USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Regresión Esférica

Roi Rendo Blanco

Curso 2024/2025

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

Regresión Esférica

Roi Rendo Blanco

Xullo, 2025

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Traballo proposto

Área de Coñecemento: Estatística e Investigación Operativa
Título: Regresión esférica
Breve descrición do contido
O modelo de regresión linear simple asume que tanto a variable resposta como a explicativa teñen soporte nun intervalo da recta real. Porén, na práctica, as variables do modelo poden ter como soporte un espazo non Euclídeo. En concreto, é frecuente atopar variables aleatorias con soporte na esfera unidade. Neste traballo proponse o estudo dun modelo de regresión onde tanto a variable explicativa como a resposta son esféricas. Para iso revisaranse os conceptos básicos relacionados coas variables esféricas, para posteriormente introducir o modelo de regresión esférico. Presentarase un caso práctico onde se aplicará o modelo de regresión esférico a datos reais.
Recomendacións
Cursar as materias de Inferencia Estatística e Modelos de Regresión e Análise Multivariante.
Outras observacións

Índice

Resumo **VIII**

1. Introdución á regresión **1**

1.1. Regresión lineal simple 1

1.2. Regresión lineal múltiple 4

1.3. Regresión non lineal 6

2. Introdución aos datos esféricos **9**

2.1. Medidas poboacionais e mostrais 9

2.2. Modelos de distribución destacados 14

2.2.1. Distribución de von Mises-Fisher 14

2.2.2. Distribución de Kent 16

2.2.3. Distribucións rotacionalmente simétricas 17

3. A regresión esférica **19**

3.1. Correlación esférica 19

3.2. Regresión empregando rotación 20

3.3. Estudo de simulación 22

3.3.1. Escenarios de simulación 22

3.3.2. Resultados 24

4. Conclusións e discusións **29**

4.1. Síntese do traballo 29

4.2. Posibles limitacións e extensións 30

A. Códigos de simulación **33**

B. Códigos gráficos **35**

Bibliografía **39**

Resumo

Este traballo estuda un modelo de regresión para variables esféricas, é dicir, aquelas que están definidas na superficie dunha esfera. Comeza cunha introdución aos conceptos e resultados básicos e necesarios da regresión lineal simple, da múltiple e da non lineal; para posteriormente comezar a tratar con datos esféricos, incorporando as definicións fundamentais. Para isto, axudarémonos de representacións gráficas. Incorporaremos tamén as distribucións máis importantes, onde nos resultará de especial interese a de von Mises-Fisher, xa que será a que empregaremos nos posteriores capítulos.

Unha vez presentado todo o anterior, poñerémolo en práctica realizando un estudo de simulación con R. Neste estudo introducimos o modelo de rotación para a regresión esférica, expoñendo algunhas das súas principais propiedades e interpretando os resultados obtidos.

Para rematar, discutiremos e explicaremos as posibles limitacións do traballo (uso exclusivo de datos simulados, referencias a demostracións, etc). Ademais disto, falaremos sobre como se podería estender este traballo se, por exemplo, cambiamos a unha dimensión maior, ou incluso mencionando outros modelos de distribucións coñecidos.

Abstract

This project studies a regression model for spherical variables, in other words, those that are defined on the surface of a sphere. It begins with an introduction to the basic and necessary concepts and results of simple linear regression, multiple and nonlinear regression; and then begin to work with spherical data, incorporating the fundamental definitions. For this purpose, we will use graphical representations. We will also incorporate the most important distributions, with the von Mises-Fisher distribution being of particular interest, as it will be the one we will use in subsequent chapters.

Once all the prior knowledge has been presented, we will put it into practice by conducting a simulation study using R. In this study, we introduce the rotation model for spherical regression, explaining some of its main properties and interpreting the results obtained.

Finally, we'll discuss and explain the potential limitations of this project (exclusive use of simulated data, references to proofs, etc). We'll also discuss how this work could be extended, for example, by switching to a larger dimension or even by mentioning other well-known distribution models.

Capítulo 1

Introdución á regresión

A regresión é unha técnica estatística empregada na análise de datos para tratar de explicar e predicir a relación entre distintas variables. Máis concretamente, trata de modelar a relación entre unha variable resposta fronte a unha ou varias variables explicativas.

Por outra parte, é unha técnica moi empregada en moitas outras áreas, como na economía e finanzas, medicina, enxeñaría, ciencias ambientais ou incluso nos deportes. Isto débese a que a regresión permite identificar e estudar as variables para logo ser capaces de tomar decisións, xa que será moito mais fácil predicir e estimar futuros valores, así como observar o impacto dos cambios nas variables.

Dentro da regresión distinguimos dous grandes tipos segundo o número de covariables empregadas: a simple e a múltiple. Neste capítulo imos a introducir ambos tipos, centrándonos na hipótese de linealidade (Seccións 1.1 e 1.2) e a maiores introduciremos a regresión non lineal na Sección 1.3. Os contidos que se tratarán neste capítulo baséanse en [1], [9] e [2].

1.1. Regresión lineal simple

Este modelo trata de explicar a relación entre unha variable resposta Y e unha única variable explicativa X. Adóitase formalizar como a media condicionada da variable resposta en función do valor da explicativa, é dicir, trataríase da seguinte función:

$$m(x) = \mathbb{E}(Y|X = x) \text{ para cada } x \in X.$$

Deste xeito, podemos expresar a variable resposta como:

$$Y = m(X) + \varepsilon,$$

onde ε é o erro, que representa a variabilidade da variable resposta por erros de medición ou outros factores non relacionados con X, e cumpre que $\mathbb{E}(\varepsilon|X = x) = 0$.

Por outra parte, considerando unha mostra $\{(x_1, Y_1), ..., (x_n, Y_n)\}$ de dimensión n, para realizar tarefas inferenciais este modelo debe cumprir unhas hipóteses básicas, que son as seguintes:

- Normalidade: $\varepsilon \in N(0, \sigma^2)$.
- Homocedasticidade: $Var(\varepsilon|X = x) = \sigma^2$.
- Linealidade: a función de regresión debe ser unha liña recta, co cal:

  $$Y = \beta_0 + \beta_1 X + \varepsilon$$

  onde β_0 é a ordenada na orixe, β_1 é a pendente da recta e ε o erro.
- Independencia: os erros aleatorios $\varepsilon_1, ..., \varepsilon_n$, son mutuamente independentes.

Unha vez visto o anterior, obteremos unha estimación dos parámetros β_0 e β_1 previamente mencionados, asi como da varianza do erro σ^2, empregando o chamado método de mínimos cadrados. Para isto, necesitaremos datos experimentais, onde previamente distinguiremos dous tipos de deseño experimental:

- Deseño fixo: os valores da variable explicativa están fixados polo experimentador, con vista á viabilidade do estudo.
- Deseño aleatorio: tanto a variable explicativa como a resposta son aleatorias.

Unha vez visto o anterior, realizaremos a estimación dos parámetros xa mencionados polo método de mínimos cadrados. Para isto, imos a supoñer as hipóteses de linealidade, homocedasticidade, normalidade e independencia, xunto con un deseño fixo.

Por outra parte, supóñase que se teñen estimadores xa calculados $\hat{\beta}_0$ e $\hat{\beta}_1$ de β_0 e β_1 respectivamente. Entón, poderiamos considerar os chamados residuos, definidos da seguinte forma:

$$\hat{\varepsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \quad \text{para } i \in \{1, ..., n\},$$

que representan a diferenza entre o valor observado Y_i para a observación x_i, e a súa predición $\hat{\beta}_0 + \hat{\beta}_1 x_i$.

Entón, o método de mínimos cadrados tratará de obter os estimadores que xeren os residuos máis pequenos posibles. Ademais, para evitar que se compensen os residuos positivos cos negativos, minimizaremos a suma de cadrados dos residuos. Deste xeito, estaremos buscando $\hat{\beta}_0$ e $\hat{\beta}_1$ tales que

$$\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \min_{\beta_0,\beta_1} \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2.$$

As expresións destes estimadores obtéñense facendo as derivadas parciais da ecuación anterior respecto a cada parámetro e igualando a cero. Logo, facemos as segundas derivadas e comprobamos que estamos ante mínimos absolutos. Vexámolo:

$$\frac{\partial \sum_{i=1}^{n}(Y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)^2}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0,$$

$$\frac{\partial \sum_{i=1}^{n}(Y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)^2}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0.$$

Deste xeito chegamos ás chamadas ecuacións normais, dadas por:

$$\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0,$$

$$\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0.$$

Resolvendo estas ecuacións para $\hat{\beta}_0$ e $\hat{\beta}_1$ chegamos ás expresións dos estimadores, como queriamos probar:

$$\hat{\beta}_0 = \bar{Y} - \frac{S_{xY}}{S_x^2}\bar{x}; \qquad \hat{\beta}_1 = \frac{S_{xY}}{S_x^2},$$

onde $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ son as medias mostrais da variable explicativa e resposta respectivamente, $S_{xY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})$ é a covarianza mostral e $S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ é a varianza mostral da variable explicativa.

Unha vez obtidos estes estimadores, podemos calcular tamén o estimador da varianza do erro:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon_i}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2,$$

dividindo por $(n-2)$ en lugar de por n, para que o estimador non teña nesgo.

Estes estimadores, obtidos polo método de mínimos cadrados, poderíanse obter tamén coa estimación por máxima verosimilitude. Isto débese a que ambas formas son equivalentes, grazas á suposición de normalidade. Para velo imos comezar considerando a función de verosimilitude dos coeficientes:

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(Y_i-\beta_0-\beta_1 x_i)^2}\right],$$

cuxo logaritmo será:

$$\log L(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2.$$

Entón, os valores de β_0 e β_1 onde alcanza o máximo este logaritmo coinciden cos valores onde a suma de cadrados $\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2$ alcanza o seu mínimo.

Se quixeramos facer inferencia sobre estes parámetros, deberiamos estudar as súas propiedades para posteriormente aplicalas nestas tarefas de inferencia. Para deducir estas propiedades imos supoñer as hipóteses de linealidade, homocedasticidade, independencia, normalidade e deseño fixo. Se realizamos este estudo para o estimador $\hat{\beta}_1$ chegaremos a que segue a seguinte distribución:

$$\hat{\beta}_1 \in N\left(\beta_1, \frac{\sigma^2}{nS_x^2}\right).$$

Do mesmo xeito podemos obter a distribución de $\hat{\beta}_0$:

$$\hat{\beta}_0 \in N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{nS_x^2}\right)\right),$$

e tamén:

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \in \chi^2_{n-2}.$$

Xa para rematar, cabe mencionar que, grazas aos resultados anteriores, seriamos capaces de obter de forma sinxela intervalos de confianza para os parámetros deste modelo de regresión, así como de realizar contrastes de hipóteses sobre eles.

1.2. Regresión lineal múltiple

Podemos estender o modelo de regresión lineal simple a casos mais xerais e complexos, onde haxa mais dunha variable explicativa. Deste xeito, partindo do modelo simple, podemos obter un modelo múltiple considerando unha combinación lineal das variables explicativas:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon,$$

onde sabemos que Y é a variable resposta, $X_1, ..., X_{p-1}$ as variables explicativas e ε o erro. É moi frecuente escribir esta fórmula en forma vectorial, polo que, no caso dun deseño fixo, podemos denotar $\mathbf{x}_i = (1, x_{i,1}, ..., x_{i,p-1})$ (vector fila) e $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{p-1})^\top$ (vector columna). Deste xeito, se consideramos o vector de variables explicativas $\mathbf{x}_i$ como vector columna, a función de regresión sería $m(\mathbf{x}_i) = \boldsymbol{\beta}^\top \mathbf{x}_i$.

Por outro lado, tamén é moi cómodo empregar unha notación matricial para escribir este modelo. Así, a forma $Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i$ sería equivalente ao seguinte:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}. \tag{1.1}$$

Á súa vez, chegariamos a unha expresión abreviada da forma

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

onde $\mathbf{Y}$ é o vector resposta, $\mathbf{X}$ é a chamada matriz de deseño, unha matriz de orde $n \times p$ onde cada fila representa a un individuo e cada columna a certa característica, $\boldsymbol{\beta}$ o vector de parámetros e $\boldsymbol{\varepsilon}$ un vector de erros que verifica que $\boldsymbol{\varepsilon} \in N_n(\mathbf{0}, \sigma^2 I_n)$.

Da mesma forma que realizamos co modelo simple, tamén podemos estimar os coeficientes por mínimos cadrados. En canto a $\boldsymbol{\beta}$, escolleremos como estimador aquel $\hat{\boldsymbol{\beta}}$ onde se alcance

$$\min_{\beta} \textstyle\sum_{i=1}^{n} (Y_i - \mathbf{x}_i\boldsymbol{\beta})^2,$$

que se pode expresar de forma equivalente en notación matricial:

$$\min_{\beta} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \min_{\beta} \phi(\boldsymbol{\beta}),$$

sendo ϕ a función obxectivo. Derivando esta respecto de $\boldsymbol{\beta}$ e igualando a cero, obtemos as chamadas ecuacións normais de regresión:

$$\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\top \mathbf{Y}.$$

Co cal, despexando $\boldsymbol{\beta}$, obtemos o estimador

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y},$$

sempre e cando a matriz $(\mathbf{X}^\top \mathbf{X})^{-1}$ teña determinante distinto de 0. Agora soamente falta calcular a estimación da varianza. Primeiro imos definir os residuos da seguinte forma, igual que fixemos no modelo simple:

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}} \quad i \in \{1, ...n\}.$$

Ademais, podemos crear un vector de residuos:

$$\mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y} = \mathbf{M}\mathbf{Y},$$

sendo $\mathbf{M} = \mathbf{I}_n - \mathbf{H}$ a chamada matriz xeradora de residuos.

Así, o estimador da varianza do erro sería

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}\hat{\varepsilon_i}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}})^2 = \frac{RSS}{n-p},$$

onde $RSS = \hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}} = \mathbf{Y}^\top\mathbf{M}\mathbf{Y}$, e dividimos por $(n-p)$ para que sexa un estimador sen nesgo.

Por outro lado, poderiamos realizar inferencia sobre estes parámetros previamente obtidos. Para iso necesitariamos coñecer as distribucións dos seus estimadores, co cal incorporamos o seguinte teorema:

Teorema 1.1. *Sexan $\varepsilon_1, ..., \varepsilon_n$ erros independentes e con distribución $N(0, \sigma^2)$, e $\mathbf{X}$ unha matriz $n \times p$ de rango p. Entón:*

I. $\hat{\boldsymbol{\beta}} \in N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$,

II. $\frac{RSS}{\sigma^2} = \frac{(n-p)\hat{\sigma}^2}{\sigma^2} \in \chi^2_{n-p}$,

III. $\hat{\boldsymbol{\beta}}$ *e* RSS *(ou* $\hat{\sigma}^2$*) son independentes.*

Podemos atopar a demostración deste teorema en [13]. Ademais, este resultado permítenos realizar inferencia para outro tipo de modelos que se atopen dentro do modelo de regresión lineal xeral.

1.3. Regresión non lineal

A diferenza do estudado ata aquí, a regresión non lineal modela unha relación entre variables que non pode ser descrita mediante unha liña recta. Deste xeito, mentres que a regresión lineal axusta unha liña recta cunha ecuación da forma $y = mx + b$, a regresión non lineal axusta unha curva que pode ser de moitas formas, dependendo do modelo que se queira axustar. Algúns exemplos son:

I. Exponencial: $y = a \cdot b^x$.

II. Logarítmico: $y = a + b \cdot ln(x)$.

III. Hiperbólico: $y = a + \frac{b}{x}$.

O modelo xeral nestes casos será

$$Y = m(X) + \varepsilon,$$

onde ε é o erro aleatorio que verifica que $\mathbb{E}(\varepsilon|X = x) = 0$, representando a variabilidade da variable resposta debido a erros de medición ou outras causas.

En canto ao seu axuste, o procedemento en base ao método de mínimos cadrados necesita dunha adaptación da expresión do Erro Cadrático Medio á correspondente ecuación e a súa posterior optimización:

$$\text{ECM} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - m(x_i))^2.$$

De todos modos, existen funcións non lineais que podemos "linealizar", é dicir, reescribila en forma dunha ecuación lineal. Exemplo disto é o seguinte: se a función a estudar é unha exponencial, $y = a \cdot e^{bx}$, podemos tomar logaritmos e así obteriamos $ln(y) = ln(a) + bx$, de xeito que, chamándolle $y' = ln(y)$ e $a' = ln(a)$, obtense o modelo de regresión lineal simple $y' = a' + bx$.

Por outra parte, en numerosos problemas será necesario empregar os chamados métodos iterativos, técnicas numéricas empregadas para resolver problemas nos que non é posible obter unha solución explícita. No noso caso, empregarémolos para estimar os parámetros do modelo axustando as solucións ata alcanzar un criterio de converxencia, que pode ser minimizar un erro ou maximizar unha función de probabilidade. Algún dos métodos iterativos máis empregados son o método de Gauss-Newton ([4]), o de Monte Carlo ([11]), o método de Nelder-Mead (Simplex) ou o de Levenberg-Marquardt ([8]).

Capítulo 2

Introdución aos datos esféricos

O concepto de datos esféricos fai referencia a un tipo de datos que se atopan distribuídos de forma esférica, é dicir, os datos teñen relación coas coordenadas nun espacio de varias dimensións, pero cunha estrutura que non depende dunha dirección particular. Os datos que son direccións poderanse representar como puntos na propia esfera, mentres que os datos que se corresponden cos eixos representaranse como pares de puntos diametralmente opostos sobre a esfera. En calquera caso, as observacións son de datos esféricos. Casos cotiáns nos que se empregan este tipo de datos son, por exemplo, estudos de imaxes da superficie terrestre, recollida de datos da temperatura ou da dirección do vento, estudos no ámbito de partículas físicas, etc. Máis exemplos atoparanse en [7], que ademais nos servirá de referencia para facer este estudo.

2.1. Medidas poboacionais e mostrais

En moitos casos, considerar unha dimensión p xeral resulta igual de sinxelo que o caso tridimensional ($p = 3$), que é o caso esférico, ao redor do cal xirará o estudo.

As direccións nunha dimensión p poderémolas representar como vectores unitarios $\mathbf{X}$, é dicir, como puntos de $S^{p-1} = \{\mathbf{X} : \mathbf{X}^\top \mathbf{X} = 1\}$, que é a esfera de dimensión $p-1$ de centro a orixe e raio unitario.

Por outra parte, no caso $p = 3$, será de utilidade empregar as chamadas coordenadas esféricas polares (θ, ϕ), definidas da seguinte forma:

$$\mathbf{X} = (\cos\theta, \operatorname{sen}\theta\cos\phi, \operatorname{sen}\theta\operatorname{sen}\phi)^\top,$$

onde θ denota a latitude e ϕ a lonxitude. Isto podémolo apreciar na Figura 2.1, onde se representan as coordenadas esféricas polares dun vector unitario $\mathbf{X}$ sobre a esfera unidade:

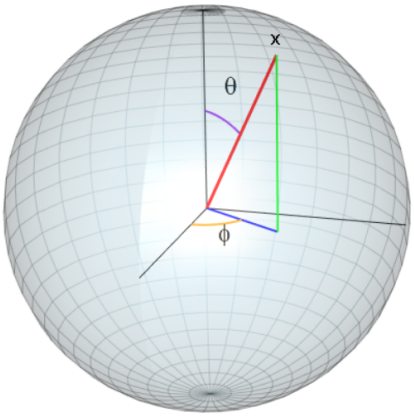

Figura 2.1: Representación da esfera unidade en $\mathbb{R}^3$, xunto cun vector unitario $\mathbf{X}$ e as súas coordenadas esféricas polares (θ, ϕ).

Ademais, para tratar de visualizar os datos na esfera unitaria tridimensional (en $\mathbb{R}^3$), resultará cómodo proxectalos sobre o plano. Entre distintas opcións, empregaremos a chamada *Proxección de áreas iguais de Lambert*:

$$(\cos\theta, \operatorname{sen}\theta\cos\phi, \operatorname{sen}\theta\operatorname{sen}\phi)^\top \quad \longmapsto \quad 2\operatorname{sen}(\tfrac{\theta}{2})(\cos\phi, \operatorname{sen}\phi)^\top,$$

que asigna a esfera unidade en $\mathbb{R}^3$ ao disco de de raio 2.

Por outro lado, un concepto de interese na análise de datos esféricos é a chamada descomposición *tanxente-normal*, que explica que calquera vector unitario $\mathbf{X}$ se pode descompoñer do seguinte xeito:

$$\mathbf{X} = t\boldsymbol{\mu} + (1-t^2)^{\frac{1}{2}}\boldsymbol{\xi},$$

con $\boldsymbol{\xi}$ tanxente unitario a S^{p-1} en $\boldsymbol{\mu}$. Podemos ver unha representación desta descomposición na Figura 2.2. Dende o punto de vista teórico, podemos atopar tres enfoques básicos para as estatísticas direccionais: o enfoque de *inmersión*, que considera á esfera S^{p-1} un subconxunto de $\mathbb{R}^p$; o *intrínseco*, no cal a esfera se considera como unha variedade en si mesma, sen facer referencia a ningunha inmersión; e o enfoque de *envoltura*, no cal os vectores $\mathbf{X}$ tanxentes á esfera en $\boldsymbol{\mu}$ están envoltos por

$$\mathbf{X} \longmapsto \operatorname{sen}(\|\mathbf{x}\mathbf{X}\|)\boldsymbol{\mu} + \cos(\|\mathbf{X}\|)\boldsymbol{\xi}$$

onde $\mathbf{X}^\top\boldsymbol{\mu} = 0$.

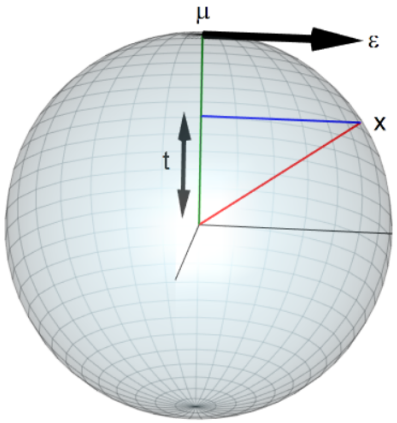

Figura 2.2: Representación na esfera unidade de $\mathbb{R}^3$ da descomposición tanxente-normal do vector $\mathbf{X}$.

Unha vez visto todo o anterior, podemos incorporar unha serie de definicións importantes. Para iso, imos a introducir o concepto de variable aleatoria esférica $\mathbf{X}$, que non é máis que un vector aleatorio de norma un.

Definición 2.1. Podemos definir a ρ para un vector unitario $\mathbf{x}$ como:

$$\rho = (\sum_{i=1}^{p} \mathbb{E}[X_i]^2)^{\frac{1}{2}} = \{\mathbb{E}[\mathbf{X}]^\top \mathbb{E}[\mathbf{X}]\}^{\frac{1}{2}}.$$

Cando $\rho > 0$, defínese a *dirección media da poboación* como

$$\boldsymbol{\mu} = \rho^{-1}\mathbb{E}[\mathbf{X}],$$

que é un vector unitario que representa a orientación promedio dos datos distribuídos na esfera. En canto a ρ, este parámetro mide que tan concentrados se atopan os datos arredor da dirección media na esfera, tomando valores no intervalo [0,1]: $\rho = 1$ quere dicir que están aliñados na mesma dirección, mentres que $\rho = 0$ indica que os datos están distribuídos uniformemente pola esfera. Como observación, no caso $\rho = 0$, $\boldsymbol{\mu}$ non está definido, xa que non hai unha dirección predominante debido á distribución uniforme ou simétrica dos datos na esfera. Na Figura 2.3 represéntanse datos simulados dunha variable esférica con distintos valores de $\boldsymbol{\mu}$ e ρ:

Agora, para o que vén, imos a considerar unha mostra $\mathbf{X}_1, ..., \mathbf{X}_n$ de observacións da esfera S^{p-1}:

Definición 2.2. A media mostral destas observacións será un vector de $\mathbb{R}^p$ do seguinte xeito:

$$\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i.$$

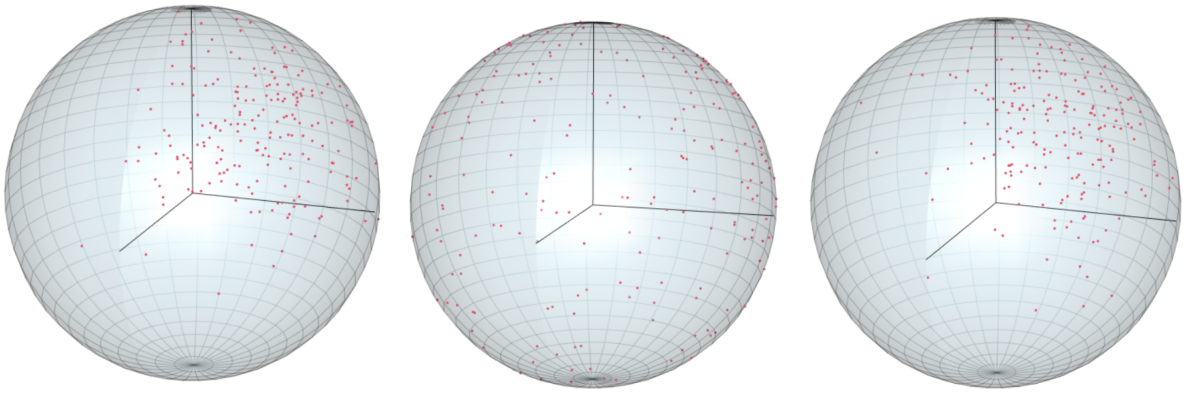

Figura 2.3: Representación na esfera unidade da simulación de 200 datos con $\boldsymbol{\mu} = \left(\frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}\right)$ e $\rho = 0{,}948$ (esquerda); $\boldsymbol{\mu} = \left(\frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}, \frac{\sqrt{3}}{3}\right)$ e $\rho = 0{,}446$ (centro) e $\boldsymbol{\mu} = (1,1,1)$ e $\rho = 0{,}948$ (dereita).

Será útil expresar este vector $\bar{\mathbf{X}}$ en forma polar como

$$\bar{\mathbf{X}} = \bar{R}\bar{\mathbf{X}}_0,$$

onde $\bar{\mathbf{X}}_0$ é un vector unitario e $\bar{R} \geq 0$, co cal $\bar{R} = \|\bar{\mathbf{X}}\|$, que se chamará a *lonxitude resultante media* e $\bar{\mathbf{X}}_0 = \|\bar{\mathbf{X}}\|^{-1}\bar{\mathbf{X}}$, que recibirá o nome de *dirección media*. Este $\bar{R}$ é unha medida de concentración, e mide que tan aliñados están os datos: cando $\bar{R} = 1$ estaremos no caso no que todos apuntan na mesma dirección, mentres que canto máis próximo se atope do 0, os datos estarán cada vez máis distribuídos uniformemente na esfera. En canto a $\bar{\mathbf{X}}_0$, é un vector unitario que apunta na dirección media dos datos. Estes parámetros serán os análogos de ρ e $\boldsymbol{\mu}$, previamente definidos.

Ademais, se os puntos $\mathbf{X}_1, ..., \mathbf{X}_n$ tiveran a mesma masa, entón o seu centro de masas sería $\bar{\mathbf{X}}$, con dirección $\bar{\mathbf{X}}_0$ e distancia á orixe $\bar{R}$. Vexámolo na Figura 2.4.

A lonxitude media resultante ten a seguinte propiedade: sexa $\mathbf{a}$ pertencente a S^{p-1}, e denotemos por $S(\mathbf{a})$ o cadrado da media aritmética das distancias euclídeas entre $\mathbf{X}_i$ e $\mathbf{a}$. Entón:

$$S(\mathbf{a}) = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \|\mathbf{X}_i - \mathbf{a}\|^2 = 2(1 - \bar{\mathbf{X}}^\top\mathbf{a}) = 2(1 - \bar{R}\bar{\mathbf{X}}_0^\top\mathbf{a}).$$

De aquí sacamos que $S(\mathbf{a})$ se minimiza cando $\mathbf{a} = \bar{\mathbf{X}}_0$ e así (sabendo que $\mathbf{a}^\top\mathbf{a} = 1$):

$$\min_{\mathbf{a}} S(\mathbf{a}) = 2(1 - \bar{R}).$$

A esta cantidade $2(1 - \bar{R})$ adóitaselle chamar a *varianza esférica* da mostra, que é unha medida de dispersión.

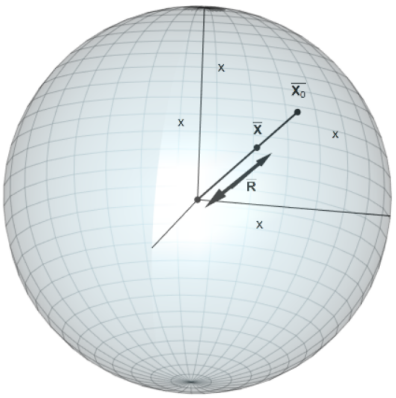

Figura 2.4: Dirección media $\bar{\mathbf{X}}_0$ e Lonxitude media resultante $\bar{R}$.

Sexa agora $\boldsymbol{\mu}$ calquera vector unitario e

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i^\top \boldsymbol{\mu},$$

a media mostral das compoñentes $\mathbf{X}_1, ..., \mathbf{X}_n$ ao largo de $\boldsymbol{\mu}$. Así:

$$2n(1-\bar{C}) = \sum_{i=1}^{n} \|\mathbf{X}_i - \boldsymbol{\mu}\|^2,$$

onde $2n(1-\bar{C})$ se pode considerar a variación total de $\mathbf{X}_1, ..., \mathbf{X}_n$ sobre $\boldsymbol{\mu}$. Esta variación pódese descompoñer do seguinte xeito:

$$2n(1-\bar{C}) = 2n(1-\bar{R}) + 2n(\bar{R}-\bar{C}).$$

Por tanto, esta expresión é aproximadamente a parte do espazo tanxente da seguinte descomposición:

$$\sum_{i=1}^{n} \|\mathbf{X}_i - \boldsymbol{\mu}\|^2 = \sum_{i=1}^{n} \|\mathbf{X}_i - \bar{\mathbf{X}}\|^2 + n\|\bar{\mathbf{X}} - \boldsymbol{\mu}\|^2.$$

Agora podemos introducir outra medida de dispersión interesante, que é a matriz

$$\bar{\mathbf{T}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top,$$

que se pode interpretar como o *tensor de inercia* arredor da orixe de partículas de peso n^{-1} en cada un dos puntos $\mathbf{X}_1, ..., \mathbf{X}_n$.

Podemos denotar como $\mathbf{S}$ a matriz de varianza mostral, que ven dada pola seguinte expresión:

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}}_0)(\mathbf{X}_i - \bar{\mathbf{X}}_0)^\top,$$

o que nos permite escribir $\bar{\mathbf{T}}$ da seguinte forma:

$$\bar{\mathbf{T}} = \frac{n-1}{n}\mathbf{S} + \bar{\mathbf{X}}_0\bar{\mathbf{X}}_0^\top.$$

2.2. Modelos de distribución destacados

Se tratamos con datos escalares podemos empregar modelos de distribución importantes como a distribución normal, pero existen outros como o modelo exponencial, beta, gamma, etc. Do mesmo xeito ocorre se tratamos con datos esféricos. Nesta sección introduciremos os modelos de distribución máis importantes para variables esféricas.

2.2.1. Distribución de von Mises-Fisher

Dentro das distribucións esféricas a máis importante é a de von Mises-Fisher. Isto débese a que modela datos direccionais empregando dous parámetros, $\boldsymbol{\mu}$ e k, o que fai que sexa sinxela de parametrizar, interpretar e simular, facilitando así a súa aplicación en problemas estatísticos.

Se $\mathbf{X}$ é unha variable aleatoria esférica con dimensión $p = 2$, a súa densidade logarítmica, que se denotará por $M(\boldsymbol{\mu}, k)$, pódese escribir do seguinte xeito:

$$\log f(\theta; \boldsymbol{\mu}, k) = k\cos(\theta - \boldsymbol{\mu}) - \log I_0(k) - \log(2\pi) = k\boldsymbol{\mu}^\top\mathbf{x} - \log I_0(k) - \log(2\pi),$$

onde $\mathbf{x} = (\cos\theta, \operatorname{sen}\theta)^\top$, $\boldsymbol{\mu} = (\cos\boldsymbol{\mu}, \operatorname{sen}\boldsymbol{\mu})^\top$ e I_0 denota a función de Bessel modificada de primeiro tipo e orde 0 (ver [12] para máis información). Polo tanto, unha xeneralización apropiada para S^{p-1} consiste nas distribucións con densidades logarítmicas lineais en $\mathbf{x}$, é dicir, que teñen densidades $f(x; \boldsymbol{\mu}, k)$ satisfacendo:

$$\log f(\mathbf{x}; \boldsymbol{\mu}, k) = k\boldsymbol{\mu}^\top\mathbf{x} + \textit{constante}. \tag{2.1.1}$$

Ademais, diremos que un vector unitario $\mathbf{X}$ segue a distribución de von Mises-Fisher de dimensión $(p-1)$, $M_p(\boldsymbol{\mu}, k)$, se a súa función de densidade é da seguinte forma:

$$f(\mathbf{x}; \boldsymbol{\mu}, k) = \left(\frac{k}{2}\right)^{p/2-1} \frac{1}{\Gamma(p/2)I_{p/2-1}(k)} \exp(k\boldsymbol{\mu}^\top\mathbf{x}), \tag{2.1.2}$$

onde k ≥ 0, $\|\boldsymbol{\mu}\| = 1$ e I_v denota a función de Bessel modificada de primeiro tipo e orde v, como xa dixemos anteriormente. Podemos ver unha representación desta función de densidade

na Figura 2.5, onde vemos que estas densidades teñen contornos circulares, e os puntos máis próximos á dirección media teñen unha maior densidade.

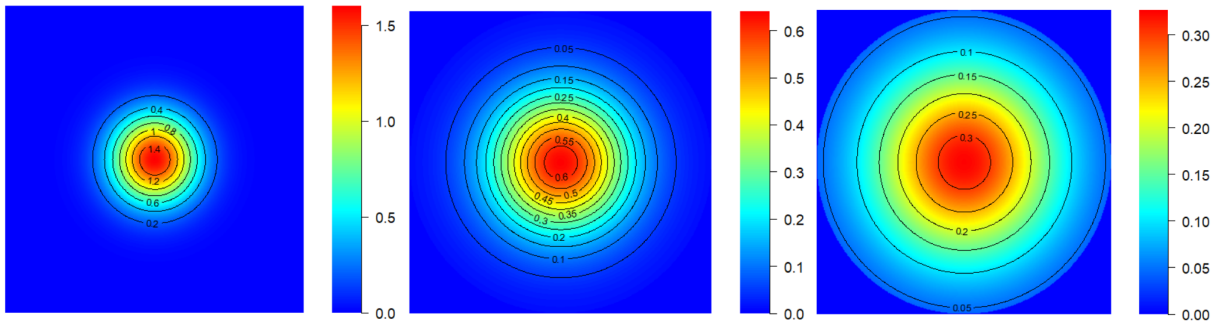

Figura 2.5: Representación da densidade de von Mises-Fisher para $k = 10$, $k = 4$ e para $k = 2$, respectivamente, considerando $\boldsymbol{\mu} = (1, 0, 0)$. A cor vermella indica unha maior densidade.

Podemos ver que esta distribución pertence á familia das exponenciais. Para iso imos recordar que unha distribución pertence a esta familia se a súa función de densidade, no caso xeral, se pode escribir da seguinte maneira:

$$f(x;\theta) = h(x)\exp(\eta(\theta)^\top T(x) - A(\theta)),$$

onde, no noso caso, $x = \mathbf{X}$, $\theta = \{\boldsymbol{\mu}, k\}$, $h(x) = \left(\frac{k}{2}\right)^{p/2-1} \frac{1}{\Gamma(p/2)I_{p/2-1}(k)}$ e $k\boldsymbol{\mu}^\top\mathbf{X}$ realiza o papel de $\eta(k,\boldsymbol{\mu})^\top T(x)$.

Aos parámetros $\boldsymbol{\mu}$ e k coñéceselles cos nomes de dirección media e parámetro de concentración (que é non negativo), respectivamente. Este $\boldsymbol{\mu}$ é o mesmo que o definido na sección anterior. Tamén podemos relacionar o parámetro k coa ρ anteriormente definida mediante a seguinte expresión:

$$\rho = A_p(k) = \frac{\int_{-1}^{1} t^2 e^{kt}(1-t^2)^{(p-3)/2}dt}{\int_{-1}^{1} e^{kt}(1-t^2)^{(p-3)/2}dt} = \frac{I_{p/2}(k)}{I_{p/2-1}(k)}. \tag{2.1.3}$$

Este parámetro k toma valores maiores ou iguais a 0: $k = 0$ indícanos unha distribución uniforme na esfera, mentres que a medida que k aumenta os puntos concéntranse máis arredor da dirección media $\boldsymbol{\mu}$.

Para rematar, acabaremos cunha idea sobre a estimación por máxima verosimilitude destes parámetros. Como a expresión da función de densidade vista en (2.1.2) é un modelo exponencial, o estimador do parámetro $k\boldsymbol{\mu}$ obtense partindo de

$$\hat{\rho}\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}.$$

Disto séguese que

$$\hat{\rho} = \bar{R}; \qquad \hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}_0,$$

co cal $A_p(\hat{k}) = \bar{R}$, é dicir, $\hat{k} = A_p^{-1}(\bar{R})$. Sabendo a expresión de $A_p(k)$, obtense que

$$\hat{k} = \frac{p-1}{2(1-\bar{R})} + \frac{p-3}{4} + O(n(1-\bar{R})), \text{ cando } \bar{R} \to 1.$$

Entón, para $\bar{R} \simeq 1$,

$$\hat{k} \simeq \frac{p-1}{2(1-\bar{R})}.$$

No caso $p = 3$, esta aproximación é válida para $\bar{R} \geq 0{,}9$.

Remataremos este apartado resaltando a importancia desta distribución de von Mises-Fisher. Esta distribución é fundamental na estatística e na análise de datos grazas á súa capacidade para tratar con datos direccionais sobre a esfera. Ademais, relaciónase ás veces coa distribución normal, xa que teñen en común que ambas funcións de densidade son unimodais e simétricas, a media coincide coa moda e ademais o estimador de máxima verosimilitude de $\boldsymbol{\mu}$ coincide co estimador de momentos. Engadir tamén que ten aplicacións en numerosos campos como a astronomía, a bioloxía ou a xeometría computacional.

2.2.2. Distribución de Kent

Como existen datos que non se axustan coa distribución de Fisher, provoca que teñamos a sospeita de que xorden de distribucións con contornos de densidade ovalada. Para tratar de explicar isto, introducimos un novo modelo con densidade

$$f(\mathbf{x}; \boldsymbol{\mu}, k, \mathbf{A}) = \frac{1}{a(k,\mathbf{A})} e^{k\boldsymbol{\mu}^\top \mathbf{x} + \mathbf{x}^\top \mathbf{A} \mathbf{x}},$$

onde $\mathbf{A}$ é unha matriz simétrica de dimensión $p \times p$ tal que $\mathrm{tr}\mathbf{A} = 0$ e $\mathbf{A}\boldsymbol{\mu} = 0$. Isto pódese ver detalladamente en [6].

No caso $p = 3$, poderemos escribir a matriz $\mathbf{A}$ como

$$\mathbf{A} = \beta(\boldsymbol{\xi_1}\boldsymbol{\xi_1}^\top - \boldsymbol{\xi_2}\boldsymbol{\xi_2}^\top),$$

con $\beta \geq 0$ e $\boldsymbol{\xi_1}, \boldsymbol{\xi_2}, \boldsymbol{\mu}$ ortogonais. Con isto, podemos reescribir a densidade da seguinte maneira:

$$f(\mathbf{x}; k, \beta, \boldsymbol{\Gamma}) = \frac{1}{c(\beta,k)} e^{kx_1 + \beta(x_2^2 - x_3^2)},$$

onde $\mathbf{\Gamma} = (\boldsymbol{\mu}, \boldsymbol{\xi_1}, \boldsymbol{\xi_2})$ é unha matriz ortogonal e $(x_1, x_2, x_3) = (\boldsymbol{\mu}^\top \mathbf{x}, \boldsymbol{\xi_1}^\top \mathbf{x}, \boldsymbol{\xi_2}^\top \mathbf{x})$.

Podemos ver unha representación da función de densidade na Figura 2.6. Aquí vemos como os puntos se concentran máis arredor da dirección media $\boldsymbol{\mu}$ a medida que aumenta k. En canto a β, a medida que aumenta móstrase unha maior asimetría, afastándose da dirección media.

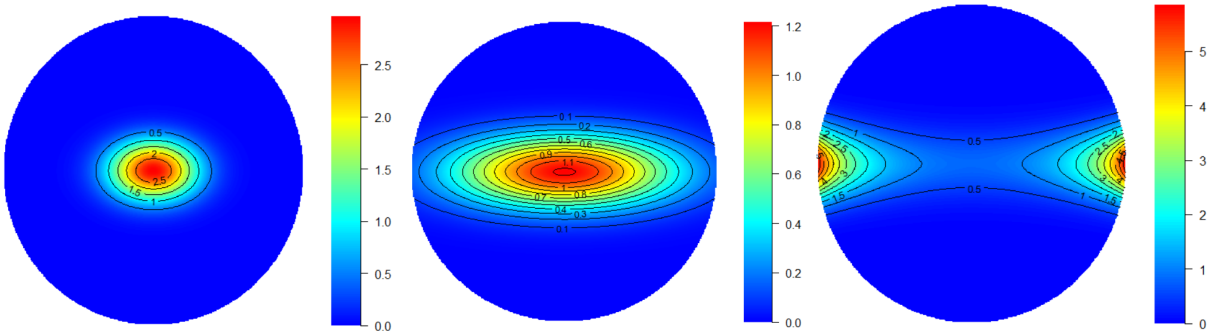

Figura 2.6: Representación gráfica da densidade de Kent para $k = 20$ e $\beta = 4$, para $k = 10$ e $\beta = 4$, e para $k = 10$ e $\beta = 6$, respectivamente. A cor vermella indica unha maior densidade. β debe ser menor ou igual a k para que a distribución sexa unimodal, pola contra sería bimodal.

Como observación, cabe destacar que a distribución de von Mises-Fisher é un caso particular desta distribución de Kent no caso de que $\beta = 0$.

2.2.3. Distribucións rotacionalmente simétricas

Unha propiedade moi importante das distribucións de von Mises-Fisher é que son rotacionalmente simétricas sobre as súas direccións modais. Entre estas distribucións, as que son continuas teñen funcións de densidade da forma

$$f(\mathbf{X}) = g(\boldsymbol{\mu}^\top \mathbf{X}),$$

sendo g unha función coñecida. Para distribucións con simetría rotacional con dirección $\boldsymbol{\mu}$ teremos

$$\mathbb{E}[\mathbf{X}] = \mathbb{E}[t]\boldsymbol{\mu}$$

e

$$\text{var}(\mathbf{X}) = \text{var(t)}\boldsymbol{\mu}\boldsymbol{\mu}^\top + \tfrac{1-\mathbb{E}[t^2]}{p-1}(I_p - \boldsymbol{\mu}\boldsymbol{\mu}^\top),$$

onde $t = \mathbf{X}^\top \boldsymbol{\mu}$.

Para o caso $p = 3$, a función de densidade de colatitude θ será

$$f(\theta; \theta_0, k) = \frac{k \operatorname{sen} \theta}{2\pi \operatorname{senh} k} e^{k \cos\theta \cos\theta_0} I_0(k \operatorname{sen}\theta \operatorname{sen}\theta_0),$$

con $0 \leq \theta_0 \leq \frac{\pi}{2}$.

Para rematar, cómpre mencionar que existen máis distribucións esféricas ademais das mencionadas, como por exemplo a distribución de Watson, a de Fisher-Bingham ou a uniforme esférica. Podemos mencionar que esta última se obtén como caso particular das anteriores, por exemplo, cando a distribución de von Mises-Fisher ten un parámetro de concentración k que tende a 0; ou que tanto β coma k sexan 0 no caso da distribución de Kent.

Capítulo 3

A regresión esférica

Moitas veces gustaríanos saber se existe algunha relación entre dúas variables esféricas distintas, de forma análoga ao que fixemos no primeiro capítulo coa regresión lineal. Neste caso temos o mesmo problema, pero agora non é trivial, xa que non ten sentido aplicar neste contexto un modelo lineal. Para iso imos introducir un coeficiente de correlación e un modelo de regresión para este tipo de variables esféricas. Todos os contidos que trataremos neste capítulo fan referencia a [7].

Dividiremos este capítulo en varias seccións. Comezaremos incorporando ao estudo unha serie de coeficientes de correlación que serán útiles na regresión esférica. Logo afondaremos no caso da regresión empregando a rotación, algo que quedará plasmado no último apartado, realizando un detallado estudo de simulación. Aquí, incorporaremos unha serie de datos simulados co obxectivo de obter estimacións co menor erro posible, así como interpretacións e gráficas onde quede perfectamente explicado todo este procedemento.

3.1. Correlación esférica

Podemos obter os coeficientes de correlación para variables aleatorias esféricas do seguinte xeito: sexan $(\mathbf{X}_1, \mathbf{Y}_1)$,..., $(\mathbf{X}_n, \mathbf{Y}_n)$ observacións independentes de vectores aleatorios $\mathbf{X}$ e $\mathbf{Y}$ de S^{p-1} e S^{q-1} respectivamente. Sexa agora

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix} \tag{3.1}$$

e

$$\mathbf{S}^* = \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} \mathbf{X}_i \mathbf{X}_i^\top & \mathbf{X}_i \mathbf{Y}_i^\top \\ \mathbf{Y}_i \mathbf{X}_i^\top & \mathbf{Y}_i \mathbf{Y}_i^\top \end{pmatrix} = \begin{pmatrix} \mathbf{S}_{11}^* & \mathbf{S}_{12}^* \\ \mathbf{S}_{21}^* & \mathbf{S}_{22}^* \end{pmatrix}, \tag{3.2}$$

que denotan a matriz de varianzas da mostra e a media mostral do produto de matrices de ($\mathbf{X}$, $\mathbf{Y}$).

Así, podemos definir o coeficiente de correlación r^2 como:

$$r^2 = tr(\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}),$$

(tendo en conta que $\mathbf{S}_{11}$ e $\mathbf{S}_{22}$ son matrices non singulares, ver [5] para máis información). Este coeficiente é unha medida da relación entre as covarianzas das variables. Ao facerlle a traza a esta matriz (a suma dos elementos diagonais) obtemos un escalar: un valor de r^2 preto a 1 indica que unha gran parte da variabilidade da variable resposta está explicada polas variables explicativas. Baixo a hipótese de independencia, teremos que

$$nr^2 \sim \chi^2_{pq} \quad \text{cando } n \to \infty$$

(tendo en conta tamén que a matriz de varianzas de ($\mathbf{X}$, $\mathbf{Y}$) é non singular). Ademais, no caso particular no que teñamos $p = q$ podemos escribir

$$\hat{\rho}_T = \frac{|\mathbf{S}_{12}^*|}{(|\mathbf{S}_{11}^*||\mathbf{S}_{22}^*|)^{1/2}},$$

que poderemos velo en [3].

Por outra parte, podemos definir outro coeficiente de correlación distinto do anterior para o caso esférico $p = q$, e denotarémolo por $\hat{\Delta}_n$. Para os pares de puntos en S^{p-1} $(\mathbf{X}_1, \mathbf{Y}_1)$,..., $(\mathbf{X}_n, \mathbf{Y}_n)$ defínese do seguinte xeito:

$$\hat{\Delta}_n = \binom{n}{p+1}^{-1} \sum_{\{i_1,\ldots,i_{p+1}\}} \delta_{i_1,\ldots,i_{p+1}}, \tag{3.3}$$

onde $\delta_{i_1,\ldots,i_{p+1}} = \text{sign}|\mathbf{X}_{i_1}, ..., \mathbf{X}_{i_{p+1}}| \times \text{sign}|\mathbf{Y}_{i_1}, ..., \mathbf{Y}_{i_{p+1}}|$. Unha interpretación xeométrica disto sería: $\delta_{i_1,\ldots,i_{p+1}}$ é 1 se o símplex de vértices $\mathbf{X}_{i_1}, ..., \mathbf{X}_{i_{p+1}}$ ten a mesma orientación que o símplex de vértices $\mathbf{Y}_{i_1}, ..., \mathbf{Y}_{i_{p+1}}$, e será -1 no outro caso.

3.2. Regresión empregando rotación

Realizar estudos de regresión dunha variable esférica $\mathbf{Y}$ de S^{p-1} sobre unha variable explicativa esférica $\mathbf{X}$ de S^{p-1} é algo moi común en diversos campos. Por exemplo, un importante problema cristalográfico consiste en relacionar un eixo $\mathbf{Y}$ dun cristal cun eixo $\mathbf{X}$ dun sistema de coordenadas estándar (Mackenzie, 1957); outro sería determinar a orientación dun satélite

espacial (Wahba, 1966) comparando as direccións $\mathbf{Y}$ das estrelas coas correspondentes direccións $\mathbf{X}$ no sistema de coordenadas terrestre; e outro igual de curioso pode ser a súa aplicación no ámbito das máquinas de visión (Kanatani, 1993), onde se comparan as direccións $\mathbf{Y}$ de distintos obxectos detectados por un sensor coas correspondentes direccións $\mathbf{X}$ percibidas por outro sensor.

Neste caso, a función de regresión é a dirección media condicionada ao valor de $\mathbf{X}$. Así, o modelo adopta a forma:

$$m(\mathbf{X}) = \mathbf{AX} \tag{3.4}$$

para algunha matriz de rotación $\mathbf{A}$ de $SO(p)$ (ortogonal e $\det(\mathbf{A})=1$). Un tipo sinxelo de modelos de regresión con funcións da forma (3.4) son os modelos de Chang ([14]), onde a distribución condicionada de $\mathbf{Y}$ dada $\mathbf{X}$ é circularmente simétrica con dirección media $\mathbf{AX}$, por iso, a densidade condicionada ten a forma

$$f(\mathbf{Y}; \mathbf{A}|\mathbf{X}) = g(\mathbf{Y}^\top \mathbf{AX}). \tag{3.5}$$

Por outra parte, dados os pares de observacións $(\mathbf{X}_1, \mathbf{Y}_1)$,..., $(\mathbf{X}_n, \mathbf{Y}_n)$, consideramos

$$\mathbf{S}_{12}^* = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{Y}_i^\top \tag{3.6}$$

e a súa descomposición en valores singulares $\mathbf{U\Lambda V}^\top$, é dicir, $\mathbf{S}_{12}^* = \mathbf{U\Lambda V}^\top$, onde $\mathbf{U}$ e $\mathbf{V}$ pertencen a $SO(p)$, e $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_{p-1}, \lambda_p)$, cumpríndose que $\lambda_1 \geq ... \geq |\lambda_p|$. Ademais, se a función g de (3.5) é crecente en $[-1, 1]$, a estimación de mínimos cadrados de $\mathbf{A}$, é dicir, o valor de $\mathbf{A}$ que minimiza

$$\textstyle\sum_{i=1}^{n} \|\mathbf{Y}_i - \mathbf{AX}_i\|^2,$$

será

$$\hat{\mathbf{A}} = \mathbf{VU}^\top. \tag{3.7}$$

Para demostrar isto comezaremos centrando $\mathbf{X}$ e $\mathbf{Y}$ do seguinte xeito:

$$\mathbf{X}' = \mathbf{X} - \tfrac{1}{n}\mathbf{X1}; \quad \mathbf{Y}' = \mathbf{Y} - \tfrac{1}{n}\mathbf{Y1},$$

onde $\mathbf{1}$ é un vector columna de uns. Logo, construiremos a matriz de covarianza $\mathbf{M} = \mathbf{X}'\mathbf{Y}'^\top$, da que realizaremos a súa descomposición en valores singulares

$$\mathbf{M} = \mathbf{U\Sigma V}^\top,$$

onde $\mathbf{U}$ e $\mathbf{V}$ son matrices ortogonais, e $\mathbf{\Sigma}$ é unha matriz diagonal que contén aos valores singulares. De aquí séguese que $\hat{\mathbf{A}} = \mathbf{VU}^\top$ minimiza o erro cadrático, xa que $\mathbf{V}$ describe como deben rotarse

os puntos de $\mathbf{Y}'$ para aliñalos con $\mathbf{X}'$, e viceversa para $\mathbf{U}^\top$. Deste xeito, multiplicando ambas, obtemos unha matriz de rotación óptima para aliñar $\mathbf{X}'$ con $\mathbf{Y}'$.

En particular, se a distribución condicionada é de von Mises-Fisher con concentración constante

$$\mathbf{Y}|\mathbf{X} \sim M_p(\mathbf{AX}, k), \tag{3.8}$$

entón, o estimador de máxima verosimilitude $\hat{\mathbf{A}}$ de $\mathbf{A}$ está dado por (3.7). Ademais, baixo (3.8), teremos que

$$2nk(r - tr(\mathbf{AS}_{12}^*)) \sim \chi^2_{p(p-1)/2} \quad \text{cando } n \to \infty,$$

onde $r = tr(\hat{\mathbf{A}}\mathbf{S}_{12}^*)$.

3.3. Estudo de simulación

A continuación poñeremos en práctica estes coñecementos previos, realizando un estudo baseado nuns datos simulados co programa *R*. Neste estudo simularemos n datos dunha variable explicativa $\mathbf{X}$, que segue unha distribución de von Mises-Fisher, e elixiremos unha matriz de rotación $\mathbf{A}$, tendo en conta que debe cumprir coa ortogonalidade e con ter determinante igual a 1. Posteriormente, simularemos a variable resposta $\mathbf{Y}$, que recordemos tamén seguirá unha distribución von Mises-Fisher:

$$\mathbf{Y}|\mathbf{X} \sim M_p(\mathbf{AX}, k), \tag{3.9}$$

onde recordemos que estamos a tratar con $p = 3$.

Para rematar, obteremos unha estimación da matriz $\mathbf{A}$, que denotaremos por $\hat{\mathbf{A}}$, e logo de repetir unha cantidade fixada B este mesmo proceso, calcularemos o erro cadrático medio cometido nesta estimación da matriz, empregando o método de Monte Carlo, para ver a calidade desta estimación. Explicaremos os valores e datos tomados en detalle na seguinte sección.

3.3.1. Escenarios de simulación

Teremos tres escenarios distintos, en función da matriz de rotación que empreguemos. Á súa vez, en cada un dos casos, teremos varios subcasos, segundo o número de datos n que decidamos simular, e segundo o parámetro k que empreguemos para simular a variable explicativa $\mathbf{X}$. Isto outórganos un total de 27 escenarios de simulación distintos, os cales poderemos comparar en función do erro cometido en cada un deles.

Será preciso facer unha breve observación sobre a notación que imos a empregar. Recordemos que, ata este punto, ambas variables $\mathbf{X}$ e $\mathbf{Y}$ teñen un parámetro de concentración denotado por k. Para diferencialo, a partir de agora denotarémolo por k' no caso da variable resposta $\mathbf{Y}$. Ademais, en cada un destes estudos, empregaremos $\boldsymbol{\mu} = (1, 0, 0)$ e unha $k' = 2$ fixados.

Comecemos definindo a matriz elixida para cada escenario. No primeiro modelo (denotado por M1) empregaremos a seguinte matriz:

$$\mathbf{A1} = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{3.10}$$

Podemos comprobar de maneira sinxela que esta é unha matriz de rotación, xa que

$$\begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}^{\top} = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{3.11}$$

e $\det(\mathbf{A1})$=1.

No segundo caso (M2), imos a utilizar a matriz

$$\mathbf{A2} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}, \tag{3.12}$$

a cal podemos ver que é unha matriz de rotación empregando o mesmo proceso que realizamos no anterior caso:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}^{\top} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ 0 & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{3.13}$$

e $\det(\mathbf{A2})$=1.

Por último, no terceiro caso (M3), a matriz escollida é

$$\mathbf{A3} = \begin{pmatrix} \frac{\sqrt{3}}{2} & 0 & -\frac{1}{2} \\ 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{\sqrt{3}}{2} \end{pmatrix}, \tag{3.14}$$

que, igual que dixemos para os dous modelos anteriores, se comproba de forma análoga que efectivamente é unha matriz de rotación.

Unha vez introducidas as tres matrices de rotación que nos distinguirán os tres modelos (M1, M2, M3), en cada un dos escenarios simularemos tres cantidades de datos distintas. Estes

números serán, por exemplo, $n = 50$, $n = 100$ e $n = 250$, elixidos desta maneira para que se distinga ben o estudo. A maiores, tamén teremos tres opcións distintas para o parámetro k á hora de simular a variable explicativa, onde recordamos que

$$\mathbf{X} \sim M_3(\boldsymbol{\mu}, k). \tag{3.15}$$

Deste xeito, mantendo $\boldsymbol{\mu}$ fixado para todos os casos, teremos, por exemplo, os casos $k = 1$, $k = 2$ e $k = 3$. Para rematar este procedemento, soamente nos faltaría calcular o erro cometido. Para isto, sumaremos cada un dos elementos de cada matriz, e posteriormente calcularemos a media aritmética destes B números (no noso caso será $B = 500$). Isto pódese escribir da seguinte maneira: sexan B matrices denotadas por $\mathbf{A_i}$ con elementos $a_{i,jk}$ para cada $i = 1, ..., B$. Estas matrices xorden de facer o cadrado da resta entre a matriz $\hat{\mathbf{A}}$ e a matriz correspondente. Entón obteremos a suma dos seus elementos como

$$S_i = \sum_{j,k} a_{i,jk}, \tag{3.16}$$

onde j e k recorren todos os elementos da matriz $\mathbf{A_i}$ correspondente. Posteriormente, realizaremos a media aritmética de todas estas matrices do seguinte xeito:

$$\frac{1}{B}\sum_{i=1}^{B} S_i = \frac{1}{B}\sum_{i=1}^{B}\sum_{j,k} a_{i,jk}. \tag{3.17}$$

Cómpre mencionar que a libraría de R que empregamos en todo este proceso de simulación, é a libraría *Directional*, procedente do paquete *Directional*, que permite analizar e modelar datos direccionais e esféricos, empregando neste caso a distribución de von Mises-Fisher.

3.3.2. Resultados

Agora, unha vez definidos os distintos escenarios de simulación cos respectivos valores dos parámetros, comezaremos coa labor de programar no programa R. Como ben dixemos anteriormente, este proceso consiste en simular n datos dunha variable explicativa $\mathbf{X}$ que segue unha distribución de von Mises-Fisher. Ademais, necesitaremos definir a matriz de rotación $\mathbf{A}$, para poder simular a variable resposta $\mathbf{Y}$. Remataremos obtendo unha estimación da matriz e calculando o erro cadrático medio empregando o método de Monte Carlo.

Observación 3.1. No desenvolvemento deste proceso de simulación, á hora de obter a estimación de $\mathbf{A}$, $\hat{\mathbf{A}}$, necesitamos calcular $\mathbf{S}_{12}^*$ do mesmo xeito que fixemos en (3.6), e a súa descomposición en valores singulares $\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\top$. Para iso, podemos empregar o comando *spher.reg* de R, de onde obtemos directamente a matriz buscada $\hat{\mathbf{A}}$. Pero, ademais desta forma, podemos calcular $\mathbf{S}_{12}^*$ e posteriormente obter a súa descomposición en valores singulares, sabendo que

$$\hat{\mathbf{A}} = \mathbf{V}\mathbf{U}^\top. \tag{3.18}$$

O código de todo este proceso, así como do procedemento completo, pódese ver no Anexo A deste traballo.

Despois deste proceso, obteremos o erro que se produce en cada estimación do seguinte xeito: unha vez obtidas as B matrices, neste caso imos a considerar $B = 500$, recordemos que imos a calcular a suma de todos os elementos de cada matriz para, posteriormente, calcular a media aritmética desa listaxe de números. Todo isto quedou explicado anteriormente en (3.16) e (3.17). Con iso, obteremos o erro cadrático medio de cada escenario, o que se pode recoller no seguinte Cadro 3.1:

		$k=1$	$k=3$	$k=5$
M1	$n=50$	0.17757	0.18223	0.21042
	$n=100$	0.08664	0.09486	0.09922
	$n=250$	0.03315	0.03657	0.04046
M2	$n=50$	0.16179	0.18207	0.21881
	$n=100$	0.08601	0.09088	0.10676
	$n=250$	0.03306	0.03652	0.04108
M3	$n=50$	0.18001	0.18889	0.21339
	$n=100$	0.08851	0.09178	0.09843
	$n=250$	0.03399	0.03673	0.04371

Cadro 3.1: Recompilación dos erros nos distintos casos de simulación empregando o método de Monte Carlo.

Cómpre recordar que para todos estes resultados empregamos un $\boldsymbol{\mu} = (1, 0, 0)$ fixado, xunto cunha $k' = 2$ para simular a variable resposta $\mathbf{Y}$.

Destes resultados poderemos obter unhas interpretacións que resultan de interese. Podemos observar como a elección da matriz de cada un dos distintos escenarios non provoca grandes cambios no erro, xa que teñen a mesma orde en cada un dos distintos casos: ordes de 10^{-1} e 10^{-2}. En canto ao número de datos n, podemos observar que, independentemente da elección de k, canto maior sexa este n mellor nos aproximamos á matriz inicial, é dicir, menos erro temos. Por último, en canto ao k que escollemos á hora de definir a variable explicativa mediante a distribución seguinte:

$$\mathbf{X} \sim M_3(\boldsymbol{\mu}, k), \tag{3.19}$$

podemos observar como, fixando unha n, vai aumentando o erro cometido a medida que aumentamos o valor de k. Vemos tamén que aumenta en maior proporción canto máis baixa é a

cantidade n. Isto dinos que, canto menor sexa o k coa que definimos a variable explicativa, menor será o erro cometido, e co cal mellor será a estimación da matriz. Representaremos graficamente estes erros na Figura 3.1, onde vemos que a proporción dos correspondentes erros apenas se distinguen. Para ver a diferencia dos valores, representaremos tamén os tres modelos xuntos:

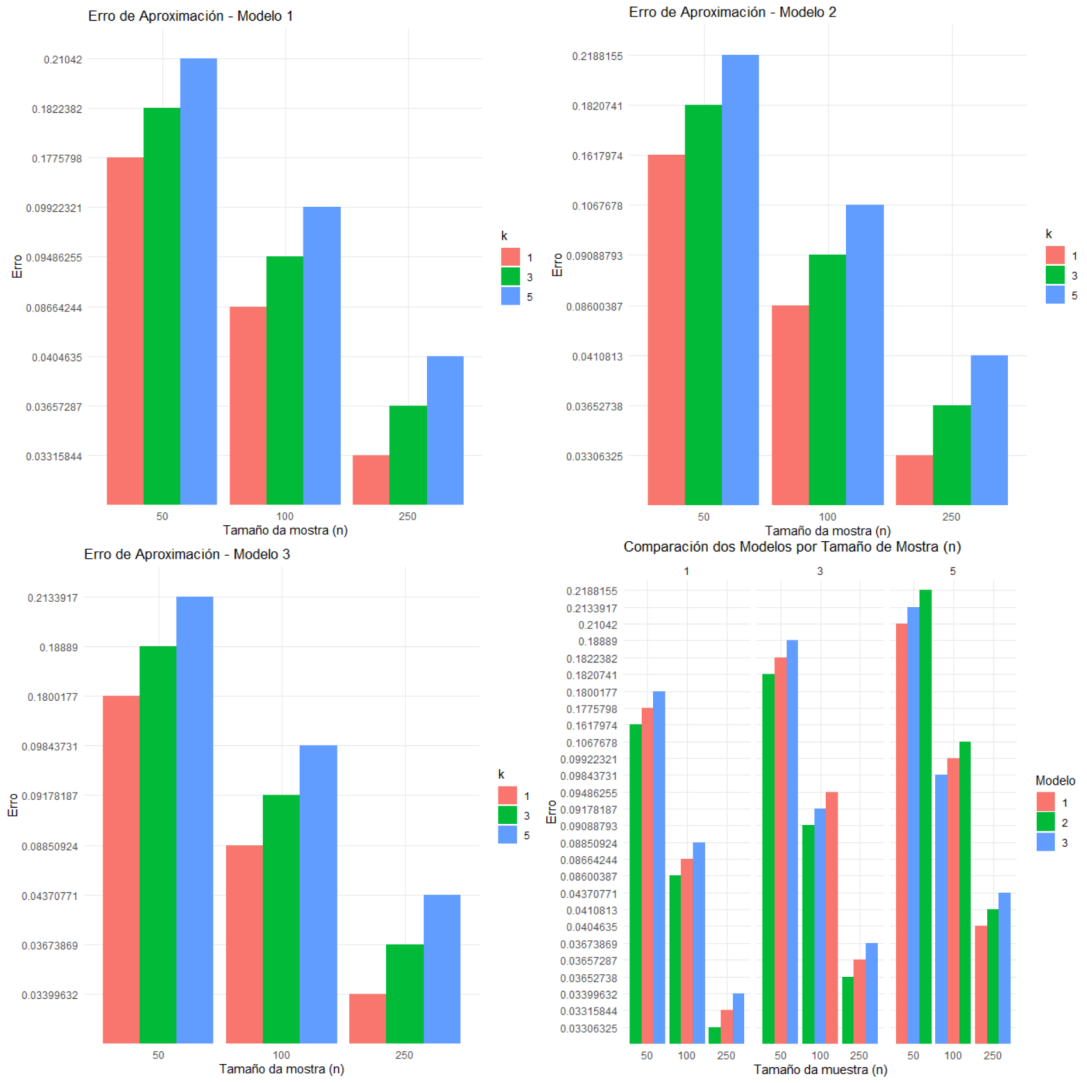

Figura 3.1: Representación mediante gráficos de barras do erro cometido en cada un dos casos de cada modelo para apreciar a súa similitude, así como a súa diferenza a escala.

Podemos seguir interpretando graficamente estes resultados de numerosas maneiras, dependendo das nosas preferencias e gustos. Xa vimos anteriormente a súa representación mediante gráficos de barras. Incorporamos agora outro tipo de gráfica, xunto coa súa explicación e interpretación. Vexámolo na Figura 3.2.

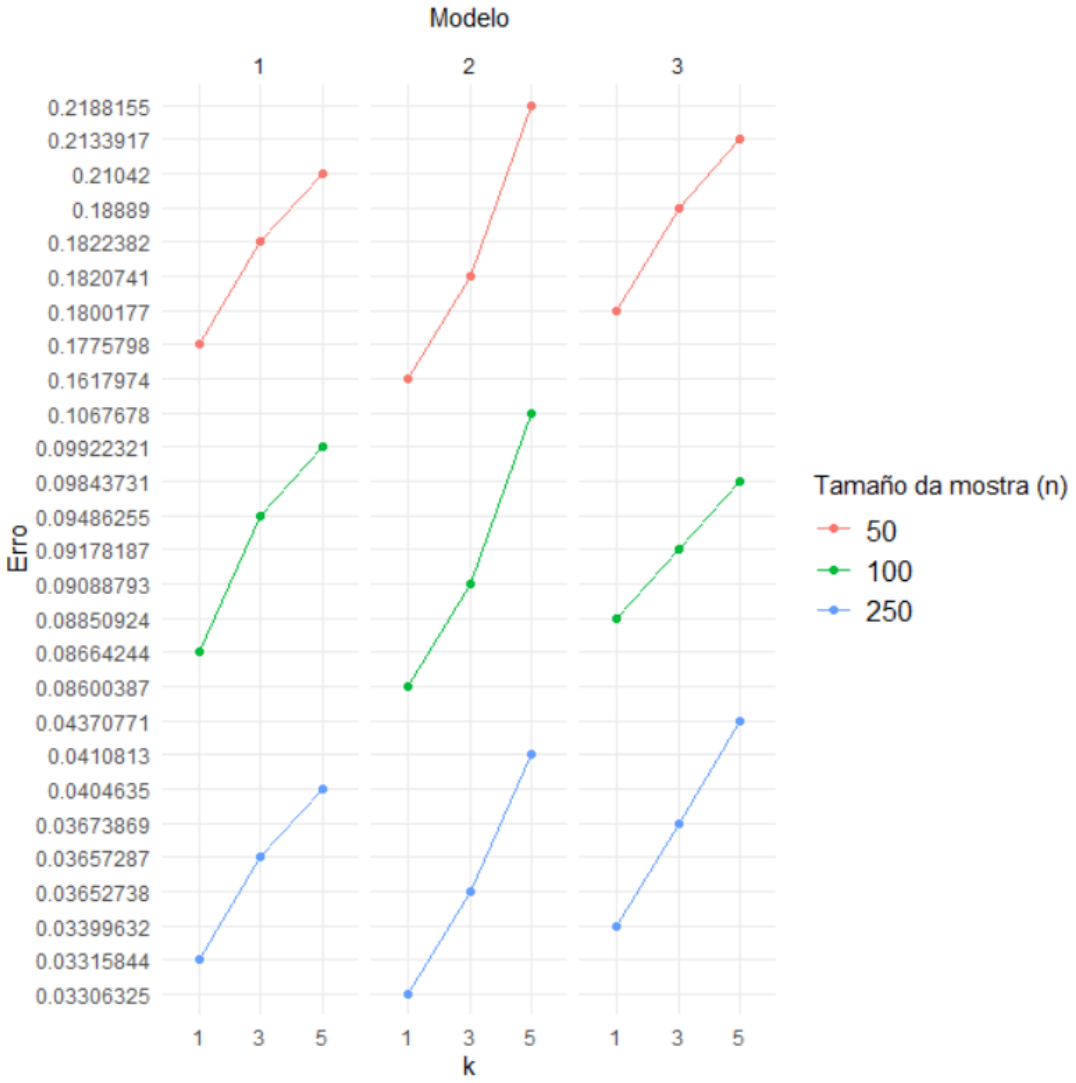

Figura 3.2: Gráfico de liñas múltiples para visualizar a variación do erro segundo o parámetro k e o tamaño da mostra n.

Nesta gráfica podemos ver doutra forma distinta como varía o erro en cada caso. Así, teremos unha escala que vai dende o menor dos erros cometidos ata o maior, e distinguiremos tres columnas verticais en función do modelo no que nos atopemos. A maiores, teremos tres zonas horizontais en función do tamaño mostral, e as liñas dos erros dependerán do valor do parámetro k escollido para a variable explicativa.

Capítulo 4

Conclusións e discusións

Neste último capítulo realizaremos unha revisión do traballo realizado ata este momento, reflexionando sobre o seu contido co principal obxectivo de analizar os resultados acadados e as posibles extensións que poderiamos engadirlle.

4.1. Síntese do traballo

Ao longo do traballo fomos introducindo diversos conceptos e resultados para poder adentrarnos na regresión esférica. Estruturamos o traballo en tres grandes bloques:

1. Introdución á regresión: neste capítulo incorporamos os conceptos clave da regresión lineal, da múltiple e da non lineal. Introducimos a notación que empregamos ao longo do traballo, así como resultados básicos e necesarios.

2. Introdución aos datos esféricos: comezamos a tratar con datos esféricos, definindo conceptos básicos como a dirección media da poboación, a lonxitude resultante media ou a dirección media. Ademais disto, incorporamos os modelos de distribución máis destacados, onde nos centramos especialmente na distribución de von Mises-Fisher, xa que a empregamos posteriormente no estudo de simulación.

3. A regresión esférica: con todo o introducido ata este punto, empezamos a facer regresión. Dividimos este capítulo en varias seccións. Comezamos definindo certos coeficientes de correlación, para posteriormente asentar as ideas xerais da regresión empregando a rotación. Isto plásmase na última sección, na que realizamos un estudo de simulación con R, explicando con detalle os resultados.

4.2. Posibles limitacións e extensións

Aínda que este traballo ofrece un achegamento importante á regresión esférica, podemos recoñecer certas limitacións. Unha destas limitacións será a ausencia de algunhas demostracións, xa que se saen dos obxectivos do traballo. Un exemplo pode ser a demostración do Teorema 1.1, no que se deixa [13] como referencia ao lector por se resulta de interese a súa demostración.

Outra limitación podería ser o uso exclusivo de datos simulados. Na realización do estudo de regresión empregamos integramente simulacións, sen aplicar o método a datos reais. Consideramos que, para este traballo, é máis que suficiente incorporar o estudo con datos simulados, aínda que podería resultar interesante engadir un estudo do mesmo estilo pero con datos reais. Do mesmo xeito, cómpre destacar que o noso estudo restrinxiuse a dimensión $p = 3$. Se consideraramos dimensións superiores teríamos unha maior complexidade computacional, xa que habería matrices máis grandes, ademais de que o comportamento das distribucións esféricas (como a von Mises-Fisher) poderían diferir en dimensións altas.

Ademais, outro caso podería ser a optimización da parte de programación que se levou a cabo no programa R. Neste traballo realizáronse numerosos códigos de programación, tanto para obter resultados como para obter gráficas. Estes códigos realizáronse da forma máis clara posible, aínda que seguramente se poidan optimizar e adaptarse para futuros estudos.

Para rematar, baseándonos en [10] e [15], cómpre comentar que existen outros modelos de regresión para variables esféricas a maiores dos incorporados neste traballo. Un exemplo disto será o modelo ESAG (Elliptically Symmetric Angular Gaussian). Para unha $\boldsymbol{\mu} \in \mathbb{R}^p$ e unha matriz de covarianzas $\mathbf{V}$, consideramos $\mathbf{z} \sim N(\boldsymbol{\mu}, \mathbf{V})$. Defínese coas condicións

$$\mathbf{V}\boldsymbol{\mu} = \boldsymbol{\mu}; \quad \det(\mathbf{V}) = 1, \tag{4.1}$$

e ten a seguinte función de densidade:

$$f(\mathbf{y}|\boldsymbol{\mu}, \mathbf{V}) = \frac{1}{2\pi(\mathbf{y}^\top \mathbf{V}^{-1}\mathbf{y})^{3/2}} \times \exp\left[\frac{1}{2}\left\{\frac{(\mathbf{y}^\top\boldsymbol{\mu})^2}{\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{y}} - \boldsymbol{\mu}^\top\boldsymbol{\mu}\right\}\right] \times M\left\{\frac{\mathbf{y}^\top\boldsymbol{\mu}}{(\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{y})^{1/2}}\right\},$$

sendo $\mathbf{y}$ un vector da esfera e $M(\alpha) = \alpha\phi(\alpha) + (1+\alpha^2)\psi(\alpha)$, sendo ϕ e ψ a función de densidade dunha normal estándar e a función de distribución acumulada respectivamente. Podemos empregar este modelo en ámbitos como a xeoloxía, para modelar orientacións de capas rochosas en función de variables ambientais; na meteoroloxía para as direccións do vento, etc.

Outro exemplo será, entre moitos outros, o modelo SIPC (Spherical Isotropic Projected Cauchy distribution). Esta distribución será rotacionalmente simétrica e terá a seguinte función de densidade:

$$\mathrm{f}(\mathbf{y}) = \frac{(\gamma^2+1)\sqrt{\delta}\left[\arctan 2(\sqrt{\delta},-\alpha)-\arctan 2(\sqrt{\delta},\alpha)+\pi\right]+2\alpha\delta}{4\pi^2\delta^2},$$

onde $\delta = \gamma^2 + 1 - \alpha^2$, con $\alpha = \mathbf{y}^\top \boldsymbol{\mu}$ e $\gamma = \|\boldsymbol{\mu}\|$. Este modelo pódese empregar en campos como a física, bioloxía e astronomía. A finalidade de engadir estes dous exemplos de outros modelos distintos non é máis que a de poñer en manifesto a cantidade de modelos de regresión esférica que nos podemos atopar e que se empregan en distintos e numerosos ámbitos.

En resumo, este traballo intenta ofrecer ao lector unha visión xeral da regresión esférica, dende os seus fundamentos teóricos ata a súa implantación práctica mediante simulacións. Recopilamos as limitacións que ten, así como as posibles extensións para ver máis alá do realizado no traballo. Os resultados incorporados non só destacan a utilidade da regresión esférica, senón que tamén sentan as bases para futuras investigacións neste campo.

Anexo A

Códigos de simulación

Neste anexo imos a incorporar os códigos realizados no programa R que serven para obter a estimación da matriz $\mathbf{A}$ de xeito manual mediante a fórmula, xunto co código dun dos 27 escenarios de simulación realizados:

- Obtención da matriz $\hat{\mathbf{A}}$ mediante a fórmula:

```
S12=matrix(0, nrow = w, ncol = w)
for (i in 1:n) {
S12=S12 + matrix(X[i,], nrow=w)%*%matrix(Y[i,], nrow=1)
}
S12=(1/n)*S12; S12
svd(S12)
V=svd(S12)$v
U=svd(S12)$u
Â=V%*%t(U); Â
```

- Código completo do proceso de simulación para, por exemplo, o escenario coa matriz **A1**, $n = 50$ e $k = 1$:

```
n=50; k=1

X=rvmf(n,c(1,0,0),k); X
w=length(c(1,0,0))

A=matrix(c(0,1,0,-1,0,0,0,0,1), ncol=w)
```

```
Y=matrix(NA, nrow = n, ncol = 3); Y
for (i in 1:n) {
 Y[i,]=rvmf(1,A%*%matrix(X[i,],nrow=3),2)
}
Y

modelo=spher.reg(Y, X)
modelo$A

erro=rep(0,length=n);
matrices=list()
B=500;
for(b in 1:B) {
 X=rvmf(n,c(1,0,0),k);
 w=length(c(1,0,0));
 A=matrix(c(0,1,0,-1,0,0,0,0,1), ncol=w);
 Y=matrix(NA, nrow = n, ncol = 3);
 for (j in 1:n) {
  Y[j,]=rvmf(1,A%*%matrix(X[j,],nrow=3),2)
 }
 modelo=spher.reg(Y, X)
 modelo$A
 matrices[[b]]=(modelo$A-A)^2
}
suma=lapply(matrices, sum)
mean(unlist(suma))
```

Anexo B

Códigos gráficos

Do mesmo xeito que fixemos no anterior anexo, neste introduciremos os códigos que empregamos para as figuras incorporadas ao longo do traballo:

- Código para xerar a esfera:

```
library(rgl); library(misc3d)
zoom<-0.75
windowRect<-c(500,50,0,0)
windowRect[3]=windowRect[1]+256*2
windowRect[4]=windowRect[2]+256*2
open3d(zoom = zoom, windowRect=windowRect)
plot_sphere <- function(radius = 1, center = c(0, 0, 0), color = "lightblue") {
 theta <- seq(0, pi, length.out = 30)
 phi <- seq(0, 2 * pi, length.out = 30)
 x <- outer(sin(theta), cos(phi)) * radius + center[1]
 y <- outer(sin(theta), sin(phi)) * radius + center[2]
 z <- outer(cos(theta), rep(1, length(phi))) * radius + center[3]
 persp3d(x, y, z, col = color, alpha = 0.3, back = "lines", box=FALSE, axes=FALSE,
 xlab="",ylab="",zlab="")
 segments3d(rbind(c(0,0,0),c(0,0,1)))
 segments3d(rbind(c(0,0,0),c(0,1,0)))
 segments3d(rbind(c(0,0,0),c(1,0,0)))
}
plot_vector <- function(theta, phi, color = "red", lwd=3) {
 x <- sin(theta) * cos(phi)
 y <- sin(theta) * sin(phi)
```

```
 z <- cos(theta)
 segments3d(rbind(c(0, 0, 0), c(x, y, z)), col = color, lwd = lwd, box=FALSE)
 segments3d(rbind(c(0, 0, 0), c(x, y, 0)), col = "blue", lwd = 2, lty = "dashed")
 segments3d(rbind(c(x, y, 0), c(x, y, z)), col = "green", lwd = 2, lty = "dashed")
 theta_arc <- seq(0, theta, length.out = 30)
 arc_x <- 0.5*sin(theta_arc) * cos(phi)
 arc_y <- 0.5*sin(theta_arc) * sin(phi)
 arc_z <- 0.5*cos(theta_arc)
 lines3d(arc_x, arc_y, arc_z, col = "purple", lwd = 2)
 phi_arc <- seq(0, phi, length.out = 30)
 arc_x_phi <- 0.5 * sin(theta) * cos(phi_arc)
 arc_y_phi <- 0.5 * sin(theta) * sin(phi_arc)
 arc_z_phi <- rep(0, length(phi_arc))
 lines3d(arc_x_phi, arc_y_phi, arc_z_phi, col = "orange", lwd = 2)
}
plot_sphere()
```

- Figura 2.1:

```
theta <- pi / 6; phi <- pi / 3
plot_vector(theta, phi)
text3d(sin(theta)*cos(phi)+0.01,sin(theta)*sin(phi)+0,cos(theta)+0.1,"x",cex=1.2)
text3d(0.35,0.2,0,expression(phi),cex=1.5)
text3d(0.05,0.15,0.65,expression(theta),cex=1.5)
```

- Figura 2.2:

```
segments3d(rbind(c(0, 0, 0), c(0, 0, 1)), col="green", lwd = 2, lty = "dashed")
text3d(0,0,1.1,expression(mu),cex=1.5)
segments3d(rbind(c(0, 0, 1/sqrt(3)), c(0, sqrt(2)/sqrt(3), 1/sqrt(3))), col="blue",
lwd = 2, lty = "dashed")
text3d(0,sqrt(2)/sqrt(3)+0.1, 1/sqrt(3),expression(x),cex=1.5)
segments3d(rbind(c(0, 0, 0), c(0, sqrt(2)/sqrt(3), 1/sqrt(3))), col="red", lwd = 2,
lty = "dashed")
arrow3d(c(-0.1,-0.1,0.1),c(-0.1, -0.1, 1/sqrt(3)), type="rotation", col="black")
arrow3d(c(-0.1, -0.1, 1/sqrt(3)-0.1),c(-0.1,-0.1,0), type="rotation", col="black")
text3d(-0.2,-0.2,1/2*1/sqrt(3),expression(t),cex=1.5)
arrow3d(c(0,0,1),c(0,sqrt(2)/sqrt(3),1), type="rotation", col="black")
text3d(0,sqrt(2)/sqrt(3)+0.05,1,expression(epsilon),cex=1.5)
```

- Figura 2.3:

```
library(Directional)
datos= rvmf(200,mu=rep(1/sqrt(3),3),k=10)
points3d(datos,col=2)
k=10; rho = besselI(k, 1) / besselI(k, 0); rho
datos= rvmf(200,mu=rep(1/sqrt(3),3),k=1)
points3d(datos,col=2)
k=1; rho = besselI(k, 1) / besselI(k, 0); rho
datos= rvmf(200,mu=c(1,1,1),k=10)
points3d(datos,col=2)
k=10; rho = besselI(k, 1) / besselI(k, 0); rho
```

- Figura 2.4:

```
points3d(0,0,0, col="black",size=7)
points3d(0,0.5,0.5, col="black",size=7)
points3d(0,0.3,0.3, col="black",size=7)
segments(c(0,0,0), c(0,0.3,0.3))
segments3d(rbind(c(0, 0, 0), c(0,0.5,0.5)), col="black", lwd = 2, lty = "dashed")
arrow3d(c(0,0.175,0.075),c(0,0.38,0.28), type="rotation", col="black")
arrow3d(c(0,0.325,0.225),c(0,0.04,-0.04), type="rotation", col="black")
text3d(0,0.28,0.1,expression(bold(bar(R))),cex=0.9)
text3d(0,0.5,0.61,expression(bold(bar(X[0]))),cex=0.9)
text3d(0,0.3,0.4,expression(bold(bar(X))),cex=0.9)
text3d(0,-0.1,0.4,expression(x),cex=0.9)
text3d(0,0.7,0.4,expression(x),cex=0.9)
text3d(0,0.1,0.7,expression(x),cex=0.9)
text3d(0.4,0.4,0,expression(x),cex=0.9)
```

- Figura 2.5:

```
library(Directional)
vmf.contour(10)
vmf.contour(4)
vmf.contour(2)
```

- Figura 2.6:

```
library(Directional)
```

```
kent.contour(20,4)
kent.contour(10,4)
kent.contour(10,6)
```

- Figura 3.1 (análogo para o resto de casos):

```
library(ggplot2)
datos=read.table("datos.txt",header=TRUE,dec=",")
attach(datos)
datos_caso1 <- subset(datos, CASO == 1)
ggplot(datos_caso1, aes(x = factor(n), y = Erro, fill = factor(k))) +
geom_bar(stat = "identity", position = "dodge") +
labs(title = "Erro de Aproximación - Modelo 1", x = "Tamaño da mostra (n)", y = "Erro",
fill = "k") +
theme_minimal()
```

- Figura 3.2:

```
library(ggplot2)
ggplot(datos, aes(x = factor(k), y = Erro, color = factor(n), group =
interaction(n, CASO))) +
geom_line() +
geom_point() +
facet_wrap(~ CASO) +
labs(
subtitle = "Modelo", x = "k", y = "Erro", color = "Tamaño da mostra (n)") +
theme_minimal() +
theme(
plot.subtitle = element_text(hjust = 0.5),
legend.text = element_text(size = 12)
)
```

Bibliografía

[1] Faraway, J.J. (2004). *Linear Models with R.* Chapman and Hall.

[2] Faraway, J.J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models.* Chapman and Hall.

[3] Fisher, N. e Lee, A.J. (1983). *A correlation coefficient for circular data.* Biometrika, 70(2), 327-332.

[4] Gratton, S., Lawless, A. e Nichols, N. (2007). *Approximate Gauss-Newton methods for non-linear least squares problems.* SIAM Journal on Optimization (SIOPT), 18 (1), 106-132.

[5] Jupp, P.E., e Mardia, K.V. (1980). *A general correlation coefficient for directional data and related regression problems.* Biometrika, 67(1), 163–173.

[6] Kent, J.T. (1982). *The Fisher–Bingham distribution on the sphere.* Journal of the Royal Statistical Society: Series B. 44 (1), 71-80.

[7] Mardia, K.V. e Jupp, P.E. (2009). *Directional Statistics.* Wiley-Intersciencie.

[8] Moré, J.J. (1977) *The Levenberg-Marquardt Algorithm: Implementation and Theory.* Springer.

[9] Panaretos, V.M. (2016). *Stadistics for Mathematicians: A Rigurous First Course.* Birkhäuser.

[10] P. J. Paine, S. P. Preston, M. Tsagris and Andrew T. A. Wood (2020). *Spherical regression models with general covariates and anisotropic errors.* Statistics and Computing, 30(1), 153–165.

[11] R. Y. Rubinstein., Kroese D.P (2007). *Simulation and the Monte Carlo Method, 2a ed..* Wiley-Intersciencie.

[12] Simmons, George F. (1993). *Ecuaciones Diferenciales con Aplicaciones y Notas Históricas, 2a ed.* McGraw-Hill.

[13] Soch, Joram, et al. (2024). *StatProofBook: The Book of Statistical Proofs* (Version 2023). Zenodo.

[14] Ted Chang. (1986). *Spherical regression.* Annals of Statistics, 14(3), 907-924.

[15] Tsagris M. and Alzeley O. (2025). *Circular and spherical projected Cauchy distributions: a novel framework for circular and directional data modeling.* Australian and New Zealand Journal of Statistics, 67(1), 77-103.