiar trayectorias en este contexto estadístico, que es más general al usual, será necesario abordarlas—desde principios básicos—con un enfoque matemático y la herramienta para lograrlo será proporcionada por la geometría diferencial.

A continuación se facilitará la notación con la que serán referidas las trayectorias en estudio, así como las propiedades que éstas poseen. Las trayectorias consideradas serán denotadas como $\alpha(t)$ y radicarán en una variedad riemanniana M, tal que $\alpha(t) : [0,1] \longrightarrow M$. Recuérdese que una variedad riemanniana es una variedad diferenciable, la cual está equipada con un producto interno. Seguidamente, se define a $\mathcal{M}$ como el conjunto de trayectorias suaves, de manera que $\alpha(t) \in \mathcal{M}$; es decir, $\mathcal{M} = \{\alpha(t) : [0,1] \longrightarrow M \mid \alpha(t) \text{ es suave}\}$. Por consiguiente, toda trayectoria en $\mathcal{M}$ es derivable, lo que conlleva al hecho de que cada una de ellas tendrá asociado un campo velocidad. Este punto es vital, ya que se requerirá para el uso del transporte paralelo. Finalmente, a la derivada de la trayectoria $\alpha(t)$ se le denotará como $\dot{\alpha}(t)$, de forma que $d\alpha(t)/dt = \dot{\alpha}(t)$.

Las trayectorias que se estudiarán poseen básicamente dos caraterísticas vitales. Éstas son que $\alpha(t)$ tiene asociada una variabilidad temporal y no es "directamente observable". La primera característica fue explicada en la Sección 3.2 de este capítulo. La segunda característica quiere decir que lo que se observa realmente es la velocidad con la que se recorre la trayectoria $\alpha(t)$. Es términos geométricos, sólo se observa una de las posibles parametrizaciones que puede tener la curva asociada a $\alpha(t)$—revisar Definición 2.2.2.

Para aclarar la idea anterior se postula el siguiente ejemplo. Supóngase que en una carrera de motocicletas a los concursantes se les coloca un gps que indica su posicionamiento cada dos segundos. La Figura 3.4 muestra las posiciones de los motociclistas, mientras que la Figura 3.5 exhibe el recorrido conjunto de éstos.

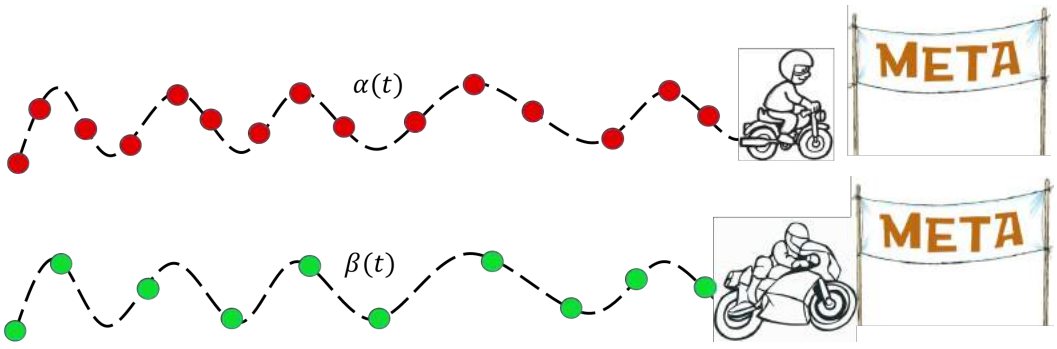


Figura 3.4: Posicionamiento de los motociclistas y trayectorias del recorrido.

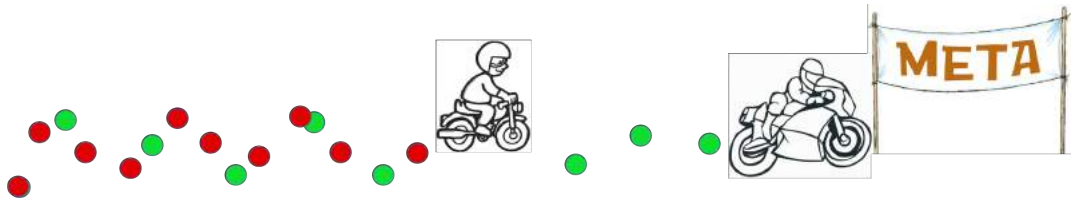


Figura 3.5: Recorrido de los motociclistas registrado por gps.

De acuerdo con la Figura 3.5, el concursante que dejó los puntos verdes llevaba mayor velocidad que el concursante de los puntos rojos, por lo cual fue observado menos veces durante la carrera. En consecuencia podría pensarse que las huellas que dejan los concursantes son diferentes. Sin embargo, la Figura 3.4 muestra que los dos motociclistas dejaron la misma traza en el recorrido, lo cual sigifica que $\alpha(t)$ y $\beta(t)$ representan a la misma curva aunque sus parametrizaciones sean diferentes. En este sentido es que se dice que las trayectorias no son directamente observables.

Es pertinente comentar que en este caso, al tratarse de un ejemplo didáctico, se sabe que la forma de las trayectorias es igual y por lo tanto representan a la misma curva. Sin embargo en datos reales, se desconoce en principio la huella que tiene una trayectoria. Por consiguiente se requieren técnicas de interpolación para determinar la forma o huella que poseen las trayectorias. Un texto que aborda el ajuste de curvas sobre variedades riemannianas es Samir et al. [2012].

Uno de los puntos a determinar en una muestra de trayectorias es identificar qué tan diferentes son entre sí. Para lograr esta tarea será necesario "estandarizarlas" vía una transformación temporal. La modificación o transformación temporal de trayectorias, también conocida como registro temporal, requerirá de una función conocida como *warping function*. Ésta, para fines de la tesis, será denominada como función de deformación temporal. La función de deformación temporal se interpretará como una reparametrización de una trayectoria $\alpha(t)$, de forma que modelará la variabilidad del tiempo en $\alpha(t)$.

La función de deformación temporal se caracterizará por ser una función desconocida y estrictamente creciente, la cual será denotada como $\gamma(t)$ tal que $\gamma : [0,1] \longrightarrow [0,1]$. Además se tiene que $\gamma \in \Gamma$, donde Γ es el conjunto de todas las orientaciones que preservan difeomorfismos[3] del intervalo $[0,1]$. Es decir, $\Gamma = \{\gamma : [0,1] \longrightarrow [0,1] \mid \gamma(0) = 0, \gamma(1) = 1, \gamma \text{ es difeomorfismo}\}$.

Para esclarecer ideas a continuación se desarrollará un ejemplo de cómo trabaja la función $\gamma(t)$. La Figura 3.6 muestra tres trayectorias de las cuales se conoce el recorrido hecho y su velocidad. Estas trayectorias corresponden al correcaminos, al coyote y a un conductor. Los puntos de color, en cada trayectoria, están asociados a las observaciones realizadas, de manera que el coyote y el correcaminos imprimen velocidades variables mientras que el automovilista lleva una velocidad constante.

[3]Un difeomorfismo es una función que tiene inversa y es diferenciable.

Lo anterior se traduce en que habrá tres funciones de deformación temporal; éstas son $\gamma_1(t)$ asociada a la trayectoria del correcaminos, $\gamma_2(t)$ asociada al coyote y $\gamma_3(t)$ asociada al conductor. Supóngase que el tiempo de recorrido en las tres trayectorias es el intervalo $[0, 1]$. Al tiempo $t = 0$ los tres personajes han recorrido un porcentaje nulo de su trayectoria total, por consiguiente las funciones $\gamma_1(t)$, $\gamma_2(t)$ y $\gamma_3(t)$ tienen el mismo punto de inicio. En el caso del automovilista, que es el que muestra la velocidad constante, al tiempo $t = .2$ habrá recorrido el 20 % de su trayectoria total, al tiempo $t = .4$ habrá recorrido el 40 % de su trayectoria total y así sucesivamente. De esa manera al tiempo $t = 1$ habrá recorrido el 100 % de su trayectoria; por dende, $\gamma_3(1) = 1$. El hecho anterior se traduce en que la función de deformación temporal para el carro es lineal, tal como muestra la Figura 3.7 con la curva $\gamma_3(1)$.

Comparando la trayectoria del coyote con la del automovilista, se observa que al tiempo $t = .2$ la velocidad del coyote es menor que la del carro. Por lo tanto, en dicho tiempo el coyote ha recorrido un menor porcentaje de su trayectoria total. Esto equivale a que el segundo punto verde, corespondiente al recorrido del coyote, quede por debajo del punto rosa que está asociado al recorrido del vehículo.

Por otra parte, el correcaminos al tiempo $t = .2$ lleva una velocidad mayor a la del conductor, casi el doble. Por ende en dicho tiempo, el correcaminos ha recorrido un mayor porcentaje de su trayectoria total. Lo anterior conlleva a que el segundo punto amarillo, de la trayectoria del correcaminos, está por encima del segundo punto rosa que corresponde a la trayectoria del automóvil. Siguiendo este razonamiento es que se obtienen las curvas restantes $\gamma_1(t)$ y $\gamma_2(t)$ de la Figura 3.7.

Por tanto, en términos coloquiales, la función de deformación temporal estirará o contraerá a las trayectorias en estudio. De esta manera tendrán el mismo punto de inicio y de fin al tiempo $t = 0$ y $t = 1$, respectivamente. En vista de los hechos comentados, la función $\gamma(t)$ se puede entender como un recurso que permite comparar trayectorias, vía una deformación temporal de éstas. A éste tipo de análisis se le conoce como *análisis elástico de la forma de una trayectoria*. Dicha temática puede ser consultada con mayor profundidad en Joshi et al. [2016] y Tucker et al. [2013].

Con base en lo que se ha comentado hasta este punto, es primordial notar que el dato que realmente se observa es $\alpha(\gamma(t))$. Esta afirmación es consecuencia de la siguiente situación: si se conoce la huella de la trayectoria $\alpha(t)$ y se recorre con distintas tasas de evolución $\gamma(t)$, lo que se obtendrá es un conjunto diferente de observaciones por cada $\gamma(t)$ empleada. Esto da lugar a la noción de curva, descrita en la Deficinición 2.2.2 del Capítulo 2 de la presente tesis. En la Figura 3.8 se puede apreciar de manera más clara el hecho comentado[4]. Por otra parte y como fue mencionado con anterioridad, para conocer la huella de una trayectoria es necesario realizar de manera previa un proceso de interpolación o ajuste sobre los puntos que conforman a las trayectorias. Este asunto puede ser considerado como un punto adverso del presente enfoque, por el reto técnico y computacional que esta tarea implica. Finalmente a la composición $\alpha(\gamma(t))$ se le denotará como $(\alpha \circ \gamma)(t)$.

[4]Las imágenes usadas en la Figura 3.8 se pueden encontrar en Srivastava et al. [2011a].

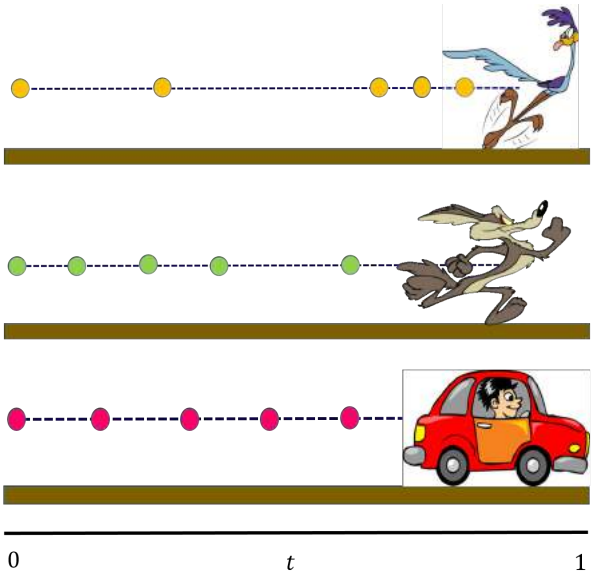


Figura 3.6: Trayectoria del correcaminos, el coyote y el conductor.

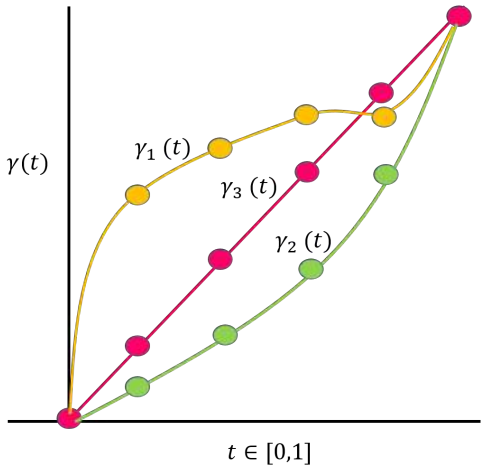


Figura 3.7: Función $\gamma(t)$ para el correcaminos, el coyote y el conductor.

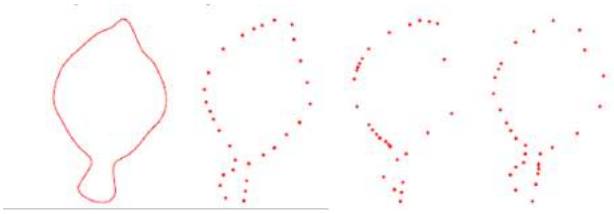


Figura 3.8: Forma de una hoja—primera figura del lado izquierdo—recorrida con tres diferentes tasas de evolución $\gamma(t)$.

Una vez esbozados y aclarados puntos que serán vitales en el análisis estadístico que será desarrollado posteriormente, se explicarán brevemente los pasos previos.

Primero, se usará el TSRVF (ver Sección 2.2.5) para representar las trayectorias en un espacio conocido. Posteriormente se empleará la función de deformación temporal, que ayudará a hacer comparables las trayectorias, es decir estandarizarlas. Finalmente, se encontrará una medida con la cual se puedan establecer distancias entre trayectorias y con ello sea posible calcular la trayectoria media y la varianza asociada a un conjunto de trayectorias. Por consiguiente se tienen las siguientes tres tareas a desarrollar:

- Representar las trayectorias en un "buen" espacio.
- Alinear las trayectorias.
- Encontrar una medida para calcular distancias entre trayectorias alineadas.

Con el desarrollo de éstos tres puntos, se plantea que el lector alcance a percibir de manera natural la utilidad de algunas herramientas que se desarrollaron en el Capítulo 2. También se pretende exhibir el reto teórico que hay al extender nociones como la de una medida o la de clases de equivalencia, considerando ciertas transformaciones.

Espacio para representar trayectorias

El principal motivo para buscar un nuevo espacio donde se puedan representar las trayectorias, es para medir las diferencias que puedan existir entre ellas considerando una reparametrización del tiempo $\gamma(t)$. El argumento anterior es consecuencia de que la distancia intrínseca de una variedad M, no es invariante a reparametrizaciones temporales. Esto significa que, $d(\alpha_1(t), \alpha_2(t)) \neq d((\alpha_1 \circ \gamma)(t), (\alpha_2 \circ \gamma)(t))$, tal que $d(\cdot, \cdot)$ representa la distancia intrínseca de la variedad M; este hecho será aclarado con mayor profundidad más adelante. Por consiguiente, es de vital importancia encontar dicho espacio, pues como se mencionó es de interés trabajar con las trayectorias estandarizadas. En el fondo lo que se desea es poder representar las trayectorias de interés en un espacio lineal, facilitando de esta manera su tratamiento matemático y estadístico.

En el Capítulo 2 se abordó una noción geométrica llamada TSRVF, que se caracterizaba por ser un tipo de transporte paralelo. Dicho concepto permitía representar trayectorias de una variedad M en un espacio tangente T_cM, tal que éste último es un espacio vectorial. Por tanto el lugar donde se representarán las trayectorias es T_cS^2, el plano tangente a un punto c en la esfera. Es relevante recordar que este transporte paralelo requiere el campo velocidad de la trayectoria $\alpha(t)$, y por tal motivo es que al principio de esta sección se comentó que se trabajaría con trayectorias suaves. Así mismo, esta noción geométrica necesita un punto de referencia $c \in S^2$. Tal punto es el lugar donde se definirá el plano tangente y por ende el lugar donde se hará el transporte paralelo de las trayectorias de interés.

El TSRVF será la herramienta estrella de este capítulo, pues permitirá representar las trayectorias de un espacio no lineal—como es el caso de la esfera—a uno que sí lo es. De esa manera es que se ganará intuición del análisis estadístico que se realizará.

Se agrega al hecho de que se contarán con varias herramientas tanto estadísticas como probabilísticas. Un ejemplo de ello son las métricas que se conocen para $\mathbb{R}^2$, como es el caso de la norma L^2. Dicha medida desempeñará un rol esencial en el desarrollo de la métrica para comparar trayectorias.

Una vez que las trayectorias fueron transportadas, sigue definir una modificación temporal de ellas, que es lo que se explicará a continuación.

Registro temporal y alineación de trayectorias

El registro temporal, como ya fue mencionado, es una transformación del tiempo que involucra el uso de la función de deformación temporal. Esta transformación establece una correspondencia uno a uno entre múltiples trayectorias. Esto significa que las estandariza de forma que todas tengan el mismo punto de inicio y el mismo punto de fin. Un caso de registro y alineación de trayectorias fue ejemplificado en las Figuras 3.6 y 3.7. El proceso de alineación entre trayectorias ofrece la ventaja de evitar un incremento en la varianza, causado por la disparidad de las observaciones. Por tanto, dicho parámetro puede ser usado en un modelo de probabilidad, como será visto en la Sección 3.4.3.

A pesar de la ganancia estadística que se obtiene al procesar un registro de trayectorias, tal procedimiento complica el cómputo. La principal razón es que para encontrar $\gamma(t)$ se requiere el uso de programación dinámica. Dicha materia representa un reto, en principio por el bagaje técnico que requiere y en segundo por el proceso de optimización implícito en esta metodología. Aunado a lo anterior, el registro de trayectorias complica encontrar una métrica entre trayectorias debido a las diferentes tasas de evolución $\gamma(t)$ con las que puede ser modelada una trayectoria.

Para tener una idea más precisa de lo que es el registro temporal, se puede consultar Ramsay [2006] y Kneip and Ramsay [2008]. El Capítulo 7 de la primera referencia trata ampliamente el tema de registro de trayectorias. La segunda referencia es un artículo que explica, en términos sencillos, en qué consiste el registro de trayectorias identificando los retos que existen. Por tal motivo, aborda algunos procedimientos de registro los cuales ejemplifica con datos reales. Este artículo es un excelente texto introductorio para aquellos que desean adentrarse en la temática del registro temporal.

Distancia entre trayectorias

Una vez que las trayectorias fueron alineadas, sigue especificar una métrica que sea invariante a reparametrizaciones temporales, es decir,

$$d(\alpha_1(t), \alpha_2(t)) = d((\alpha_1 \circ \gamma)(t), (\alpha_2 \circ \gamma)(t)).$$

¿Por qué quiero una métrica que cumpla esa característica? La respuesta, en esencia, obedece al hecho que será ejemplificado a continuación. Supóngase que se tienen dos carreteras, las cuales son recorridas varias veces al día por vehículos que llevan

distintas velocidades. La forma que tienen estas carreteras no cambia, independientemente de la velocidad con la que han sido recorridas por cada vehículo. Algo semejante ocurre con las trayectorias $\alpha_1(t)$ y $\alpha_2(t)$; es decir, la huella de una trayectoria no cambia sólo porque fue recorrida de una manera "x" o "y". Por ende, la distancia entre trayectorias no debe de cambiar, independientemente del cómo fueron recorridas. Con esta noción de invarianza es que se formaliza la idea de que uno de los principales objetos de estudio, en el análisis estadístico de trayectorias, es la huella que de manera *per se* trazan éstas.

A continuación se muestran algunas propuestas que se consideraron en el artículo base para ser la métrica principal en el estudio estadístico. Sin embargo, por motivos que serán expuestos más adelante, fueron desechadas. La finalidad de mostrar estas métricas es que el lector gane intuición de las características que debe tener la métrica de interés. Las propuestas fueron las siguientes:

1. **Extención de la distancia riemannina**.
   La idea de esta métrica es comparar cualesquiera dos trayectorias, $\alpha_1(t)$ y $\alpha_2(t)$, directamente sobre la variedad M. Se define como

   $$d_x(\alpha_1(t), \alpha_2(t)) = \int_0^1 d_m\left(\alpha_1(t), \alpha_2(t)\right),$$

   donde d_m es la distancia intrínseca de la variedad M.

   La ventaja que presenta esta métrica es que no exige una transformación previa de las trayectorias para hacer comparaciones entre ellas. Sin embargo, no es invariante a transformaciones temporales; es decir, $d_x(\alpha_1(t), \alpha_2(t)) \neq d_x((\alpha_1 \circ \gamma)(t), (\alpha_2 \circ \gamma)(t))$. Por este motivo queda descartada.

2. **Distancia riemanniana más un término de corrección**.
   Se define como

   $$\min_{\gamma} \left(\int_0^1 d_m\left(\alpha_1(t), \alpha_2(\gamma(t))\right)^2 dt + \lambda \mathcal{R}(\gamma) \right),$$

   donde $d_m(\cdot, \cdot)$ representa nuevamente a la distancia intrínseca de la variedad M, γ es la función de deformación temporal, $\mathcal{R}(\gamma)$ es un término de regularización y $\lambda > 0$ es una constante asociada a $\mathcal{R}(\gamma)$.

   La intuición que justifica esta métrica es encontrar una deformación temporal sobre la variedad M, de manera que dicha deformación sea controlada con el término $\mathcal{R}(\gamma)$. Por consiguiente, $\mathcal{R}(\gamma)$ será interpretado como un índice del grado de elasticidad de una trayectoria. En otras palabras, $\mathcal{R}(\gamma)$ indicará qué tanto se puede deformar una trayectoria con respecto a otra.

   La desventaja que presenta esta propuesta es que no es una distancia propia, por ende no es una métrica. Aunado a lo anterior, la distancia entre $\alpha_1(t)$ y $\alpha_2(t)$, considerando la reparametrización temporal, no es igual a la distancia

entre $\alpha_2(t)$ y $\alpha_1(t)$. Los detalles de la prueba pueden ser revisados en Christensen and Johnson [2001].

3. **Log-Mapeo**.
Definido y estudiado brevemente en el Capítulo 2, la idea de esta propuesta es representar una trayectoria $\alpha(t)$ en el espacio T_cM, vía el mapeo logarítmico. En el caso de la presente tesis, tal espacio es T_cS^2. El problema que posee el log-mapeo, bajo esta concepción, es que arroja resultados inconsistentes. Un ejemplo de ello es la proyección estereográfica, pues tres puntos cercanos en el polo norte quedarían distantes al proyectarlos en el plano tangente, como muestra la Figura[5] 3.9.

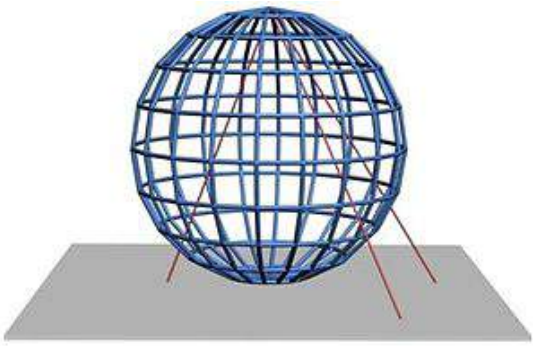


Figura 3.9: Proyección estereográfica de tres puntos.

Como ya se dijo, a pesar de que las propuestas anteriores no fueron fructíferas, ayudaron a concebir atributos deseables en la métrica de interés. El primer atributo es, que la métrica en consideración indique cuán suave o rugosa fue la transformación temporal de la trayectoria en cuestión . El segundo atributo hace referencia a que el lugar donde quede definida tal métrica, tenga la estructura de un espacio vectorial.

Previamente, en el apartado titulado "Espacio para representar trayectorias", se comentó que T_cS^2 sería el sitio donde se estudiarían las trayectorias de interés. Este plano tangente se caracteriza por ser un espacio vectorial; en consecuencia es posible trabajar con métricas conocidas, de manera específica una modificación de la norma L^2 que incorporará al TSRVF. Esta "nueva métrica" se llamará $d_h(\cdot,\cdot)$ y tiene la siguiente definición.

Definición 3.3.1 *Sean $\alpha_1(t)$ y $\alpha_2(t)$ dos trayectorias suaves sobre M y sean $h_{\alpha_1}(t)$ y $h_{\alpha_2}(t)$ sus correspondientes TSRVFs. La distancia entre las trayectorias, se define como*

$$d_h(h_{\alpha_1}(t), h_{\alpha_2}(t)) = \left(\int_0^1 \mid h_{\alpha_1}(t) - h_{\alpha_2}(t) \mid^2 dt \right)^{1/2}. \tag{3.1}$$

La ventaja que presenta la métrica $d_h(\cdot,\cdot)$ es que contempla la transformación que sufrió la trayectoria $\alpha(t)$, para poder trabajarla en el espacio vectorial T_cM, o bien

[5]Fuente *https://es.wikipedia.org/wiki/Proyección_estereográfica.*

en el caso particular de la presente tesis en T_cS^2. Además, tal métrica satisface ser invariante a trasformaciones temporales. Este último hecho es el que será formalizado a continuación.

Teorema 3.3.1 *Para cualquier $\alpha_1(t), \alpha_2(t) \in M$ y $\gamma \in \Gamma$, la distancia $d_h(\cdot,\cdot)$ satisface*

$$d_h(h_{\alpha_1\circ\gamma}(t), h_{\alpha_2\circ\gamma}(t)) = d_h(h_{\alpha_1}(t), h_{\alpha_2}(t)). \tag{3.2}$$

La implicación en términos geométricos de este teorema es que la distancia entre trayectorias, considerando el TSRVF, es igual sin importar la deformación temporal que sea utilizada.

Para la prueba del Teorema 3.3.1 es necesario notar que

$$h_{\alpha\circ\gamma}(t) = h_\alpha(\gamma(t))\sqrt{\dot\gamma(t)}. \tag{3.3}$$

Esta igualdad se sigue de sustituir $\beta(t) = (\alpha\circ\gamma)(t)$ en la ecuación (2.7), que es la definición del TSRVF, dada en el Capítulo 2. Por lo tanto

$$\begin{aligned} h_{\alpha\circ\gamma}(t) &= h_\beta(t) \\ &= \frac{\dot\beta(t)_{\beta(t)\longrightarrow c}}{\sqrt{|\ \dot\beta(t)\ |}} \\ &= \frac{(\dot\alpha(\gamma(t))\dot\gamma(t))_{\alpha(\gamma(t))\longrightarrow c}}{\sqrt{|\ \dot\alpha(\gamma(t))\dot\gamma(t)\ |}}. \end{aligned}$$

Por otra parte, ya que $\gamma : [0,1] \longrightarrow [0,1]$ se sigue

$$\begin{aligned} h_{\alpha\circ\gamma}(t) &= \frac{\dot\alpha(\gamma(t))_{\alpha(\gamma(t))\longrightarrow c}\dot\gamma(t)}{\sqrt{|\ \dot\alpha(\gamma(t))\dot\gamma(t)\ |}} \\ &= \frac{\dot\alpha(\gamma(t))_{\alpha(\gamma(t))\longrightarrow c}\sqrt{\dot\gamma(t)}}{\sqrt{|\ \dot\alpha(\gamma(t))\ |}} \\ &= h_\alpha(\gamma(t))\sqrt{\dot\gamma(t)}. \end{aligned}$$

Al TSRVF de $(\alpha\circ\gamma)(t)$ se le denotará como $(h_\alpha,\gamma)(t)$, por lo cual $(h_\alpha,\gamma)(t) = (h\circ\gamma)\sqrt{\dot\gamma}$. De esa misma forma, se resalta que la ecuación (3.3) indica que una vez realizado el TSRVF, la estandarización de la trayectoria tendrá asociado un término de penalización, el cual indicará que tan rugosa o suave fue la transformación temporal de $h_\alpha(t)$. Por tal razón, se podrá saber en qué medida se deformó el transporte paralelo de la trayectoria en cuestión. Aclarado el punto anterior sigue la prueba del teorema ya citado.

Demostración:

Tomando como punto de partida las igualdades (3.1) y (3.3), la demostración del teorema se reduce a realizar algunas sustituciones algebraicas. Ello se muestra a continuación:

$$
\begin{aligned}
d_h(h_{\alpha_1\circ\gamma}, h_{\alpha_2\circ\gamma}) &= \left(\int_0^1 \mid h_{\alpha_1\circ\gamma}(t) - h_{\alpha_2\circ\gamma}(t) \mid^2 dt\right)^{1/2} \\
&= \left(\int_0^1 \left|h_{\alpha_1}(\gamma(t))\sqrt{\dot\gamma(t)} - h_{\alpha_2}(\gamma(t))\sqrt{\dot\gamma(t)}\right|^2 dt\right)^{1/2} \\
&= \left(\int_0^1 \left|\Big(h_{\alpha_1}(\gamma(t)) - h_{\alpha_2}(\gamma(t))\Big)\sqrt{\dot\gamma(t)}\right|^2 dt\right)^{1/2} \\
&= \left(\int_0^1 |h_{\alpha_1}(\gamma(t)) - h_{\alpha_2}(\gamma(t))|^2\, \dot\gamma(t) dt\right)^{1/2}.
\end{aligned}
$$

Considerando el cambio de variable $s = \gamma(t)$ se sigue $ds = (d\gamma(t)/dt)\, dt = \dot\gamma(t)dt$. Por ende

$$
\begin{aligned}
d_h(h_{\alpha_1\circ\gamma}, h_{\alpha_2\circ\gamma}) &= \left(\int_0^1 |h_{\alpha_1}(s) - h_{\alpha_2}(s)|^2\, ds\right)^{1/2} \\
&= d_h(h_{\alpha_1}, h_{\alpha_2}),
\end{aligned}
$$

con lo cual queda concluída la prueba. ∎

El siguiente paso es trabajar las trayectorias considerando todas las tasas de evolución temporal con las que pueden ser modeladas. Por tanto, a continuación se introducirá la noción de clases de equivalencia entre trayectorias. Dos trayectorias $\alpha_1(t)$ y $\alpha_2(t)$ se dirá que son equivalentes, $\alpha_1 \backsim \alpha_2$, si

1. $\alpha_1(0) = \alpha_2(0)$.
2. Existe una sucesión $\{\gamma_k\} \in \Gamma$ tal que $\lim_{k\longrightarrow\infty} h_{\alpha_1\circ\gamma_k} = h_{\alpha_2}$ bajo la métrica L^2.

Lo anterior significa que dos trayectorias son equivalentes si tienen el mismo punto de inicio y via transformaciones temporales se puede llegar de $h_{\alpha_1}(t)$ a $h_{\alpha_2}(t)$. La presentación de las clases de equivalencia entre trayectorias consituye otra de las aportaciones realizadas por el artículo Su et al. [2014a].

A continuación se definirá formalmente a las clases de equivalencia con las que se trabajará.

Definición 3.3.2 *Sea $h_\alpha \in \mathcal{H}$ el TSRVF de $\alpha(t) \in \mathcal{M}$, tal que $h_\alpha \in \mathcal{H}$ donde $\mathcal{H}$ es el conjunto de transportes paralelos de trayectorias $\alpha(t)$, se tiene que la clase de equivalencia de h_α está dada por*

$$[h_\alpha] = \{(h_\alpha, \gamma) \mid \gamma \in \Gamma\}.$$

La ventaja que ofrece la Definición 3.3.2 es que trabaja con la noción de curva tomando en cuenta el TSRVF. Por lo tanto se compararán trayectorias vía la curva que les subyace y no propiamente sobre los puntos observados. De acuerdo con lo que se ha cubierto hasta este punto, sigue definir la distancia más corta $d_h(\cdot,\cdot)$, que logre cuantificar las diferencias que existen entre estas clases de equivalencia.

Definición 3.3.3 *La distancia $d_s(\cdot,\cdot)$ sobre $\mathcal{H}/\sim$ es la distancia más corta $d_h(\cdot,\cdot)$ entre las clases de quivalencia en $\mathcal{H}$, está dada por*

$$\begin{aligned} d_s([h_{\alpha_1}],[h_{\alpha_2}]) &= \inf_{\gamma_1,\gamma_2\in\Gamma} d_h((h_{\alpha_1},\gamma_1),(h_{\alpha_2},\gamma_2)) \\ &= \inf_{\gamma\in\Gamma}\left(\int_0^1 \mid h_{\alpha_1}(\gamma_1(t))\sqrt{\dot{\gamma}_1(t)} - h_{\alpha_2}(\gamma_2(t))\sqrt{\dot{\gamma}_2(t)} \mid^2 dt\right)^{1/2}. \end{aligned}$$

Esta distancia cumple con ser una distancia propia ya que es simétrica, positiva definida y satisface la desigualdad del triángulo. La prueba se hace desde principios básicos y puede ser consultada en la página 11 de Su et al. [2014a]. Cabe resaltar que la definición de $d_s(\cdot,\cdot)$, así como la prueba de que es una distancia propia constituye otra de las aportaciones del artículo base.

La métrica $d_s(\cdot,\cdot)$ también cumple la propiedad de ser invariante a deformaciones temporales. Más aún, es invariante a deformaciones temporales simultáneas. Es decir

$$d_s([h_{\alpha_1\circ\gamma_1}],[h_{\alpha_2\circ\gamma_2}]) = d_s([h_{\alpha_1}],[h_{\alpha_2}]).$$

La prueba puede ser consultada en Su [2013].

Por tanto, se ha encontrado una métrica entre trayectorias que es invariante a diferentes tasas de evolución $\gamma(t)$. Como se ha anunciado previamente, el principal uso de esta métrica será para encontrar una trayectoria media de un conjunto $\{\alpha_i(t)\}_{i=1}^n$ de trayectorias, así como para cuantificar la varianza que se le asocia.

Con este punto concluído se dan por finalizados los pasos previos al desarrollo estadístico, el cual será abordado a continuación. Es importante notar que en general el principal reto de esta sección fue definir la métrica $d_s(\cdot,\cdot)$, en gran parte por los diferentes requerimientos que debía cumplir ésta y por ende por toda la herramienta que se necesitó desarrollar y probar.

Algunas recomendaciones bibliográficas, para enriquecer la presente sección, son:

1. Tucker et al. [2013]. *Generative models for functional data using phase and amplitude separation.*
   Este texto motiva la necesidad de capturar la estructura o geometría que puede ostentar una curva. Como consecuencia de ello es que implementa un estudio estadístico de curvas, tal que la principal herramienta es una técnica llamada *análisis eslástico de la forma de una curva.* Algunas ideas de dicho enfoque son extendidas y empleadas por el artículo que fue tomado como base. También, esta referencia aborda algunos algoritmos parecidos a los que se expondrán a continuación y los ejemplifica con el uso de datos reales.

2. Srivastava et al. [2011b]. *Registration of functional data using Fisher-Rao metric.*
   Introduce nociones geométricas en el análisis de curvas, bajo el contexto de

datos funcionales. Su principal aportación es proponer el uso de funciones que ayuden a comparar trayectorias, de forma que la métrica de Fisher-Rao pueda ser usada bajo cierta transformación. Este artículo es uno de los precursores en el análisis estadístico de trayectorias sobre variedades, por lo cual puede considerarse como una lectura previa al artículo base. Cabe mencionar que la idea de trabajar con la norma L^2 modificada surge de este trabajo. Para aquellos lectores que deseen conocer y ahondar en la temática que refiere a la métrica de Fisher-Rao se recomienda leer Maybank [2008].

3. Srivastava et al. [2007]. *Riemannian analysis of probability density functions with applications in vision.*
   Es uno de los primeros artículos en el área de ciencias de la computación en comentar que hay un reto y una necesidad en desarrollar herramientas para hacer inferencia estadística en espacios no lineales. El principal objetivo de este texto es encontrar una métrica que habilite un cómputo eficiente de herramientas estadísticas, de manera que la metodología desarrollada pueda ser aplicada en el análisis de visión computacional.

3.4. Análisis estadístico de trayectorias

Una vez que se establecieron todas las herramientas matemáticas necesarias, sigue hacer el análisis estadístico de las trayectorias. Por lo tanto, en esta sección se expondrán los algoritmos para encontrar la trayectoria media de un conjunto de trayectorias y la varianza asociada a éste. Una vez calculados éstos parámetros, se abordará un modelo de probabilidad para una trayectoria $\alpha(t)$.

3.4.1. Trayectoria media.

El algoritmo con el cual se obtendrá dicha trayectoria estará basado principalmente en la siguiente función objetivo:

$$h_\mu = \underset{[h_\alpha]\in\mathcal{H}/\sim}{\operatorname{argmín}} \sum_{i=1}^{n} d_s([h_\alpha],[h_{\alpha_i}])^2. \tag{3.4}$$

La función 3.4 es análoga a la función (1.1), que es la media de Karcher para datos puntuales que se encuentran en una variedad M. Las piezas que cambian, en esta nueva función, son la distancia y los elementos sobre los cuales se realizará el proceso de minimización. Por tanto, la intuición de esta media sigue siendo encontrar aquel elemento en $\mathcal{H}$, bajo la relación de equivalencia $\sim$, que minimice la distancia entre los elementos $[h_{\alpha_i}]$ que pertenecen a dicho espacio. Es valioso percatarse que para definir h_μ—el TSRVF de la trayectoria media—es que se requirió determinar la distancia $d_s(\cdot,\cdot)$.

El siguiente algoritmo explica el procedimiento para encontrar la trayectoria media de un conjunto de trayectorias.

Algoritmo 3.4.1.1. Trayectoria media de un conjunto $\{\alpha_i(t)\}_{i=1}^n$

Datos de entrada:

- El conjunto de trayectorias observadas $\{\alpha_i(t)\}_{i=1}^n$.
- Un punto de referencia c.

Se recuerda que las trayectorias $\{\alpha_i(t)\}_{i=1}^n$ deben de ser suaves y no pasar por el punto antípodo a c.

Datos de salida:

- Trayectoria media $\mu(t)$.
- El conjunto de trayectorias $\{\alpha_i(t)\}_{i=1}^n$ alineadas.

Pasos:

1. Encontrar la media de Fréchet de los puntos $\{\alpha_i(0)\}_{i=1}^n$. A este punto se le denotará como $\mu(0)$.
   Recuérdese que dicha media fue definida en el Capítulo 1, mediante la ecuación (1.1). Por otro lado, es fundamental aclarar que únicamente para este paso será usada la métrica de la variedad M con la que se esté trabajando. En el caso de la esfera unitaria se usará la distancia definida en (2.1).

2. Del conjunto de trayectorias $\{\alpha_i(t)\}_{i=1}^n$ seleccionar una trayectoria como $\mu(t)$. Posteriormente hallar $h_\mu(t)$, es decir el TSRVF de $\mu(t)$.
   En este paso es que se requiere el punto de referencia c, pues es el lugar donde se hará el TSRVF de las trayectorias $\{\alpha_i(t)\}_{i=1}^n$ es T_cS^2.

3. Obtener $h_{\alpha_i}(t)$ para $i = 1, \ldots, n$.

4. Alinear cada $h_{\alpha_i}(t)$ con base en h_μ.
   Para el desarrollo de este paso se requerirá encontrar la función de deformación temporal, $\gamma_i^*(t)$, que satisfaga la siguiente igualdad

$$\gamma_i^* = \underset{\gamma_i \in \Gamma}{\text{argmín}} \left(\int_0^1 \mid h_\mu(t) - h_{\alpha_i}(\gamma_i(t)) \sqrt{\dot{\gamma}_i(t)} \mid^2 dt \right)^{\frac{1}{2}}. \tag{3.5}$$

   La igualdad anterior es similar a la ecuación **??** tomando $\gamma_1(t) = \mathrm{Id}(t)$, donde $\mathrm{Id}(t)$ es la función identidad. En la ecuación 3.5 se presenta que la deformación temporal se hará tomando como base el TSRVF de aquella trayectoria que se tomó como media.

5. Obtener $\tilde{\alpha}_i = \alpha_i \circ \gamma_i^*$, tal que $i = 1, \ldots, n$.
   En este caso $\{\tilde{\alpha}_i(t)\}_{i=1}^n$, representará el conjunto de trayectorias alineadas. También se aclara que en el caso de la trayectoria α_i que fue elegida como la trayectoria media se tiene que $\tilde{\alpha}_i = \alpha_i(\mathrm{Id}(t))$; es decir $\gamma_i^* = \mathrm{Id}(t)$.

6. Hallar $h_{\tilde{\alpha}_i}(t)$, donde $i = 1, \ldots, n$.

7. Actualizar $h_\mu(t)$, como una curva en T_cS^2, de acuerdo con

$$h_\mu(t) = \frac{1}{n}\sum_{i=1}^{n} h_{\tilde{\alpha}_i}(t).$$

   Nótese que en este paso es dónde se aprovecha al máximo que T_cS^2 es un espacio vectorial, ya que la media $h_\mu(t)$ se calcula igual que una media muestral en $\mathbb{R}^n$.

8. Regresar la trayectoria media a la variedad S^2, vía la ecuación diferencial

$$\frac{d\mu(t)}{dt} =\mid h_\mu(t) \mid h_\mu(t)_{c\longrightarrow\mu(t)},$$

   con condición inicial $\mu(0)$.
   Es de apreciar que esta ecuación es quivalente a (2.8), sustituyendo el campo vectorial $V(t)$ por $h_\mu(t)$. En este caso $c \longrightarrow \mu(t)$ representa la curva geodésica que va de c a $\mu(t)$ para $t \in [0, 1]$.

9. Encontrar

$$E = \sum_{i=1}^{n} d_s([h_\mu], [h_{\alpha_i}])^2 = \sum_{i=1}^{n} d_h(h_\mu, h_{\tilde{\alpha}_i})^2$$

   y revisar su convergencia. Si ésta no existe regresar al paso tres del presente algoritmo.

Es relevante comentar que la función (3.4) decrece iterativamente hacia cero. Por tanto ésta siempre convergerá, con lo cual se puede asegurar la existencia de una trayectoria media.

El Algoritmo 3.4.1.1 es una de las principales aportaciones del artículo Su et al. [2014a], pues consigue definir una trayectoria media representativa sobre variedades. Esto significa que la forma de la trayectoria media se encuentra acorde con la forma de las trayectorias individuales. Cabe mencionar que dicho algoritmo es una generalización del que fue propuesto por Le and Kume [2000], el cual logró obtener la media de triángulos en el espacio de formas. Dicho texto es considerado el artículo precursor en abordar la media de una forma, así como en ofrecer un modelo de probabilidad a los vértices de una forma.

Las Figuras 3.10–3.15 ofrecen un esbozo gráfico de los pasos expuestos con anterioridad.

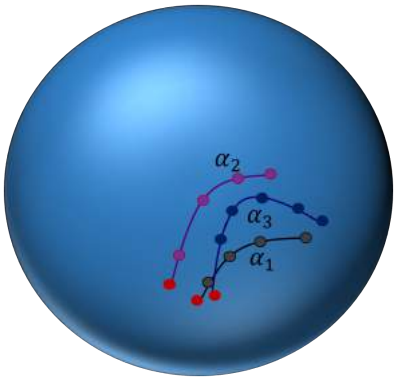


Figura 3.10: Conjunto de trayectorias con sus puntos iniciales y $\mu(0)$.

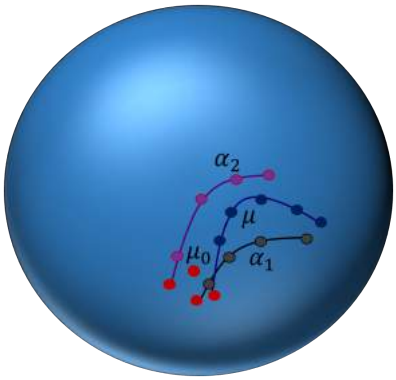


Figura 3.11: Selección de una trayectoria como la trayectoria media.

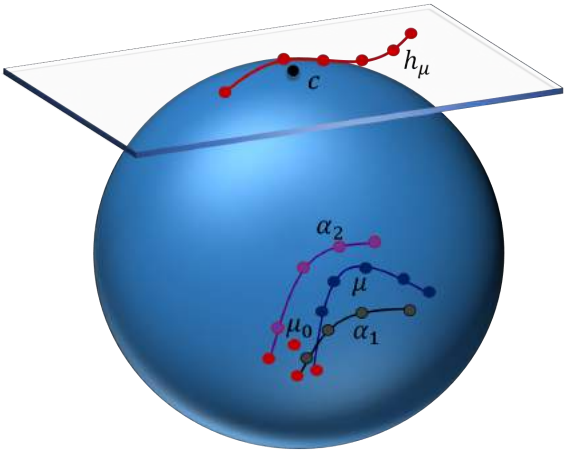


Figura 3.12: TSRVF de la trayectoria tomada como media.

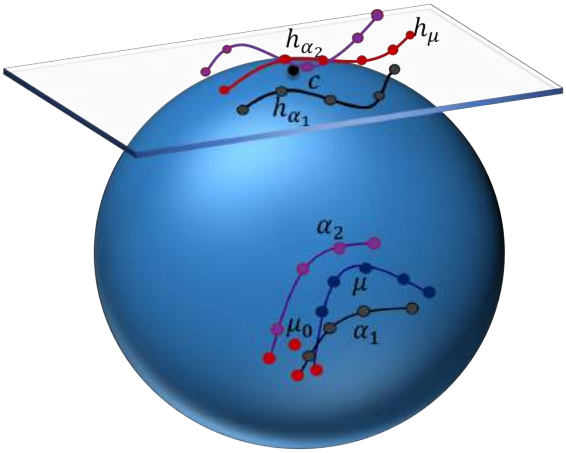


Figura 3.13: TSRVF de las demás trayectorias.

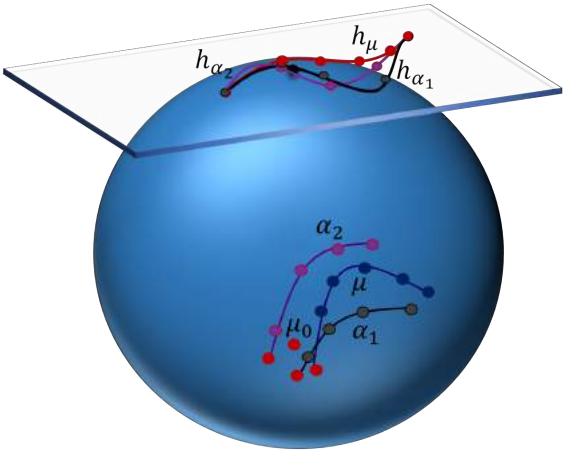


Figura 3.14: Alineación de h_{α_1} y h_{α_2} con base en h_μ.

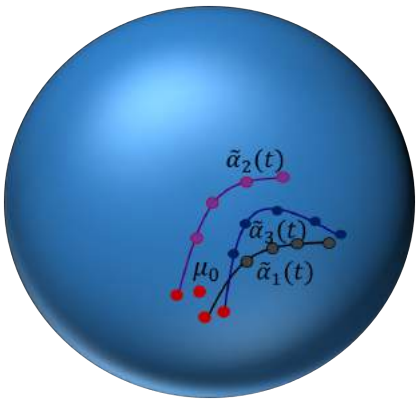


Figura 3.15: Trayectorias alineadas.

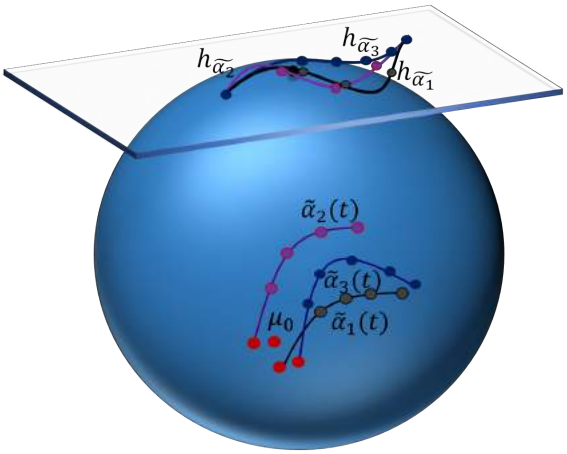


Figura 3.16: TSRVF trayectorias alineadas.

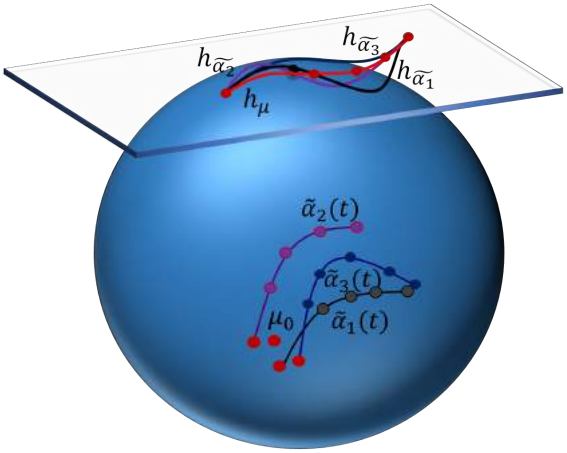


Figura 3.17: Actualización de h_μ.

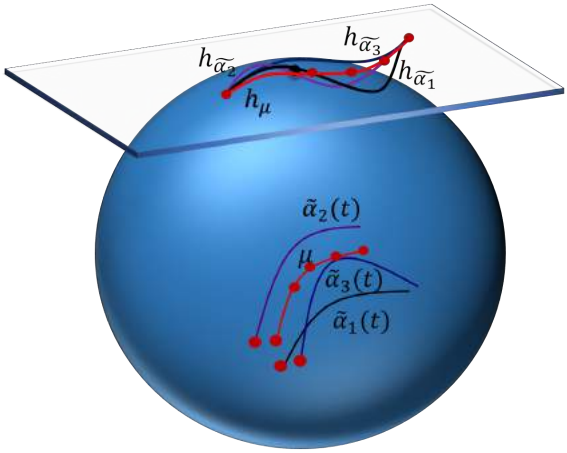


Figura 3.18: Trayectoria media sobre la esfera.

La alineación en la esfera, presentada en la Figura 3.15, se refiere a recorrido entre trayectorias. Es decir, dónde se pueden posicionar las observaciones puntuales en cada trayectoria y así encontrar medidas estadísticas representativas. La Figura 3.19, tomada del artículo base, muestra dos trayectorias previo y posterior al proceso de alineación. La Figura 3.20 muestra las trayectorias utilizadas en el esbozo previo y posterior al proceso de alineación.

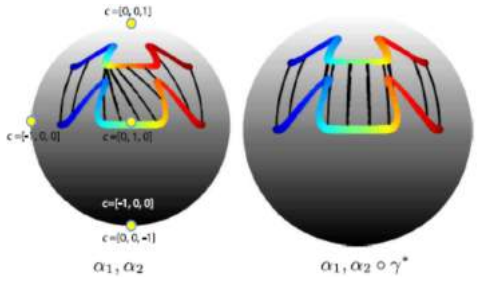


Figura 3.19: En la esfera de la izquierda dos trayectorias α_1 y α_2 sin alinear. En la esfera de la derecha la trayectoria α_2 alineada con base en la trayectoria α_1.

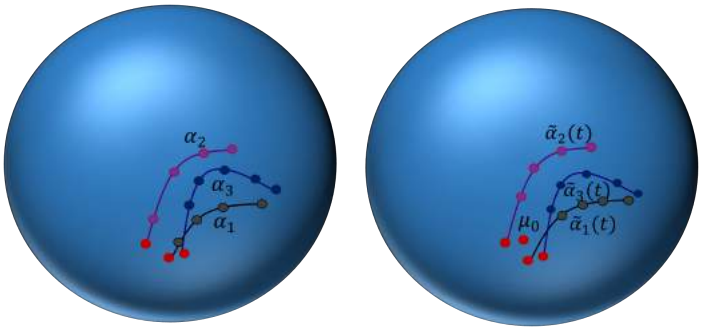


Figura 3.20: La esfera de la izquierda muestra las trayectorias sin alinear. La esfera de la derecha muestra las trayectorias alineadas con base en α_3.

3.4.2. Varianza de un conjunto de trayectorias.

La varianza de un conjunto de trayectorias $\{\alpha_i(t)\}_{i=1}^n$, a diferencia de la trayectoria media $\mu(t)$, es un conjunto de cantidades que indican qué tan semejantes son las trayectorias entre sí. Para su cálculo será necesario hacer una partición del tiempo. Es decir, considerar $\{t_j\}_{j=1}^m$ tal que $t_1 = 0, \ldots, t_m = 1$. De esa forma es que se trabajará con las trayectorias discretizadas, como se muestra a continuación.

Algoritmo 3.4.2.1 Varianza de un conjunto de trayectorias $\{\alpha_i(t)\}_{i=1}^n$.

Datos de entrada:

- Trayectoria media discretizada, $\mu(t_1), \mu(t_2), \ldots, \mu(t_m)$.
- Trayectorias alineadas discretizadas, $\{\tilde{\alpha}_i(t_j)\}_{i=1}^n$ tal que $j = 1, \ldots, m$.

Datos de salida:

- Matriz de varianzas y covarianzas estimada para cada tiempo t_j, $j = 1, \ldots, m$.

Pasos:

1. Encontrar el mapeo logarítmico de $\mu(t_j)$ a $\tilde{\alpha_i}(t_j)$. Al vector resultante se le denotará como $v_i(t_j)$ y se le denominará *shooting vector*.

   En este paso es importante notar los siguientes detalles:

   - Para hallar el mapeo logarítmico se establecerá como punto de referencia $\mu(t_j)$.
   - El lugar donde se cuantifica la varianza es $T_{\mu(t_j)}S^2$, lo cual se traduce en que $v_i(t_j) \in T_{\mu(t_j)}S^2$.
   - Para cada trayectoria $\{\tilde{\alpha}_i(t_j)\}_{i=1}^n$ existe un *shooting vector* $v_i(t_j)$.

   Un *shooting vector* podrá entenderse como un recurso puntual, para determinar la dirección principal que hay de $\mu(t_j)$ a cada una de las trayectorias $\alpha(t_j)$.

2. Encontrar la matriz de covarianzas muestral $\hat{K}(t)$, asociada a los *shooting vectors*.

$$\hat{K}(t_j) = \frac{1}{n-1}\sum_{i=1}^{n} v_i(t_j)v_i(t_j)^T. \tag{3.6}$$

   A (3.6) se le conoce como la covarianza muestral de Karcher al tiempo t_j.

3. Calcular la traza de ecuación (3.6).

$$\hat{\rho}(t_j) = \mathrm{tr}(\hat{K}(t_j)).$$

   En este caso $\hat{\rho}(t_j)$ se interpreta como una medida del nivel de alineación de las trayectorias $\{\tilde{\alpha}_i(t)\}_{i=1}^n$ en el tiempo t_j.

En las Figuras 3.21–3.31 se ejemplifica el algoritmo anterior, con tres trayectorias.

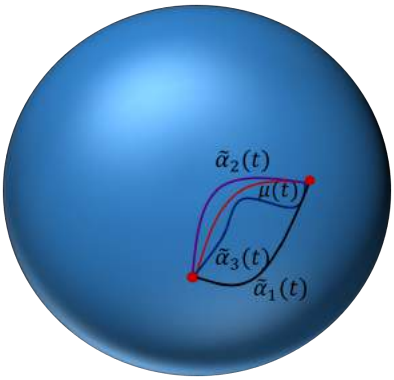


Figura 3.21: Trayectoria media y conjunto de trayectorias alineadas.

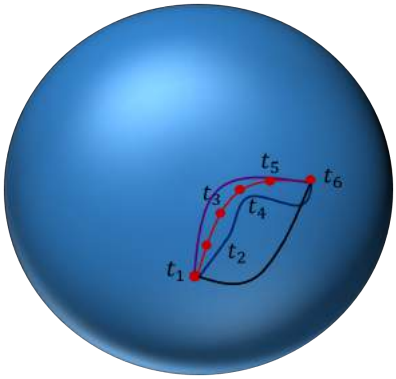


Figura 3.22: Discretización del tiempo.

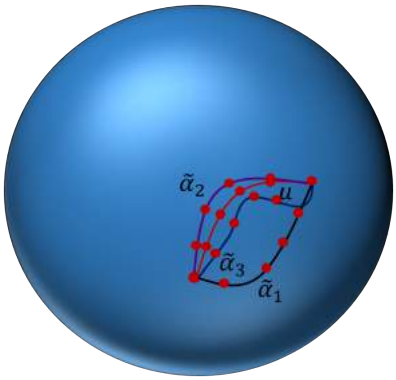


Figura 3.23: Discretización de las trayectorias.

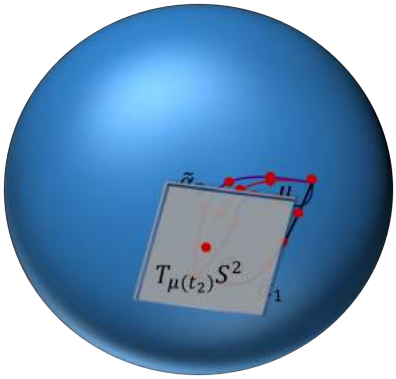
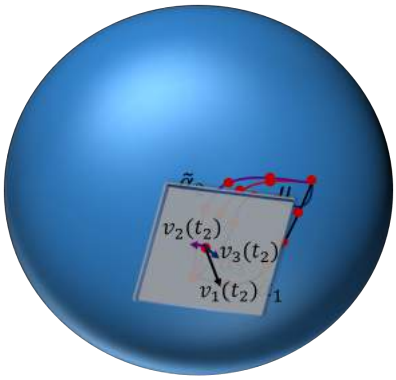


Figura 3.24: Plano tangente en $\mu(t_2)$.



$\widehat{K}(t_2) = \frac{1}{2}\sum_{i=1}^{3} v_i(t_2)v_i^T(t_2)$

Figura 3.25: *Shooting vectors* al tiempo t_2.

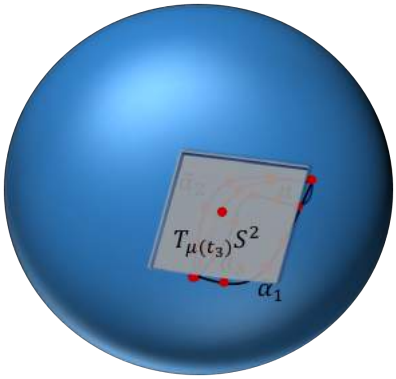
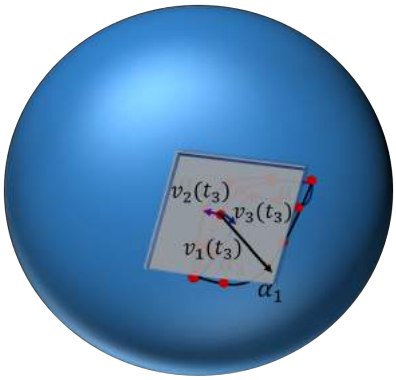


Figura 3.26: Plano tangente en $\mu(t_3)$.



$\widehat{K}(t_3) = \frac{1}{2}\sum_{i=1}^{3} v_i(t_3)v_i^T(t_3)$

Figura 3.27: *Shooting vectors* al tiempo t_3.

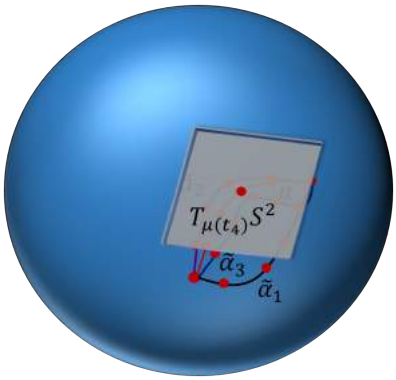
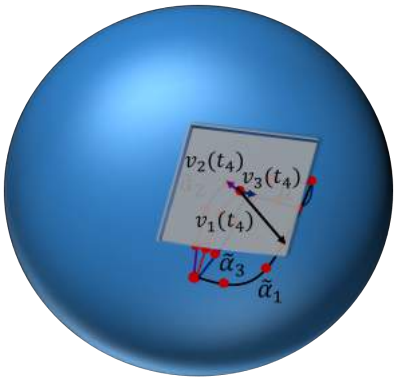


Figura 3.28: Plano tangente en $\mu(t_4)$.



$\widehat{K}(t_4) = \frac{1}{2}\sum_{i=1}^{3} v_i(t_4)v_i^T(t_4)$

Figura 3.29: *Shooting vectors* al tiempo t_4.

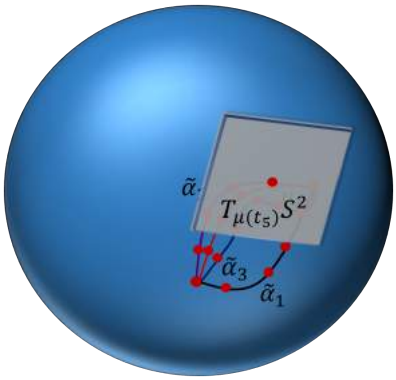
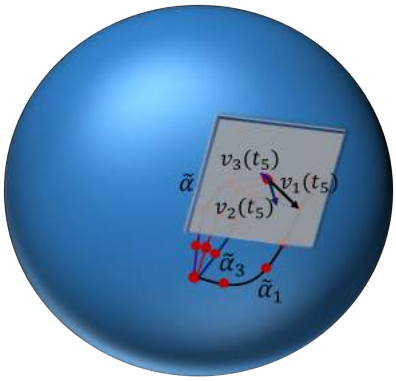


Figura 3.30: Plano tangente en $\mu(t_5)$.



$\widehat{K}(t_5) = \frac{1}{2}\sum_{i=1}^{3} v_i(t_5)v_i^T(t_5)$

Figura 3.31: *Shooting vectors* al tiempo t_5.

En el esbozo anteriormente presentado el conjunto de trayectorias $\{\tilde{\alpha}_i(t)\}_{i=1}^{3}$ tienen el mismo punto de inicio y el mismo punto de fin. En consecuencia, las varianzas correspondientes a los tiempos $t_1 = 0$ y $t_m = 1$ son cero. Sin embargo, es fundamental puntualizar que no necesariamente las trayectorias alineadas $\{\tilde{\alpha}_i(t)\}_{i=1}^{n}$ tienen el mismo punto de inicio y fin. Por tanto es necesario implementar el Algoritmo 3.4.2.1 en su totalidad.

3.4.3. Densidad de una trayectoria.

Uno de los usos más comunes que tienen la media y la varianza muestral es fungir cómo parámetros en un modelo de probabilidad, con el cual se busca capturar el comportamiento de los datos de interés. En el caso del análisis estadístico sobre variedades se tiene el mismo propósito; sin embargo es más complicado, pues el lugar donde se desea ajustar tal modelo es un espacio no lineal. Por consiguiente, dado el reto que impone esta tarea, es preferible trabajar en un espacio lineal; por ejemplo, en el caso del presente trabajo, T_cS^2. Esto implica que el lugar donde se definirá la

densidad de las trayectorias en estudio es el plano tangente a un punto en la esfera.

El modelo de probabilidad con el que se trabajará es una normal multivariada, la cual tendrá media cero y varianza $\hat{K}(t)$, tal que $\hat{K}(t)$ es la matriz de varianzas y covarianzas definida en el algoritmo anterior. Esta distribución será impuesta a los *shooting vectors* $v(t)$. Los pasos para obtener una estimación de la densidad de una trayectoria $\alpha(t)$ se enlistan a continuación.

Algoritmo 3.4.3.1 Densidad de una trayectoria $\alpha(t)$

Datos de entrada:

- Una trayectoria $\alpha(t)$ del conjunto de trayectorias observadas $\{\alpha_i(t)\}_{i=1}^{n}$.
- Trayectoria media discretizada, $\{\mu(t_j)\}_{j=1}^{m}$.
- Covarianza muestral de Karcher, $\hat{K}(t_j)$ tal que $j = 1, \ldots, m$.

La trayectoria $\alpha(t)$ debe ser discretizada, de manera que existan la misma cantidad de puntos $\alpha(t_j)$ que de puntos $\mu(t_j)$ y de matrices $\hat{K}(t_j)$. Es decir, para cada punto $\alpha(t_j)$ habrá una media $\mu(t_j)$ y una covarianza $\hat{K}(t_j)$, tal que $j = 1, \ldots, m$.

Datos de salida:

- Densidad de la trayectoria $\alpha(t)$.

Pasos:

1. Obtener los *shooting vectors* $v(t_j)$, entre $\mu(t_j)$ y $\alpha(t_j)$ tal que $j = 1, \ldots, m$. Notar que $v(t_j) \in T_{\mu(t_j)}M$.
2. Calcular una normal multivariada con los siguientes parámetros:
$$f(\alpha(t_j)) = N(v(t_j); 0, \hat{K}(t_j)).$$
3. Obtener el producto de las densidades $f(\alpha(t_j))$, como se muestra a continuación:
$$P(\alpha) = \prod_{j=1}^{m} f(\alpha(t_j)) = \prod_{j=1}^{m} N(v(t_j); 0, \hat{K}(t_j)). \tag{3.7}$$
En este caso $P(\alpha)$ representa la densidad de la trayectoria $\alpha(t)$.

El Algoritmo 3.4.3.1 puede ser útil para dar un p-valores de trayectorias simuladas. La simulación de trayectorias consiste en tomar el conjunto $\{(\mu(t_j), \hat{K}(t_j) \mid t_1 = 0, \ldots, t_m = 1\}$ y bajo alguna distribución simular los vectores $v(t_j)$. Posteriormente dichos vectores se devuelven a S^2 vía el mapeo exponencial. De esa forma se obtendrían los puntos que componen a la trayectoria simulada. Para obtener el p-valor

de una trayectoria simulada $\alpha(t)$, basta usar el método Monte Carlo. Esto significa, simular $N = 10000$ trayectorias y calcular $p(\alpha) = \sum_{i=1}^{N} \mathbf{1}_{P(X_i)<P(\alpha)}/N$, donde X_i representa a la i-ésima trayectoria simulada y $P(X_i)$ la densidad que ésta posee.

A continuación las Figuras 3.32–3.40 ejemplifican los pasos del algoritmo presentado.

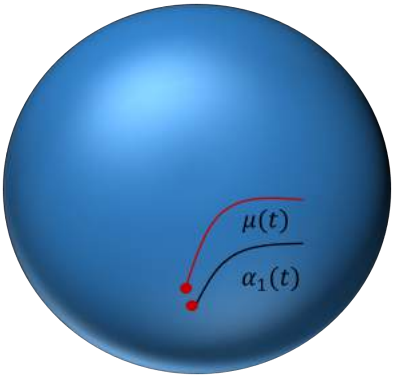


Figura 3.32: Trayectoria media y trayectoria sin alinear.

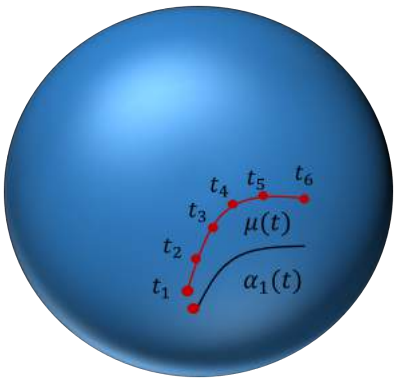


Figura 3.33: Discretizaión del tiempo igual que en el algoritmo de la varianza.

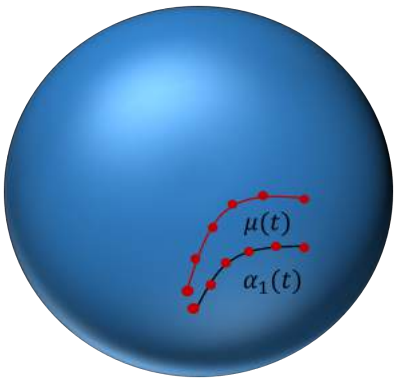
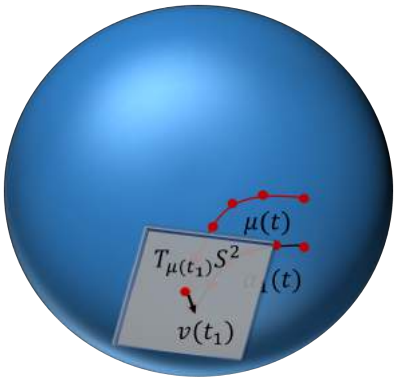
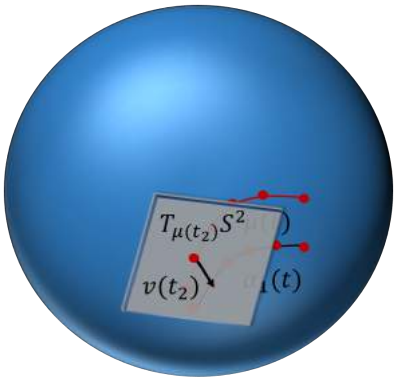


Figura 3.34: Discretización del tiempo en ambas trayectorias.



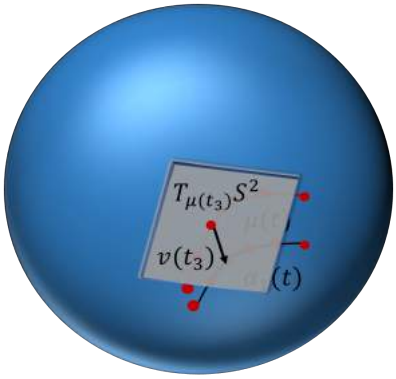
$f(\alpha(t_1)) = N(v(t_1); 0, \tilde{K}(t_1))$

Figura 3.35: *Shooting vector* al tiempo t_1 y densidad de $\alpha_1(t_1)$.



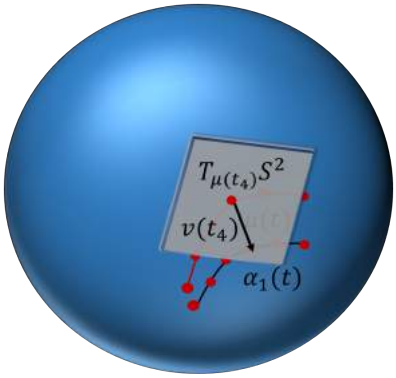
$f(\alpha(t_2)) = N(v(t_2); 0, \tilde{K}(t_2))$

Figura 3.36: *Shooting vector* al tiempo t_2 y densidad de $\alpha_1(t_2)$.



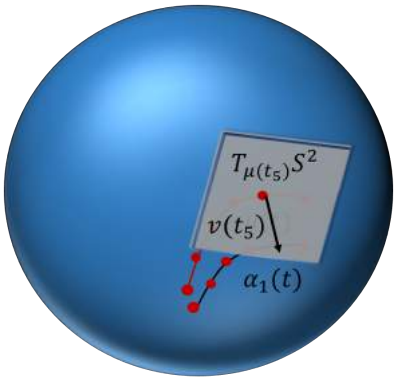
$f(\alpha(t_3)) = N(v(t_3); 0, \tilde{K}(t_3))$

Figura 3.37: *Shooting vector* al tiempo t_3 y densidad de $\alpha_1(t_3)$.



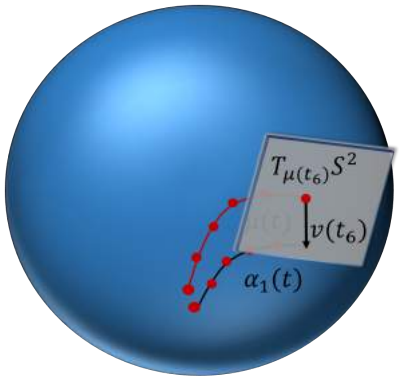
$f(\alpha(t_4)) = N(v(t_4); 0, \tilde{K}(t_4))$

Figura 3.38: *Shooting vector* al tiempo t_4 y densidad de $\alpha_1(t_4)$.



$f(\alpha(t_5)) = N(v(t_5); 0, \tilde{K}(t_5))$

Figura 3.39: *Shooting vector* al tiempo t_5 y densidad de $\alpha_1(t_5)$.



$f(\alpha(t_6)) = N(v(t_6); 0, \tilde{K}(t_6))$

Figura 3.40: *Shooting vector* al tiempo t_6 y densidad de $\alpha_1(t_6)$.

Es importante comentar que no hubo un proceso estadístico para ajustar el modelo de probabilidad normal a los vectores $v(t_j)$, de manera que esto podría considerarse como un punto sensible de este algoritmo. Por tal motivo, para un estudio de simulación, será necesario probar otras distribuciones y comparar resultados. De esa manera será posible obtener una intuición de cómo afecta la elección de la distribución a los resultados observados.

Es esencial notar que en ninguno de los algoritmos desarrollados se implementó de manera directa algún tipo de cálculo sobre S^2—excepto la media de Karcher asociada a los puntos $\{\alpha_i(0)\}_{i=1}^n$. Todos los procedimientos fueron realizados en un espacio lineal y vía alguna herramienta de geometría diferencial fueron devueltos a S^2. Esto es un indicador de la dificultad matemática y estadística que hay al trabajar en variedades no lineales. Por tanto, todavía existe teoría por refinar para hacer más accesibles herramientas y algoritmos en las áreas ya referidas.

Para concluir la presente sección se ofrecen algunas recomendaciones bibliográficas. En éstas, respectivamente, se podrá ahondar en temas como la importancia y dificultad de obtener la media de una forma, métodos numéricos para la resolución de ecuaciones diferenciales—como la que se presentó en el paso ocho del Algoritmo 3.4.1.1— y por último algunos ejemplos relacionados con análisis de imágenes donde fue empleada la metodología desarrollada en este capítulo.

1. Le and Kume [2000]. *The Fréchet mean shape and the shape of the means.*
2. Butcher [2005]. *The numerical analysis of ordinary differential equations.*
3. Su et al. [2014b]. *Rate-Invariant analysis of trajectories on riemannian manifolds with aplication in visual speech recognition.*

3.4.4. Análisis estadístico de trayectorias de huracanes

Con la finalidad de materializar y ejemplificar la utilidad de la teoría desarrollada, es que se decidió hacer un muy breve estudio de simulación. En ese mismo sentido, se planteó para mostrar el transporte paralelo y la trayectoria media de datos reales. El estudio de simulación será sobre ocho trayectorias de huracanes, las cuales se obtuvieron del siguiente sitio de Internet:

$$http://weather.unisys.com/hurricane/atlantic/.$$

Dichas trayectorias corresponden a un huracán seleccionado de los años de 1857, 1887, 1892, 1909, 1910, 1917, 1933 y 1944. Éstas se pueden observar en la Figura 3.41. Las características que comparten los huracanes se enuncian a continuación:

- Las trayectorias se encuentran en el Océano Atlántico.
- La velocidad de recorrido, en cada trayectoria, es diferente.
- Las observaciones asentadas se realizaron cada seis horas. Para ello, se consideró la latitud y longitud del lugar en el que se encontraba el huracán en dicho momento.

- Para cada trayectoria, la cantidad de observaciones es diferente.

- Las trayectorias tienen una forma similar, en el sentido de que nacen en la misma zona general del océano y su trayectoria inicial hacia el oeste, ingresando a tierra por el Golfo de México.

La elección de las trayectorias reseñadas obedeció al hecho de que comparten una curva similar, así como por otras razones que serán esclarecidas posteriormente. Basta mencionar por el momento que la motivación principal está relacionada con que el modelo probabilístico propuesto en el artículo base no resulta ser lo suficientemente flexible para albergar curvas muy disimilares.

Es relevante mencionar que no se realizó un proceso de interpolación en los datos que componen a cada trayectoria y tampoco se efectuó el proceso de alineación que propone el Algoritmo 3.4.1.1. El motivo principal fue por acotamiento del alcance de la tesis, ya que cada tarea implicaría en sí misma un proyeco sustancial de investigación e implementación computacional. Por tanto las trayectorias fueron trabajadas de forma "discreta", como se verá posteriormente.

Aclarados los puntos anteriores se procede con la implementación de los algoritmos. El primer paso es notar que la Tierra se puede concebir como una esfera. Por consiguiente las trayectorias de los huracanes se pueden representar en S^2, como muestra la Figura 3.42.

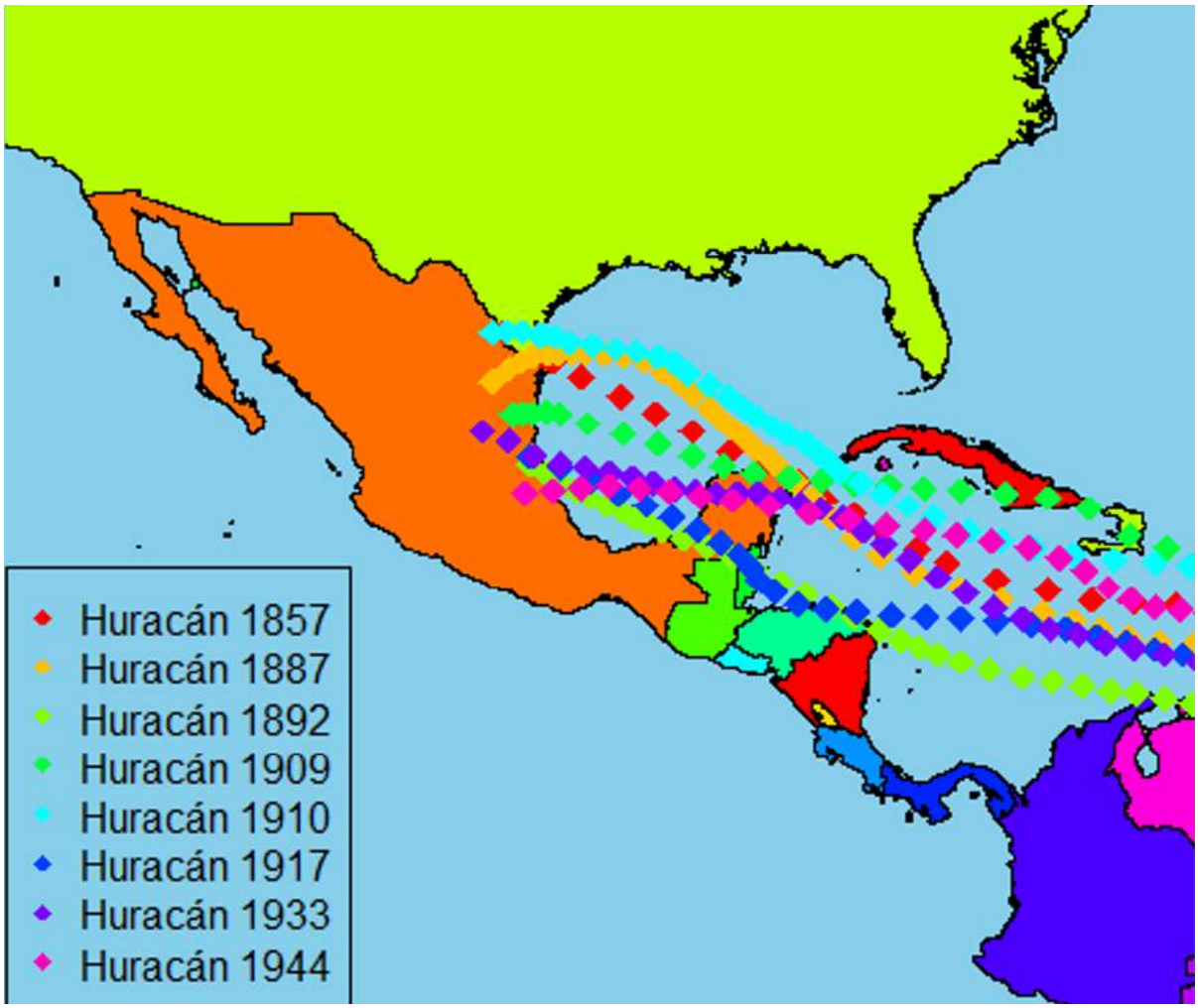


Figura 3.41: Ocho trayectorias de huracanes, pertenecientes al Oceáno Atlántico.

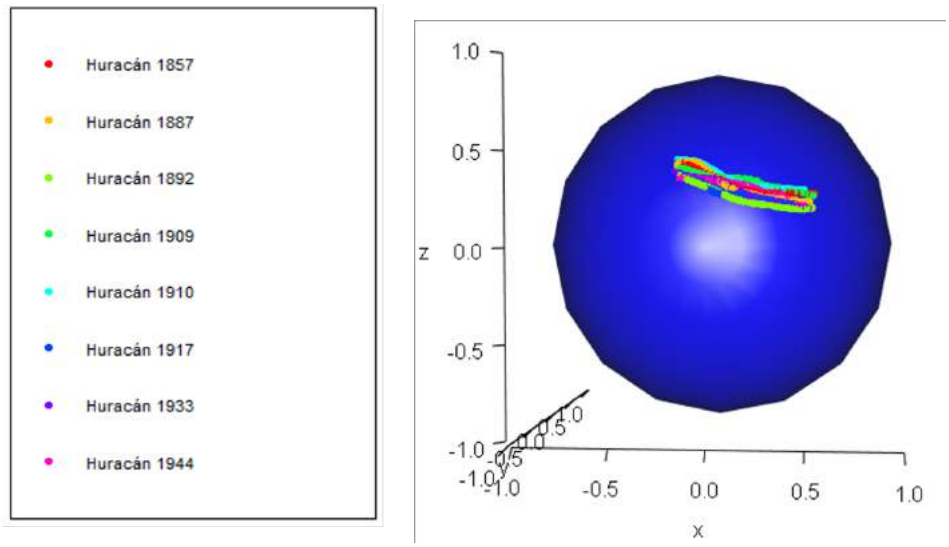


Figura 3.42: Trayectorias de huracanes sobre la esfera.

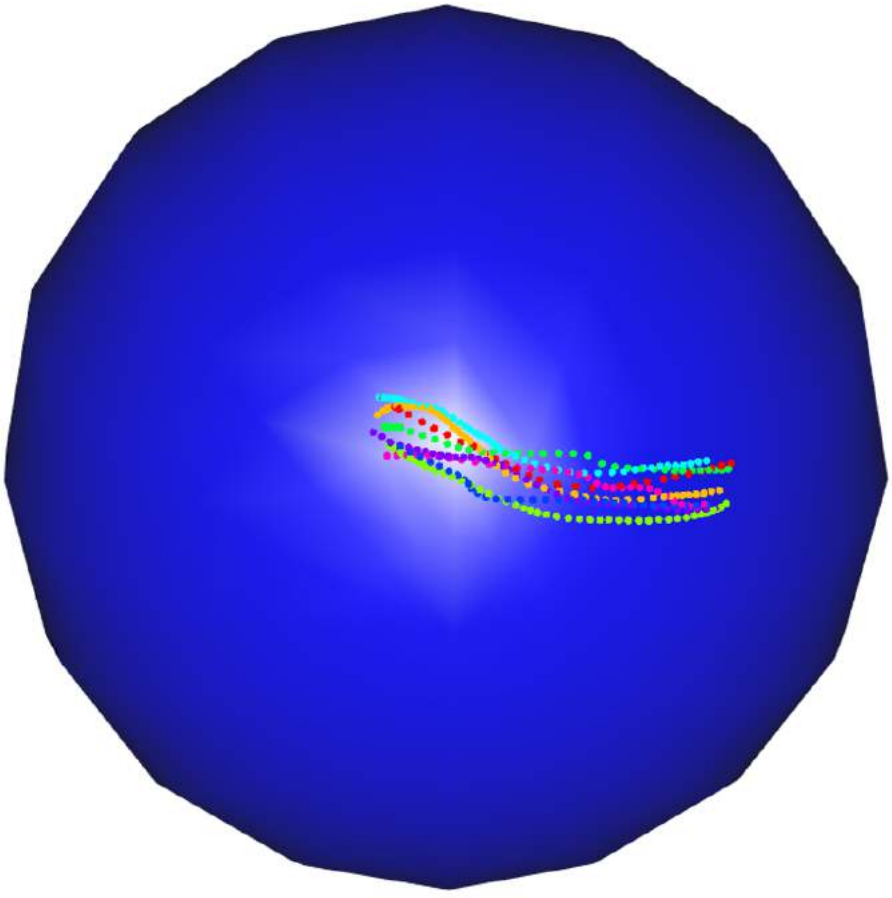


Figura 3.43: Acercamiento de las trayectorias en la esfera.

El primer algoritmo en ser implementado es el que corresponde al cálculo de la trayectoria media. Los datos de entrada son los puntos que conforman a cada una de las ocho trayectorias, así como el punto $c = (0, 0, 1)$ que representa el polo norte en la Tierra. Las ocho trayectorias serán denotadas como $\alpha_1(t), \alpha_2(t), \cdots, \alpha_8(t)$, respectivamente. La media de Fréchet (ver Sección 1.2.1) de los puntos iniciales de las trayectorias en cuestión, $\{\alpha_i(0)\}_{i=1}^{8}$, es el punto $\mu(0) = (0.5259418, -0.8174658, 0.2348080)$. Esta media al igual que los puntos $\alpha_i(0)$, donde $i = 1, \cdots, 8$, se pueden apreciar en la Figura 3.44.

Dado que no se realizó el proceso iterativo que sugiere el Algoritmo 3.4.1.1, no fue necesario elegir una trayectoria del conjunto $\{\alpha_i(t)\}_{i=1}^{8}$ para que fungiera como trayectoria inicial en el algoritmo ya citado (ver paso 2). Por tanto, bajo el contexto mencionado se calculó el TSRVF de las ocho trayectorias, con acuerdo en el paso 3, como se muestra en la Figura 3.45. Nótese que el transporte paralelo de estas trayectorias es muy parecido, lo cual es un indicador de que este concepto geométrico respeta la noción de cercanía o lejanía entre trayectorias.

Es valioso comentar que para obtener el campo velocidad, que sería usado en el transporte paralelo, se supuso que entre cada pareja de observaciones correspondientes a un huracán había una curva geodésica. Posteriormente se calculó la derivada—con respecto a t—de la función (2.4), que es una de las parametrizaciones de la curva geodésica, comentada en el Capítulo 2. Para ilustrar ideas, si una trayectoria $\alpha(t)$ está conformada por veintinueve puntos implica que se calcularán veintiocho curvas geodésicas y de cada una de ellas se obtendrá la derivada respecto a t, por consiguiente se transportarán veintiocho vectores a T_cS^2. Éstos representan el campo vectorial asociado a la trayectoria $\alpha(t)$. Por ende, dichos vectores ofrecerán una representación de la trayectoria $\alpha(t)$ en el plano tangente. Por otra parte, como consecuencia de la omisión del proceso iterativo, los pasos 4, 5 y 6 del algoritmo citado no fueron implemantados.

El siguiente paso es encontrar la trayectoria media. Para ello se eligieron veintiocho puntos "representativos" en cada $h_{\alpha_i}(t)$ tal que $i = 1, \ldots, 8$—el TSRVF de las trayectorias—. Dicha cantidad fue elegida debido a que era el menor número de puntos que conformaban a uno de los transportes paralelos. El criterio para elegir tales puntos en cada TSRVF fue vía porcentajes, se buscaron aquellos elementos que representaran[6] el $4\,\%, 7\,\%, 11\,\%, 14\,\%, 18\,\%, 21\,\%, 25\,\%, 29\,\%, 32\,\%, 36\,\%, 39\,\%, 43\,\%$, $46\,\%, 50\,\%, 54\,\%, 57\,\%, 61\,\%, 64\,\%, 68\,\%, 71\,\%, 75\,\%, 79\,\%, 82\,\%, 86\,\%, 89\,\%, 93\,\%, 96\,\%$ y $100\,\%$ del TSRVF en cuestión. Una vez realizado tal procedimiento se encontró la media muestral de los elementos $h_{\alpha_1}(t_j), h_{\alpha_2}(t_j), \ldots, h_{\alpha_8}(t_j)$ para cada tiempo t_j tal que $j = 1, \ldots, 28$; es decir $\mu(t_j) = 1/8 \sum_{i=1}^{8} h_{\alpha_i}(t_j)$. La Figura 3.46 muestra la trayectoria media en $T_{(0,0,1)}S^2$, tal que ésta se encuentra representada por los puntos negros.

[6]Los porcentajes que se muestran son resultado del desarrollo de la siguiente fórmula $\{(k \cdot 100)/28\}_{k=1}^{28}$, de manera que los números obtenidos sean redondeados.

Para representar el TSRVF de la trayectoria media en la esfera se resolvió la ecuación diferencial

$$\frac{d\mu(t)}{dt} =\mid h_\mu(t) \mid h_\mu(t) \tag{3.8}$$

correspondiente al paso ocho del Algoritmo 3.4.1.1. Para la resolución de ésta se consideró la aproximación

$$\frac{\mu(\delta) - \mu(0)}{\delta} \approx\mid h_\mu(\delta) \mid h_\mu(\delta).$$

Por consiguiente,

$$\mu(\delta) \approx \mu(0) + \delta \mid h_\mu(\delta) \mid h_\mu(\delta).$$

Usando este recurso de manera iterativa se obtuvo lo siguiente:

$$\begin{aligned}
\mu(\delta) &\approx \mu(0) + \delta \mid h_\mu(\delta) \mid h_\mu(\delta),\\
\mu(2\delta) &\approx \mu(\delta) + \delta \mid h_\mu(2\delta) \mid h_\mu(2\delta),\\
&\vdots\\
\mu(n\delta) &\approx \mu((n-1)\delta) + \delta \mid h_\mu(n\delta) \mid h_\mu(n\delta),
\end{aligned}$$

donde n es el número de puntos que conforman al TSRVF, en este caso $n = 28$. Por otro lado, para que los puntos $\mu(\delta), \mu(2\delta), \ldots, \mu(n\delta)$ cayeran en la esfera, se hizo una normalización de éstos. Es decir, se consideró la transformación

$$\mu^*(k\delta) = \frac{\mu(k\delta)}{|\mu(k\delta)|},$$

para $k = 1, \ldots, n$. De esta manera, los puntos $\mu^*(k\delta)$ fueron los que se graficaron en S^2. Como resultado se obtuvo la trayectoria de la Figura 3.47.

Con la finalidad de verificar la intuición, respecto al comportamiento de la trayectoria $\mu(t)$, se devolvieron los TSRVFs de las trayectorias de huracanes a la esfera, vía el razonamiento esbozado con anterioridad. Las trayectorias que se obtuvieron no conservan con toda exactitud la estructura de las trayectorias originales. El hecho descrito es causa de los errores numéricos, ocasionados por el método burdo que fue utilizado para resolver la ecuación diferencial. La Figura 3.48 muestra las trayectorias originales y las trayectorias que se obtuvieron vía la resolución de esa ecuación diferencial.

El siguiente algoritmo en implementarse es el 3.4.2.1, el cual refiere a la varianza asociada a un conjunto de trayectorias, como ya se había comentado este algoritmo arrojará un conjunto de cantidades que indicarán que tan semejantes son las trayectorias en ciertos tiempos.

Las covarianzas muestrales fueron obtenidas tomando como referencia cada uno de los veintiocho puntos que componen a la trayectoria media y considerando veintiocho puntos representativos en cada trayectoria $\alpha_i(t)$, $i = 1, \ldots, 8$. Dos comentarios

surgen en esta instancia; el primero es que el Algoritmo 3.4.2.1 trabaja con las trayectorias $\tilde{\alpha}_i(t)$, es decir con las trayectorias alineadas; sin embargo tal proceso no fue implementado. Por lo tanto el algoritmo citado se implementó con las trayectorias originales $\alpha_i(t)$, $i = 1, \ldots, 8$. El segundo comentario refiere a la obtención de los puntos que fueron considerados en las trayectorias $\alpha_i(t)$. Basta comentar que se tomaron aquellos puntos que representan los porcentajes considerados en el TSRVF. El Listing 1.1 muestra las matrices de varianzas y covarianzas $\hat{K}(t_{26}), \hat{K}(t_{27}), \hat{K}(t_{28})$ y las trazas de $\hat{K}(t_1), \ldots, \hat{K}(t_{28})$.

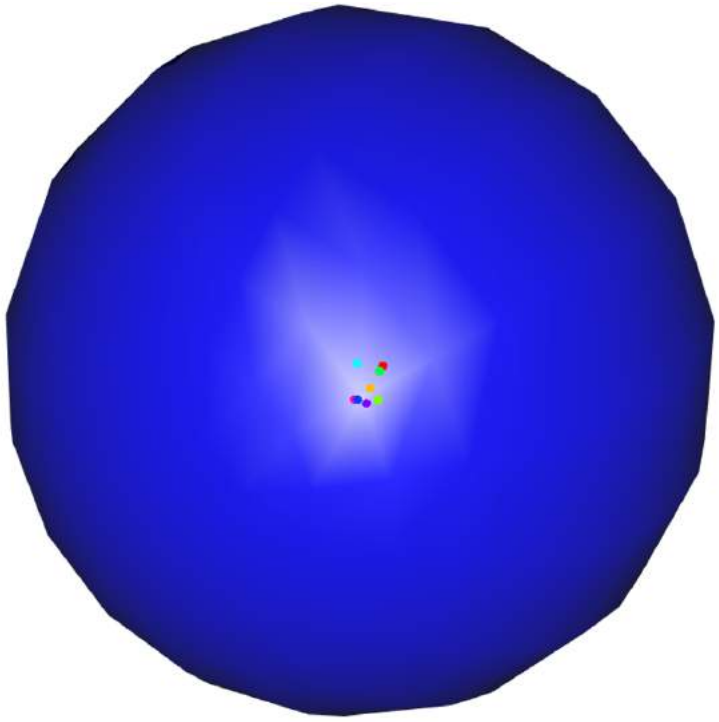
Como se puede apreciar en las covarianzas hay un cambio de signos, por ejemplo de $\hat{K}(t_{26})$ a $\hat{K}(t_{27})$. Esto indica que en el tiempo t_{27} hubo un cambio en el comportamiento de las trayectorias de huracanes y tal cambio es significativo por las unidades que hay de diferencia. Por otra parte las varianzas $\rho(t_1), \rho(t_2), \ldots, \rho(t_{28})$ son grandes, lo que indica que las trayectorias no están "bien" alineadas. Este último resultado era de esperarse, pues como se dijo no se implementó el algoritmo en cuestión con las trayectorias alineadas.

El paso final de este breve estudio es simular trayectorias de huracanes. Para ello se consideró una media $\mu(t) = (0, 0, 0)$ y matrices de varianzas y covarianzas de distintos órdenes. Dichas matrices fueron $\hat{K}(t_j), 1/10\hat{K}(t_j), 1/50\hat{K}(t_j), 1/100\hat{K}(t_j)$, tal que $j = 1, \ldots, 28$. En la Figura 3.49 se muestran las trayectorias de huracanes simuladas. Como se puede observar los puntos que conforman a la trayectoria simulada con las matrices $\{\hat{K}(t_j)\}_{j=1}^{28}$, Figura 3.49a, se encuentran totalmente dispersos principalmente en la parte final de la trayectoria. Es decir, que bajo la estructura impuesta de $\hat{K}(t_j)$ la trayectoria del huracán presenta un comportamiento errático. Para la segunda trayectoria simulada con $1/10\hat{K}(t_j)$, Figura 3.49b, los puntos que constituyen a la trayectoria siguen presentado un comportamiento errático; sin embargo se puede vislumbrar una trayectoria más "real" comparada con la anterior. Las últimas dos simulaciones poseen un comportamiento sensato, pues los puntos que las componen no están totalmente dispersos. Sin embargo, en estas dos trayectorias todavía se puede apreciar la mayor variabilidad en sus puntos terminales.

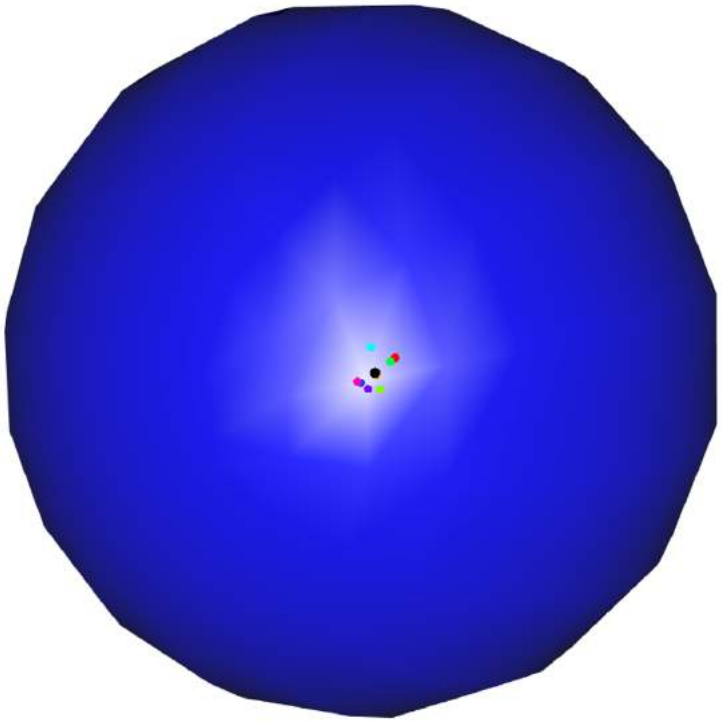
Es importante mencionar que dichas simulaciones, en general, no capturaron la estructura de los datos, pues todas las trayectoria simuladas quedaron en torno a la trayectoria media $\mu(t)$. Esto puede deberse a factores como la falta de interpolación en los datos, la ausencia de registro o que la distribución normal multivariada no es la adecuada para modelar los datos. A razón de esto es que se considera vital realizar los dos primeros procedimientos, y de esa misma forma explorar metodología para ajustar un modelo de probabilidad a los vectores $v(t)$. Todo ello con la finalidad de obtener resultados más consistentes con los datos. Para concluir, la metodología desarrollada en el presente capítulo es útil para conocer la probabilidad de que un huracán llegue a determinada costa del Oceáno Atlántico.

```
1
2  [[26]]
3                      x              y              z
4  [1,]   0.0003669986  -0.0001469341  -0.0003774803
5  [2,]  -0.0001469341   0.0015472662   0.0052943360
6  [3,]  -0.0003774803   0.0052943360   0.0181602811
7
8  [[27]]
9                      x              y              z
10 [1,]   2.593028e-04  -2.030742e-05  3.759466e-05
11 [2,]  -2.030742e-05   1.584926e-03  5.440292e-03
12 [3,]   3.759466e-05   5.440292e-03  1.871837e-02
13
14 [[28]]
15                     x              y              z
16 [1,] 0.0002343610 0.0001214343 0.0005279876
17 [2,] 0.0001214343 0.0015951797 0.0055130241
18 [3,] 0.0005279876 0.0055130241 0.0191054041
19
20 > traza
21  [1] 0.002895959 0.003021716 0.003050921 0.003175744 0.003360234
       0.003497778
22  [7] 0.003819601 0.003460979 0.004160103 0.004433891 0.005493516
       0.006702765
23 [13] 0.008176330 0.009099484 0.009618453 0.009957313 0.010922260
       0.011359632
24 [19] 0.011902401 0.013387441 0.014215386 0.015889487 0.016986101
       0.017979638
25 [25] 0.018698731 0.020074546 0.020562595 0.020934945
```

Listing 3.1: Matrices de variazas y covarianzas $\hat{K}(t_{26}), \hat{K}(t_{27}), \hat{K}(t_{28})$ y trazas de $\hat{K}(t_1), \ldots, \hat{K}(t_{28})$.



(a) Puntos iniciales de las ochos trayectorias en estudio



(b) Media Karcher representada por el punto negro.

Figura 3.44: Media de Karcher de los puntos iniciales de las ocho trayectorias de huracanes.

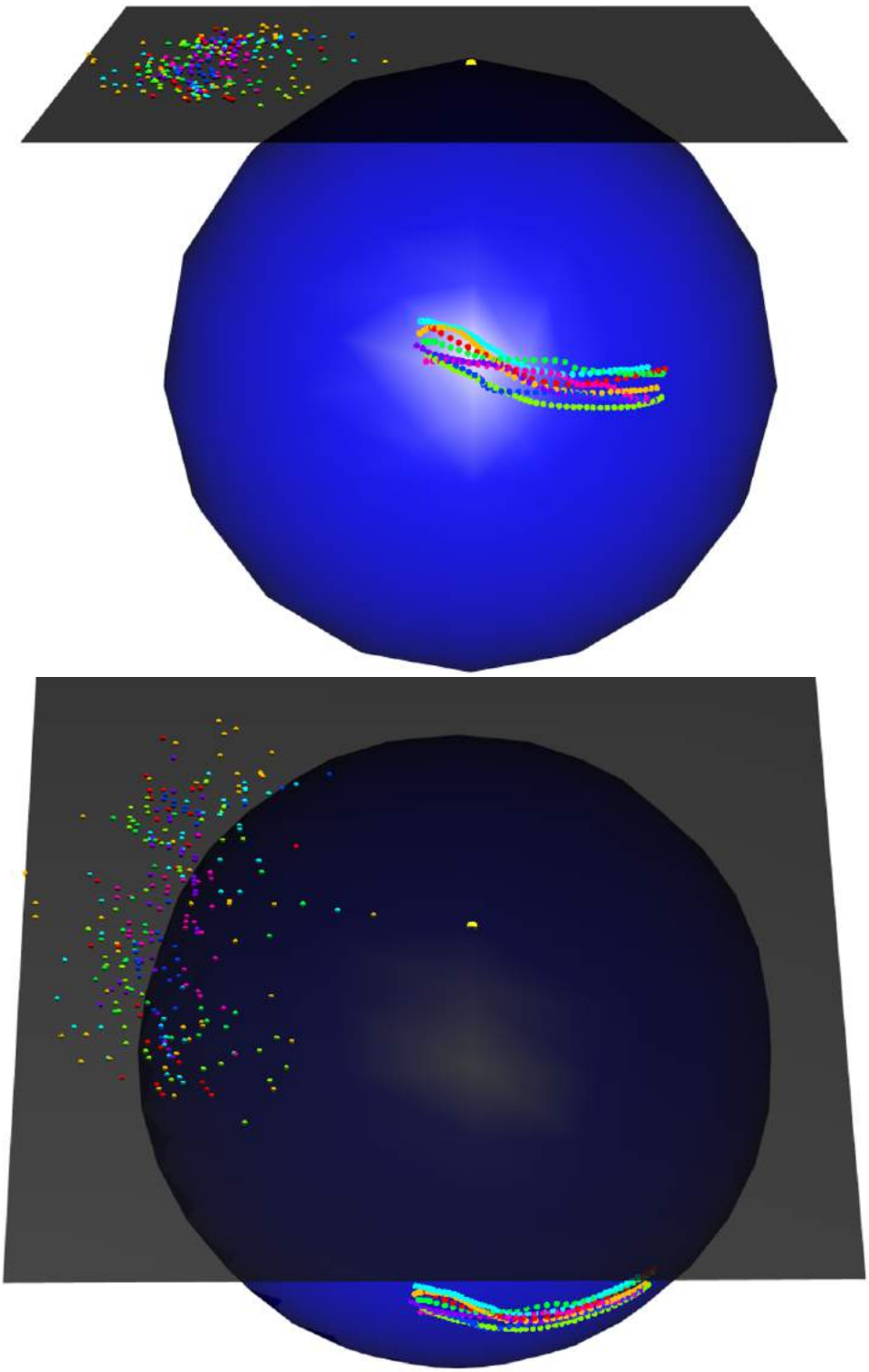


Figura 3.45: Plano tangente al $(0, 0, 1)$ y TSRVF de las ocho trayectorias de huracanes.

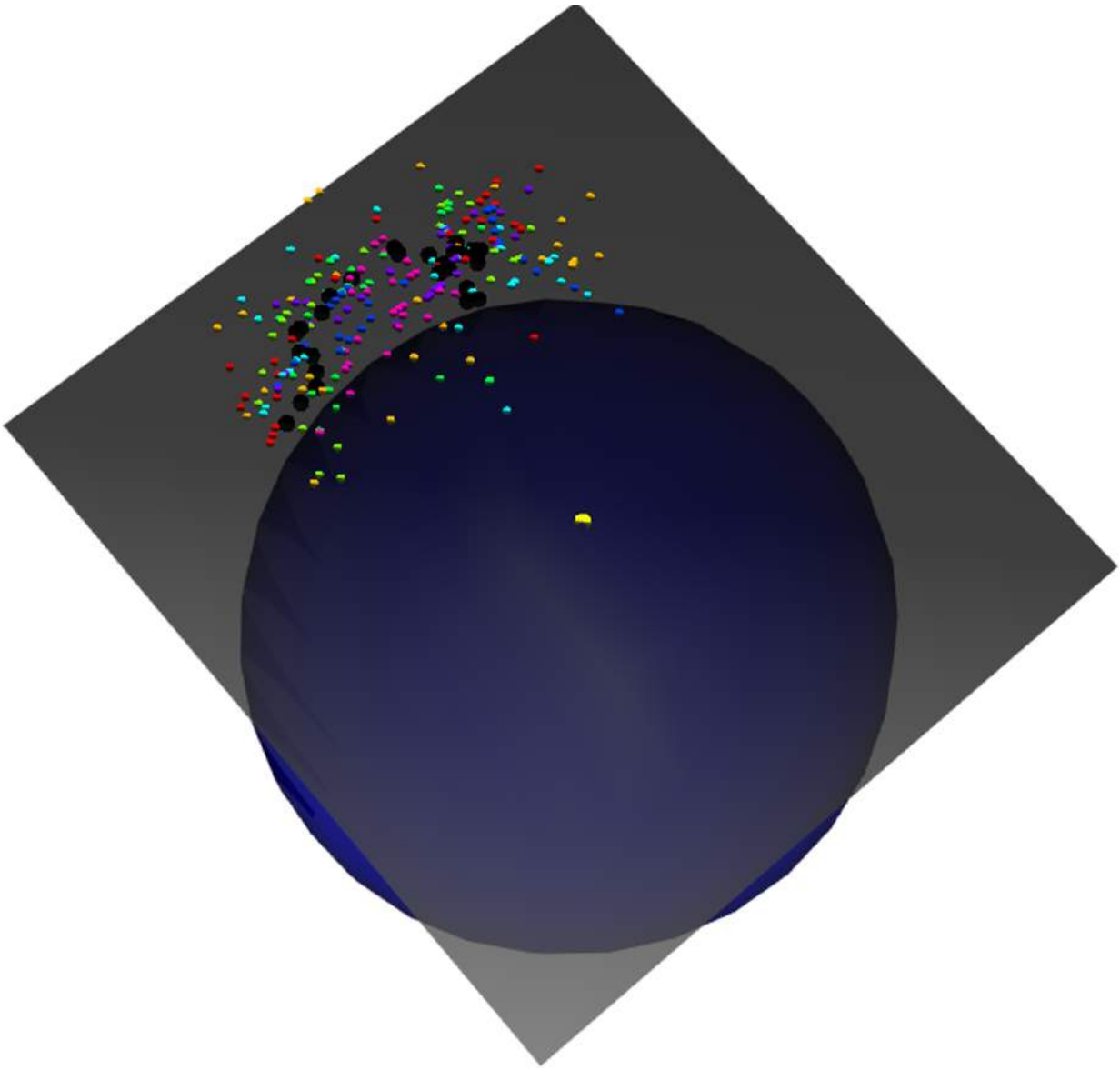


Figura 3.46: Veintiocho puntos de cada uno de los ocho TSRVFs de huracanes y la trayectoria media de dicho conjunto de TRSVFs.

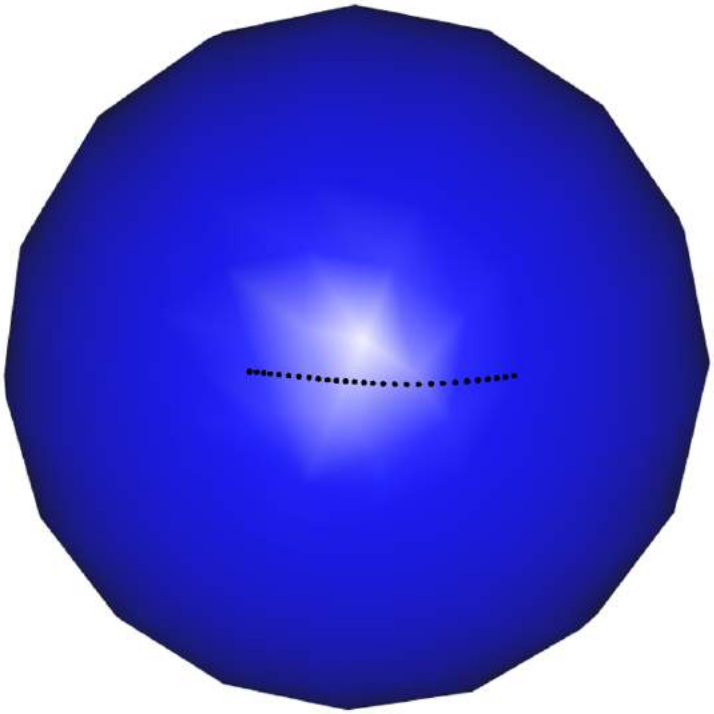
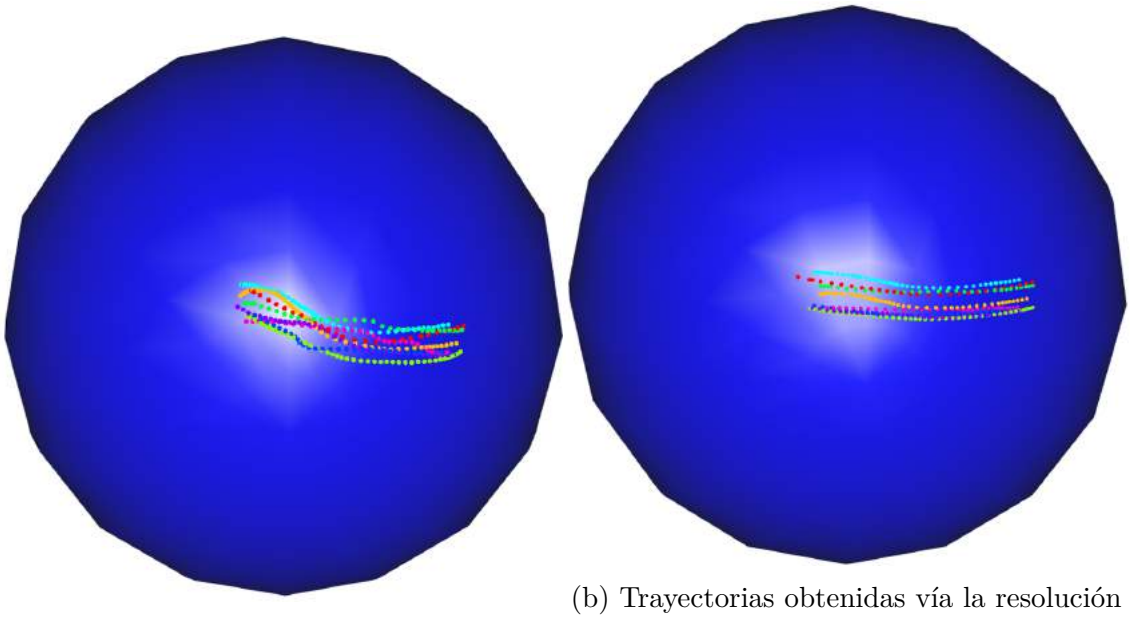


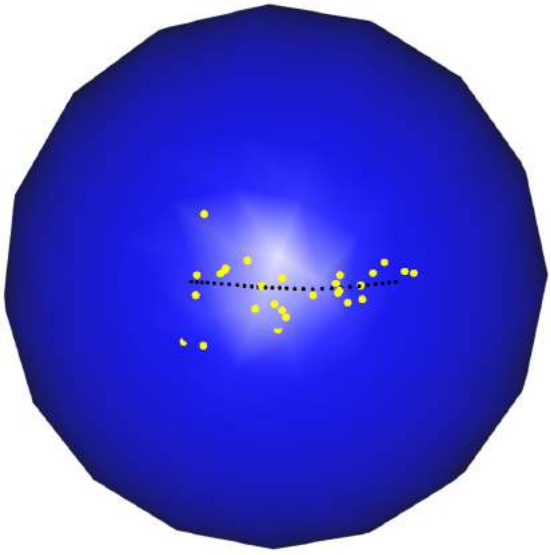
Figura 3.47: Trayectoria media en S^2.



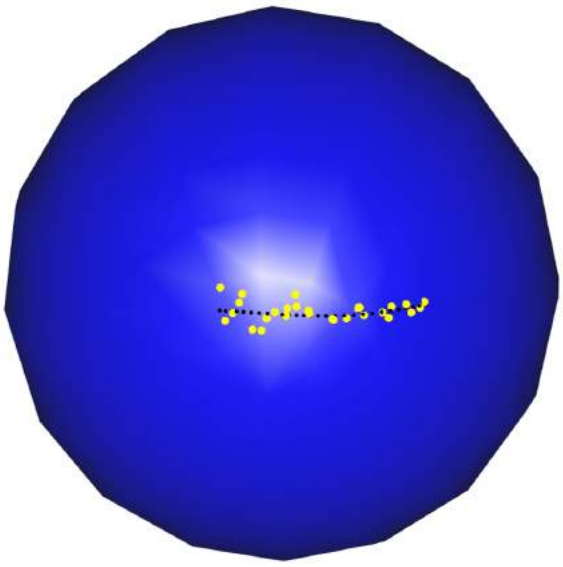
(a) Trayectorias originales.

(b) Trayectorias obtenidas vía la resolución de la ecuación diferencial.

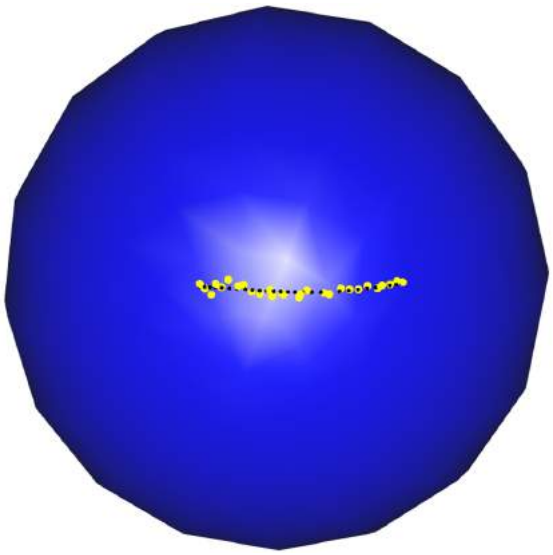
Figura 3.48: Comparación de la forma de las trayectorias de huracanes originales .



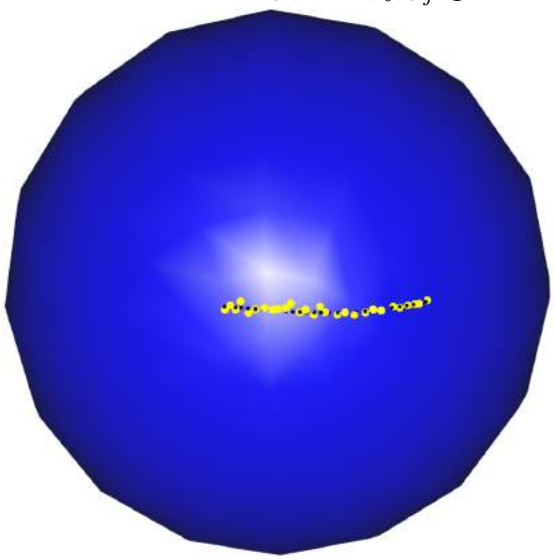
(a) Trayectoria de huracán simulada condiderando $\{\hat{K}(t_j)\}_{j=1}^{28}$.



(b) Trayectoria de huracán simulada condiderando $\{1/10\hat{K}(t_j)\}_{j=1}^{28}$.



(c) Trayectoria de huracán simulada condiderando $\{1/50\hat{K}(t_j)\}_{j=1}^{28}$.



(d) Trayectoria de huracán simulada condiderando $\{1/100\hat{K}(t_j)\}_{j=1}^{28}$.

Figura 3.49: Simulación de trayectorias de huracanes considerando distintas estructuras de varianzas y covarianzas.

3.5. Epílogo

El presente capítulo abordó la vinculación entre la geometría diferencial y la estadística y probabilidad. Así mismo trató conceptos de estadística sobre variedades, tales como el de media y varianza. De esa misma forma ofreció algunas aportaciones, las cuales se comentan a continuación:

1. Identificar el artículo base después de hacer una revisión bibliográfica de la temática.
2. Rellenar detalles técnicos del artículo base.
3. Proporcionar un resumen estructurado accesible.
4. Ofrecer un enriquecimiento bibliográfico.
5. Facilitar explicaciones heurísticas para aterrizar conceptos y terminologías.
6. Identificar y exponer conceptos técnicos.
7. Otorgar intuición de la teoría desarrollada, a lo largo del capítulo.
8. Dar conexiones con antecedentes teóricos.
9. Explicar pasajes complejos.
10. Detectar y enfatizar las aportaciones del artículo base.

Es importante mencionar que la teoría desarrollada así como los algoritmos presentados pueden generalizarse fácilmente en lo conceptual, cambiando la variedad S^2 por una variedad riemanniana M. El reto de tal generalización será la parte computacional, pues como se mencionó previamente varias nociones geométricas de interés no tienen una expresión analítica cerrada.

Para finalizar el capítulo se recomienda la lectura Turaga and Srivastava [2015]. Esta referencia, a pesar de ser propia del área de ciencias de la computación, contiene varios temas de vanguardia en lo que respecta a inferencia estadística sobre variedades. Por ejemplo PGA, análisis de regresión, manifold learning, estadística no paramétrica, entre otros.

Capítulo 4

Aportaciones y conclusiones

La motivación de la presente tesis radicó en la incursión y exploración de metodología para análisis estadístico de trayectorias sobre variedades. El principal objetivo de la tesis ha sido ofrecer un texto autocontenido que explique la teoría desarrollada en Su et al. [2014a]. Lo anterior requirió de presentar otros tópicos relacionados con el tema de estadística sobre variedades.

La inserción en la temática citada exigió una amplia búsqueda bibliográfica. Se localizaron temas y fuentes de interés concernientes a varias ramas de la estadística, en un contexto explícito de variedades. Entre ellas se destacan las siguientes, por tratarse de temas versátiles y recurrentes:

a) Modelos de probabilidad sobre variedades.
Bobrowski and Mukherjee [2014]. *The topology of probability distributions on manifolds.*

b) Manifold learning.
Lin and Zha [2008]. *Riemannian manifold learning.*
Izenman [2008]. *Modern multivariate statistical techniques*

c) Regresión sobre variedades.
Aswani et al. [2011]. *Regression on manifolds: Estimation of the exterior derivative.*

El tema de interés primordial fue materializado mediante el resumen *in extenso* del artículo Su et al. [2014a] titulado *Statistical analysis of trajectories on Riemannian manifolds: bird migration, hurricane tracking and video surveillance.* Este trabajo de síntesis fue desarrollado en el Capítulo 3, y refirió a su vez a otros temas de vanguardia en el área de estadística. Entre ellos vale la pena destacar las siguientes referencias por sus diversas aplicaciones en temas de actualidad:

a) Registro de trayectorias y sus aplicaciones.
Srivastava et al. [2011b]. *Registration of functional data using Fisher-Rao metric.*

b) Análisis elástico de curvas.
Joshi et al. [2016]. *Elastic Shape Analysis of Functions, Curves and Trajectories.*

c) Análisis de imágenes.
Nielsen and Barbaresco [2015]. *Geometric Science of Information.*
Turaga and Srivastava [2015]. *Riemannian Computing in Computer Vision.*

d) Interpolación de datos sobre variedades.
Samir et al. [2012]. *A gradient-descent method for curve fitting on Riemannian manifolds.*

Para hacer accesibles las nociones de geometría diferencial tratadas en el Capítulo 2, fue necesario hacer una excursión en ese tema tangencial. Se encontraron así libros que tratan la sinergía entre la geometría diferencial y la estadística. Como ejemplos se incluyen Shun-ichi [1985] con *Differential-geometrical methods in statistics* y Amari and Nagaoka [2007] con *Methods of information geometry.* Dichos textos fueron introducidos y reseñados por primera vez en el Capítulo 1. Constituye una aportación el haber expuesto aquellas definiciones de tal manera que fueran más accesibles para los lectores que carecen de una formación previa en geometría diferencial. Todas las ideas geométricas se abordaron en un contexto general para luego especializarlas en la esfera. Se complementó esto con una intuición verbal y gráfica, destacando de manera especial el transporte paralelo.

Por otra parte, se concluyó que para lograr una incursión exitosa en el análisis estadístico sobre variedades, es necesario contar con una formación—al menos básica—en tres áreas del conocimiento. Estas tres ramas de la matemática son vitales, ya que uno de los principales asuntos en el análisis estadístico sobre variedades es encontrar la "buena" métrica, con la cual sea posible establecer diferencias entre los datos de interés. Por tanto, es necesario identificar la estructura y propiedades del espacio en el que se encuentran. Tales ramas son las siguientes:

a) Geometría diferencial.

b) Teoría de la medida.

c) Topología elemental.

En el transcurso del estudio, surgió una recomendación indirecta para adentrarse en el área de análisis estadístico sobre variedades de manera gradual. Ésta consiste en comenzar con el estudio de métodos para datos direccionales. Estos datos se caracterizan por radicar en variedades como el círculo y la esfera. Un caso concreto fue abordado en el Capítulo 1, con realación a las tortugas terrestres. En este caso es más sencillo adoptar intución de las herramientas que son necesarias, para luego abordar la temática en un contexto general. Además, en dichos espacios las nociones topológicas y geométricas son más claras, ya que es posible contar con una representación gráfica, como es el caso de los conceptos de curva geodésica y espacio tangente a un punto, que fueron tratados en el Capítulo 2.

Para el desarrollo de la tesis requirió de identificar las ideas fundamentales para el planteamiento de los modelos descritos en Su et al. [2014a], en lo concerniente a la modelación estadística de trayectorias sobre variedades. Una vez identificadas

estas ideas se expusieron desde principios básicos. La finalidad e importancia de ello es que los puntos tratados resultaran accesibles al entendimiento y por ende clarificar y facilitar el proceso estadístico. A continuación se recapitulan éstas ideas fundamentales, las cuales pueden encontrarse en la Secciones 3.3 y 3.4.

a) Representar las trayectorias en un espacio lineal.

b) Deformar temporalmente las trayectorias, con la finalidad de hacerlas comparables.

c) Establecer una métrica para comparar trayectorias considerando deformaciones temporales.

d) Calcular la trayectoria media y matrices de varianzas y covarianzas de un conjunto de trayectorias en un espacio lineal, de manera análoga para la densidad y simulación de una trayectoria.

e) Regresar la trayectoria media y la trayectoria simulada a la esfera vía la resolución de una ecuación diferencial o herramientas de geometría diferencial, respectivamente.

Un resultado secundario del trabajo fue lograr un dimensionamiento del grado de dificultad del tema bajo consideración. Con base en la lectura realizada se obtuvo una concepción más clara de la dificultad del tema, así como de las herramientas y conocimientos previos que eran requeridos para su entendimiento. Por tal motivo, a lo largo del Capítulo 3 se proporcionaron referencias en las que se puede ahondar en temáticas como interpolación de datos en variedades riemannianas, registro temporal de trayectorias, métricas entre curvas y trayectorias sobre espacios no lineales, *etc.* Se recomendaron lecturas clasificadas por niveles de dificultad en la materia de estadística sobre variedades. Algunos libros citados y resumidos en el Capítulo 1, se enlistan a continuación en un orden que obedece a su dificultad progresiva. De esa forma, es posible notar cómo se enlanzan los conceptos estadísticos y geométricos desde sus principios fundamentales.

a) Mardia and Jupp [1999]. *Directional statistics.*

b) Patrangenaru [2015]. *Nonparametric Statistics on Manifolds and Their Applications to Object Data Analysis.*

c) Bhattacharya and Bhattacharya [2012]. *Nonparametric inference on manifolds: with applications to shape spaces.*

Se aprovechó la gran diversidad de materiales a los que hubo que dar lectura para encauzar una bibliografía anotada. En particular se recomendó lectura previa que enriqueciera los conocimientos del lector en lo que refiere al área de estadística sobre variedades. Algunas muestras de ello se dieron a lo largo del Capítulo 1, con el esbozo de los diferentes tópicos estadísticos que se han extendido a espacios no lineales como PCA, clustering, estadística no paramétrica, entre otros. De las temáticas citadas se proporcionaron las referencias pertinentes para lecturas más profundas (ver Sección 1.2). Por otra parte, a posterior elección de un artículo arbitrario referente al área, la

compilación de materiales preliminares presentada en esta tesis permite establecer con mayor facilidad muchos puntos esenciales. Así, esta revisión bibliográfica facilita la asimilación de la heurística y de las herramientas teóricas requeridas.

Uno de los principales retos que presentó la inserción en esta temática fue la labor computacional. En el área de estadística sobre variedades se carece de riqueza en cuanto a *software* implementado y accesible. Uno de los retos computacionales de la presente tesis fue implementar los conceptos de geometría diferencial abordados en el Capítulo 2. Principalmente giraron en torno a la noción de transporte paralelo, que jugó un rol esencial en el desarrollo estadístico. Gracias a este importante concepto, junto con los de mapeo exponencial y log-mapeo fue posible la descripción probabilística de trayectorias muestra. Esto a su vez formó la base para la simulación de huracanes y el examen de la ideosincrasia de trayectorias modeladas. Es importante resaltar que el cómputo de la función de deformación temporal $\gamma(t)$ no fue llevado al cabo, ya que por sí mismo amerita un enfoque computacional *ad hoc*, pues como se mencionó en el Capítulo 3 requiere del uso de programación dinámica. Esto significa que dentro del alcance de esta tesis no fue posible valorar la magnitud del efecto que pueda tener esta función, no obstante que en la literatura complementaria se hace alusión a que este concepto es vital.

En virtud del aprendizaje obtenido de la tesis, surgen algunos comentarios y conclusiones. Éstos son, en parte, un señalamiento crítico de ciertos pasos que no son comentados en el artículo base. Éstos pasos afectan al desarrollo de la teoría, y a los resultados calculados y su interpretación.

1. Existe una noción implícita de preprocesamiento en los datos. Es decir, para aplicar la metodología desarrollada en el artículo Su et al. [2014a], los datos de interés deben pasar por un proceso previo de interpolación. Ésto en sí mismo es un reto, pues no existe una amplia gama de herramientas para la interpolación de datos sobre variedades no lineales. A lo anterior se le auna el hecho computacional, ya que se requiere de un cómputo exhaustivo y la implementación no es inmediata.

2. La teoría de datos funcionales permite entender, en primera instancia, la esencia del artículo tomado como base. Esto se debe a que durante el desarrollo del texto se mencionan conceptos que son de uso frecuente en el área de FDA, por ejemplo, *variabilidad de fase* o *función de deformación temporal*. Lo anterior obedece al hecho de que uno de los puntos a desarrollar en el artículo es modelar la variabilidad temporal de las trayectorias. A decir el estudio de datos funcionales, bajo ciertos enfoques como el de Tucker et al. [2013], está íntimamente relacionado con el análisis de formas.

3. El registro temporal es una parte primordial del análisis estadístico de trayectorias. En el artículo se obvia el hecho de que el registro temporal es uno de los pasos primordiales en el estudio de trayectorias. De esa misma forma soslaya que la implementación de este procedimiento no es trivial y que en sí mismo el mecanismo para alinear trayectorias constituye un amplio tópico de investigación. Además, no se aclara que al realizar un proceso de registro hay cierta

pérdida de información. Por lo tanto, es necesario contemplar aquel registro temporal en el que se pierda la menor cantidad de información representativa de una trayectoria.

4. Debido a que el registro temporal es un paso vital en el análisis estadístico de trayectorias, se sugiere probar varias técnicas de registro y alineamiento de trayectorias para así adoptar aquella que sea más *ad hoc* con los datos. Lo anterior parece un hecho inocuo, y hasta quizás evidente. Sin embargo es vital porque no todas las trayectorias admiten la misma deformación temporal. En ese mismo sentido se desconoce cuánto puede impactar la elección de un método sobre otro en los resultados observados.

5. Es indispensable contemplar varias opciones distribucionales. En el artículo se impone un modelo de probabilidad normal para modelar el comportamiento de las trayectorias de huracanes. Sin embargo, como se observó en las simulaciones, dicho modelo no necesariamente captura con fidelidad el comportamiento de los datos. Esto no debería ser una sorpresa, pues al calcular la varianza como se mostró en el Algoritmo 3.4.2.1, no se proporciona una dirección principal a los *shooting vectors*. Aunado al hecho anterior, el modelo normal no parece acertado por la estructura *per se* que ostenta, ya que no ofrece una única dirección preferencial a la simulación de los vectores. Por tanto, se considera que modelos que contemplen colas pesadas unilaterales, como la χ-cuadrada, son más pertinentes para modelar la dirección que toma un huracán.

6. A continuación se comentan algunos puntos que pueden ser considerados como metodologías alternas para el tratamiento estadístico de trayectorias de huracanes. Estos puntos se compilan a partir de la experiencia obtenida tras el estudio de la metodología descrita en el artículo base. Toman en consideración aquellos detalles que se contemplaron como problemáticos para su implementación práctica, así como ideas diversas que fueron discernidas tras la revisión bibliográfica que esta tesis requirió.

    a) Modelar el comportamiento de las trayectorias de huracanes via una caminata aleatoria sobre la esfera. La idea subyacente es, dado que la trayectoria se encuentra en cierto punto temporal de su recorrido, con una probabilidad positiva se puede desplazar hacia "adelante" tomando alguna dirección de la esfera. Este enfoque se considera pertinente pues toma en cuenta la evolución temporal de la trayectoria, así como la probabilidad de moverse en alguna dirección particular de este espacio. Un texto que puede complementar esta propuesta es Roberts and Ursell [1960] con su trabajo titulado *Random walk on a sphere and on a Riemannian manifold*.

    b) Empleo de Cópulas para modelar la dependencia que existe entre cada punto que compone a una trayectoria. Se valora que un análisis de cópulas resulta conveniente, ya que con éste se habilita la posibilidad explorar distintas estructuras de dependecia que puede poseer la trayectoria de un huracán. Una referencia útil para el estudio de trayectorias sobre la

esfera, considerando el tratamiento ya mencionado, es Jupp [2015] con *Copulae on products of compact Riemannian manifolds*. Es importante resaltar que el enfoque sugerido es contrario al propuesto por el artículo base, ya que en este último lo que se modela es el comportamiento grupal de las trayectorias de huracanes.

Finalmente, se puede aseverar que el análisis estadístico sobre variedades es una rama joven de la estadística, lo cual conlleva que su teoría presente detalles finos por resolver. Algunos de los más comentados son los siguientes:

a) Caracterizar y ajustar un modelo de probabilidad.

b) Encontrar un criterio general para hablar de unicidad en la media.

c) Reducción del costo computacional en el desarrollo de algoritmos.

Estos puntos, por más pequeños que parezcan, han dado origen a una gran cantidad de disertaciones y charlas entre expertos del área; un ejemplo de ello es Hotz [2013], quien desarrolló un breve estudio de medias—extrínseca e intrínseca[1]—en el círculo. En este estudio comenta, cómo afecta el conocimiento de la distribución en la elección entre la media extrínseca o intrínseca en cuanto costo a cumputacional y robustez. Por tanto esta materia representa un área de oportunidad para estadísticos, computólogos, geométras y todo aquél científico que desee realizar análisis estadístico con datos más complejos que aquellos producidos en el espacio n-dimensional.

[1]Para conocer un poco de estos enfoques se sugiere consultar Bhattacharya [2013].

Apéndice A

Librerias usadas

```
1
2  ## Este script contiene todas las librerias que se usaran
3  ## para trabajar con otros scripts
4
5  library(rgl)       ## visualizaciones 3D
6  library(sphereplot)## trabajar graficos de la esfera
7  library(circular)  ## datos circulares
8  library(aspace)    ## trabajar radianes
9  library(plyr)      ## para aplicar funciones de forma sencilla
10 library(dplyr)     ## separar datos
11 library(tidyr)
12 library(mvtnorm)   ## trabajar con normal multivariada
13 library(lubridate) ## trabajar con fechas
14 library(stringr)   ## separar caracteres
15 library(MASS)      ## trabajar con la normal multivariada
```

Funciones utilizadas

```
1
2  ## Este script contiene todas las funciones que se usarán
3  ## para trabajar con otros scripts. Acontinuación se mencionan
4  ## las funciones que contiene.
5
6  ## Grafica de la esfera.
7  ## Grafica plano tangente en el punto (0,0,1).
8  ## Geodésica reparametrizada.
9  ## Log-mapeo.
10 ## Producto interno.
11 ## Derivada de una geodésica.
12 ## Norma de un vector.
13 ## Transporte paralelo.
14 ## Distancia en la esfera.
15 ## Producto de matrices.
16 ## Shooting vectors
17 ## Matriz de covarianzas
18 ## Regreso transporte paralelo
19 ## Función landmark
20
21 ## Grafica de la esfera
22 esfera<-function()
23 {
24   # crear un nuevo plot
```

```
  open3d()
  # generar la esfera
  spheres3d(x = 0, y = 0, z = 0, radius = 1,col="red",alpha =.9)
  # generar los ejes
  axes3d(c('x', 'y', 'z'))
  ## título y subtítulo
  title3d('','' ','x', 'y', 'z')
}

## Grafica plano tangente
plano<-function()
{
   f <- 0
   g <- 0
   h <- 1
   i <-  -.9999999
   planes3d(f, g, h, i, alpha = 0.8)
   points3d(0,0,1, col="yellow", size=10,lwd=10)

}

## Geodesica reparametrizada
## t=tiempo, p=punto inicio geodésica, v=dirección.
G<-function(t,p,v)
{
  nv<-sqrt(sum(v*v))
  return(cos(t*nv)*p + sin(t*nv)*(v/nv))
}

## Recordar que el mapeo exponencial es la geodesica evaluada
## en t=1

## Implementación geodesica
GC<-function(p,v,a)
{
  nv<-sqrt(sum(v*v))
  sapply(seq(0,pi/(a*nv),len=n2),G, p,v)
}

## Log mapeo
## p=punto de origen,  q0= a donde va
logM<-function(p, q0)
{
  if(all(q0==p)) return(c(0,0,0))
  return((acos(sum(p*q0)))/(sqrt(1- (sum(p*q0)^2)))*(q0-(sum(p*q0)*p)))
}

## Producto interno
Prod_int<-function(x,y) return(sum(x*y))

## Derivada de una geodesica
## t=tiempo, p=punto inicio geodésica, v=dirección.
DG<-function(t,p,v)
{
  nv<-sqrt(sum(v*v))
  return( (-sin(t*nv)*nv*p) + (cos(t*nv)*v))
```

```
}

## Norma de un vector
## x= vector
N_vec <- function(x) return(sqrt(sum(x^2)))


## Transporte paralelo
## p= punto de inicio, vl=velocidad, c= en donde se hará el transporte
TP<-function(p,vl,c)
{
  ## Norma de la suma suma de dos vectores elevado al cuadrado
  NS2<-(sum(p*p))+ (sum(c*c))+ (2*sum(p*c))
  ## Transporte paralelo
  ff<- vl -  ( (2*sum(vl*c)/NS2)*(p+c) )
  ## SRtvF
  ff<- ff/sqrt(N_vec(ff))
  return(ff)
}

## Distancia en la esfera
## p,q0= puntos de la esfera
dist_esf<-function(p,q0) return( acos(sum(p*q0)) )

## Producto matrices
## x=vector
Prod_M<-function(x) x%*%t(x)

## Shooting vectors
## SVect= Shooting Vectors
## npt= Numero puntos trayectoria
## tmu= Trayectoria mu
## ta= trayectoria a
SVect<-function(npt,tmu,ta)
{
  ## Matrix shooting vectors
  MSV<-NULL
  for(i in 1:npt)
  {
    sv<-logM(tmu[,i],ta[,1])
    MSV<-rbind(sv, MSV)
  }
  return(MSV)
}

## Matriz de covarianza
## MCov=Matriz de covarianzas
## Msv= Matriz shooting vectors
MCov<-function(Msv)
{
  ## Separar la matriz por columnas
  Msv<-as.list(split(Msv,col(Msv)))
  return(lapply(Msv, Prod_M))
}

## Regreso  transporte paralelo
```

```
Regreso_T<-function(TpT,P_ini,color)
{
  Reg_T<-matrix(0, nrow=dim(TpT)[1], ncol=3)
  Reg_T[1,]<- P_ini+ (1/dim(TpT)[1])*(N_vec(TpT[2,])*TpT[2,])
  Reg_T[1,]<- Reg_T[1,]/N_vec(Reg_T[1,])
  points3d(Reg_T[1,1],Reg_T[1,2],Reg_T[1,3])

  for(j in 2:dim(TpT)[1])
  {
    Reg_T[j,]<- Reg_T[j-1,]+ (1/dim(TpT)[1])*(N_vec(TpT[j,])*TpT[j,])
    Reg_T[j,]<- Reg_T[j,]/N_vec(Reg_T[j,])
    points3d(Reg_T[j,1],Reg_T[j,2],Reg_T[j,3], col=color, size=5,lwd=10)
  }
}

## función landmark
land<-function(por,m) round((por*m)/100) ## funcion landmarks

```

Capítulo 2

Transporte paralelo curvas geodésicas

```

##############################################
### Transporte Paralelo Curvas geodésicas ###
##############################################

esfera()
plano()

## Número de puntos en cada curva
n2<-100

## Curvas geodésicas
a<-GC(p=c(.0028,.9999,.000116),v=c(1/sqrt(2),0,1/sqrt(2)),a=3)
b<-GC(p=c(1,0,0),v=c(0,1,0),a=3)

## Gráfica curvas geodésicas
for(i in 1:n2)
{
  points3d(a[1,i],a[2,i], a[3,i], col="blue", size=5,lwd=10)
  points3d(b[1,i],b[2,i], b[3,i], col="green", size=5,lwd=10)
}

## Recorrido de las curvas geodésicas
tiempo<-function(a,v)
{
  nv<-sqrt(sum(v*v))
  seq(0,pi/(a*nv),len=n2)
}

## Campos velocidad
VectVa<-sapply(tiempo(3,v=c(0,1,0)),DG, p=c(.0028,.9999,.000116),v=c(1/sqrt(2),0,1/sqrt(2)))
```

```
VectVb<-sapply(tiempo(3,v=c(0,1,0)),DG, p=c(1,0,0),v=c(0,1,0))

## Grafica transporte paralelo curvas geodésicas
for(j in 1:n2)
{
  ## Transporte paralelo curva a
  tpa<-TP(a[,j],VectVa[,j],c=c(0,0,1))
  ## Transporte paralelo curva b
  tpb<-TP(b[,j],VectVb[,j],c=c(0,0,1))
  ##Gráfica transporte parlelo geodésica a
  points3d(tpa[1],tpa[2],tpa[3]+1, col="blue", size=5,lwd=10)
  ##Gráfica transporte parlelo geodésica b
  points3d(tpb[1],tpb[2],tpb[3]+1, col="green", size=5,lwd=10)
}
```

Transporte paralelo curva paralela

```

##############################################
#### Transporte paralelo curva paralela  ####
##############################################

esfera()
plano()

n2<-100 ## numero puntos cada curva
d<-1    ## longitud curva

### Curva paralela
M<-function(t)
{
  return( (1/2)*c(sin(t),cos(t),sqrt(3)) )
}

### Derivada de la curva paralela
DM<-function(t)
{
  return( (1/2)*c(cos(t),-sin(t),0) )
}

### Puntos curva paralela
ma<-sapply(seq(-pi,pi/d, len=n2),M)

### Grafica curva paralela
for(j in 1:n2) points3d(ma[1,j],ma[2,j], ma[3,j], col="yellow", size=5,lwd=10)

### Campo velocidad curva paralela
VectVma<-sapply(seq(-pi,pi/d, len=n2),DM)

### Grafica transporte paralelo curva paralela
for(j in 1:n2)
{
  tpa<-TP(ma[,j],VectVma[,j],c=c(0,0,1))
  points3d(tpa[1],tpa[2],tpa[3]+1, col="yellow", size=5,lwd=10)
}
```

Bibliografía

Amari, S.-i. and Nagaoka, H. (2007). *Methods of information geometry*, volume 191. American Mathematical Soc.

Aswani, A., Bickel, P., and Tomlin, C. (2011). Regression on manifolds: Estimation of the exterior derivative. *The Annals of Statistics*, pages 48–81.

Bhattacharya, A. and Bhattacharya, R. (2012). *Nonparametric inference on manifolds: with applications to shape spaces*, volume 2. Cambridge University Press.

Bhattacharya, R. (2013). A nonparametric theory of statistics on manifolds. In *Limit Theorems in Probability, Statistics and Number Theory*, pages 173–205. Springer.

Bobrowski, O. and Mukherjee, S. (2014). The topology of probability distributions on manifolds. *Probability Theory and Related Fields*, 161(3-4):651–686.

Butcher, J. C. (2005). *The numerical analysis of ordinary differential equations*. Wiley Online Library.

Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308.

Christensen, G. E. and Johnson, H. J. (2001). Consistent image registration. *Medical Imaging, IEEE Transactions on*, 20(7):568–582.

Do Carmo, M. P. (1976). *Differential geometry of curves and surfaces*, volume 2. Prentice-hall Englewood Cliffs.

Do Carmo Valero, M. P. (1992). *Riemannian geometry*.

Dryden, I. L. and Mardia, K. V. (1998). *Statistical shape analysis*, volume 4. Wiley Chichester.

Fisher, N. I. (1995). *Statistical analysis of circular data*. Cambridge University Press.

Fisher, N. I., Lewis, T., and Embleton, B. J. (1987). *Statistical analysis of spherical data*. Cambridge university press.

Fletcher, P. T., Lu, C., Pizer, S. M., and Joshi, S. (2004). Principal geodesic analysis for the study of nonlinear statistics of shape. *Medical Imaging, IEEE Transactions on*, 23(8):995–1005.

Fletcher, T. (2010). Terse notes on riemannian geometry.

Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. In *Annales de l'institut Henri Poincaré*, volume 10, pages 215–310.

Gallier, J. (2001). Basics of classical lie groups: The exponential map, lie groups, and lie algebras. In *Geometric Methods and Applications*, pages 367–414. Springer.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *Unsupervised learning*. Springer.

Hendriks, H. and Landsman, Z. (1996). Asymptotic tests for mean location on manifolds. *Comptes rendus de l'Académie des sciences. Série 1, Mathématique*, 322(8):773–778.

Hotz, T. (2013). Extrinsic vs intrinsic means on the circle. In *Geometric Science of Information*, pages 433–440. Springer.

Izenman, A. (2008). *Modern multivariate statistical techniques*, volume 1. Springer.

Joshi, S. H., Su, J., Zhang, Z., and Amor, B. B. (2016). Elastic shape analysis of functions, curves and trajectories. In *Riemannian Computing in Computer Vision*, pages 211–231. Springer.

Jung, S., Dryden, I. L., and Marron, J. (2012). Analysis of principal nested spheres. *Biometrika*, 99(3):551–568.

Jung, S., Foskey, M., and Marron, J. (2011). Principal arc analysis on direct product manifolds. *The Annals of Applied Statistics*, pages 578–603.

Jupp, P. (2015). Copulae on products of compact riemannian manifolds. *Journal of Multivariate Analysis*, 140:92–98.

Jupp, P. E. and Kent, J. T. (1987). Fitting smooth paths to speherical data. *Applied Statistics*, pages 34–46.

Karcher, H. (1977). Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541.

Kaziska, D. and Srivastava, A. (2008). The karcher mean of a class of symmetric distributions on the circle. *Statistics & Probability Letters*, 78(11):1314–1316.

Kneip, A. and Ramsay, J. O. (2008). Combining registration and fitting for functional models. *Journal of the American Statistical Association*, 103(483):1155–1165.

Kume, A. and Le, H. (2003). On fréchet means in simplex shape spaces. *Advances in Applied Probability*, pages 885–897.

Le, H. and Kume, A. (2000). The fréchet mean shape and the shape of the means. *Advances in Applied Probability*, pages 101–113.

Lee, J. M. (2006). *Riemannian manifolds: an introduction to curvature*, volume 176. Springer Science & Business Media.

Lin, T. and Zha, H. (2008). Riemannian manifold learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(5):796–809.

Liu, X. and Müller, H.-G. (2004). Functional convex averaging and synchronization for time-warped random curves. *Journal of the American Statistical Association*, 99(467):687–699.

Loring, W. T. (2008). An introduction to manifolds.

Mardia, K. V. and Jupp, P. E. (1999). Directional statistics.

Maybank, S. J. (2008). The fisher-rao metric. *Mathematics Today*, 44(6):255–257.

Nielsen, F. and Barbaresco, F. (2015). Geometric science of information.

Patrangenaru, V. (1998). *Asymptotic statistics on manifolds*. PhD thesis, Ph. D. dissertation, Indiana Univ.

Patrangenaru, Leif Ellingson, V. (2015). *Nonparametric Statistics on Manifolds and Their Applications to Object Data Analysis*. CRC Press.

Pennec, X. (2015). Barycentric subspaces and affine spans in manifolds. In *Geometric Science of Information GSI'2015*.

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

Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math.*, pages 81–91.

Rentmeesters, Q. and Absil, P.-A. (2011). Algorithm comparison for karcher mean computation of rotation matrices and diffusion tensors. In *Signal Processing Conference, 2011 19th European*, pages 2229–2233. IEEE.

Roberts, P. H. and Ursell, H. D. (1960). Random walk on a sphere and on a riemannian manifold. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 252(1012):317–356.

Samir, C., Absil, P.-A., Srivastava, A., and Klassen, E. (2012). A gradient-descent method for curve fitting on riemannian manifolds. *Foundations of Computational Mathematics*, 12(1):49–73.

Sánchez Morgado, H. and Palmas Velasco, O. A. (2007). Geometría riemanniana.

Shun-ichi, A. (1985). *Differential-geometrical methods in statistics*, volume 28. Springer Science & Business Media.

Souvenir, R. and Pless, R. (2005). Manifold clustering. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 648–653. IEEE.

Srivastava, A., Jermyn, I., and Joshi, S. (2007). Riemannian analysis of probability density functions with applications in vision. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE.

Srivastava, A., Klassen, E., Joshi, S. H., and Jermyn, I. H. (2011a). Shape analysis of elastic curves in euclidean spaces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(7):1415–1428.

Srivastava, A., Wu, W., Kurtek, S., Klassen, E., and Marron, J. (2011b). Registration of functional data using fisher-rao metric. *arXiv preprint arXiv:1103.3817*.

Su, J. (2013). Statistical analysis of trajectories on riemannian manifolds.

Su, J., Kurtek, S., Klassen, E., Srivastava, A., et al. (2014a). Statistical analysis of trajectories on riemannian manifolds: bird migration, hurricane tracking and video surveillance. *The Annals of Applied Statistics*, 8(1):530–552.

Su, J., Srivastava, A., de Souza, F. D., and Sarkar, S. (2014b). Rate-invariant analysis of trajectories on riemannian manifolds with application in visual speech recognition. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 620–627. IEEE.

Trouvé, A. and Younes, L. (2000). Diffeomorphic matching problems in one dimension: Designing and minimizing matching functionals. In *Computer Vision-ECCV 2000*, pages 573–587. Springer.

Tu, E., Cao, L., Yang, J., and Kasabov, N. (2014). A novel graph-based k-means for nonlinear manifold clustering and representative selection. *Neurocomputing*, 143:109–122.

Tucker, J. D., Wu, W., and Srivastava, A. (2013). Generative models for functional data using phase and amplitude separation. *Computational Statistics & Data Analysis*, 61:50–66.

Turaga, P. K. and Srivastava, A. (2015). Riemannian computing in computer vision.

Willard, S. (1970). General topology, addison.